



SAULO ALMEIDA MORELLATO

INFERÊNCIA ESTATÍSTICA PARA REGRESSÃO MÚLTIPLA
H-Splines

CAMPINAS

2014



UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística
e Computação Científica

SAULO ALMEIDA MORELLATO

INFERÊNCIA ESTATÍSTICA PARA REGRESSÃO MÚLTIPLA
H-Splines

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em estatística.

Orientador: Ronaldo Dias

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA
TESE DEFENDIDA PELO ALUNO SAULO ALMEIDA MORELLATO, E ORIENTADA PELO PROF. DR. RONALDO DIAS.

Assinatura do Orientador

CAMPINAS

2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

M815i Morellato, Saulo Almeida, 1983-
Inferência estatística para regressão múltipla h-splines / Saulo Almeida
Morellato. – Campinas, SP : [s.n.], 2014.

Orientador: Ronaldo Dias.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de
Matemática, Estatística e Computação Científica.

1. Modelos aditivos generalizados. 2. Spline, Teoria do. 3. Métodos MCMC. 4.
Testes de hipóteses estatísticas. 5. Análise de Regressão. I. Dias, Ronaldo, 1959-
II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e
Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Statistical inference for h-splines multiple regression

Palavras-chave em inglês:

Generalized additive models

Spline theory

MCMC methods

Statistical hypothesis testing

Regression analysis

Área de concentração: Estatística

Titulação: Doutor em Estatística

Banca examinadora:

Ronaldo Dias [Orientador]

Mariana Rodrigues Motta

Mário de Castro Andrade Filho

Carlos Alberto de Bragança Pereira

Carlos Alberto Ribeiro Diniz

Data de defesa: 14-04-2014

Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 14 de abril de 2014 e aprovada

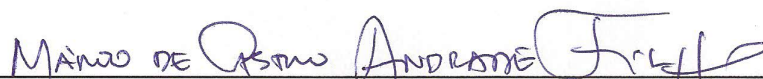
Pela Banca Examinadora composta pelos Profs. Drs.



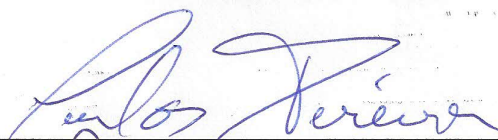
Prof(a). Dr(a). RONALDO DIAS



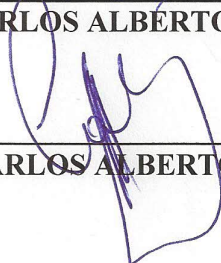
Prof(a). Dr(a). MARIANA RODRIGUES MOTTA



Prof(a). Dr(a). MÁRIO DE CASTRO ANDRADE FILHO



Prof(a). Dr(a). CARLOS ALBERTO DE BRAGANÇA PEREIRA



Prof(a). Dr(a). CARLOS ALBERTO RIBEIRO DINIZ

Abstract

In this work we discuss two inference problems related to multiple nonparametric regression: estimation in additive models using a nonparametric method and hypotheses testing for equality of curves, also considering additive models. In the estimation step, it is constructed a generalization of the h-splines method, both in the sequential adaptive context proposed by Dias (1999), and in the Bayesian context proposed by Dias and Gamerman (2002). The h-splines methods provide an automatic choice of the number of bases used in the estimation of the model. Simulation studies show that the results obtained by proposed estimation methods are superior to those achieved in the packages `gamlss`, `mgcv` and `DPpackage` in R. Two hypotheses testing are created to test $H_0 : f = f_0$. A hypotheses test that has a decision rule based on the integrated squared distance between two curves, for adaptive sequential approach, and another based on the Bayesian evidence measure proposed by Pereira and Stern (1999). In Bayesian hypothesis testing the performance measure of evidence is observed in several simulation scenarios. The proposed measure showed a behavior that is consistent with evidence favorable to H_0 . In the test based on the distance between the curves, the power of the test was estimated at various scenarios using simulations, and the results are satisfactory. At the end of the work the proposed procedures of estimation and hypotheses testing are applied in a dataset concerning to the work of Tanaka and Nishii (2009) about the deforestation in East Asia. The objective is to choose one amongst eight models. The tests point to a pair of models as being the most suitable.

Keywords: Generalized additive modelos; Spline theory; MCMC methods; Statistical hypothesis testing; Regression analysis.

Resumo

Este trabalho aborda dois problemas de inferência relacionados à regressão múltipla não paramétrica: a estimação em modelos aditivos usando um método não paramétrico e o teste de hipóteses para igualdade de curvas ajustadas a partir do modelo. Na etapa de estimação é construída uma generalização dos métodos *h-splines*, tanto no contexto sequencial adaptativo proposto

por Dias (1999), quanto no contexto bayesiano proposto por Dias e Gamerman (2002). Os métodos *h-splines* fornecem uma escolha automática do número de bases utilizada na estimação do modelo. Estudos de simulação mostram que os resultados obtidos pelos métodos de estimação propostos são superiores aos conseguidos nos pacotes `gamlss`, `mgcv` e `DPpackage` em R. São criados dois testes de hipóteses para testar $H_0 : f = f_0$. Um teste de hipóteses que tem sua regra de decisão baseada na distância quadrática integrada entre duas curvas, referente à abordagem sequencial adaptativa, e outro baseado na medida de evidência bayesiana proposta por Pereira e Stern (1999). No teste de hipóteses bayesiano o desempenho da medida de evidência é observado em vários cenários de simulação. A medida proposta apresentou um comportamento que condiz com uma medida de evidência favorável à hipótese H_0 . No teste baseado na distância entre curvas, o poder do teste foi estimado em diversos cenários usando simulações e os resultados são satisfatórios. Os procedimentos propostos de estimação e teste de hipóteses são aplicados a um conjunto de dados referente ao trabalho de Tanaka e Nishii (2009) sobre o desmatamento no leste da Ásia. O objetivo é escolher um entre oito modelos candidatos. Os testes concordaram apontando um par de modelos como sendo os mais adequados.

Palavras-chave: Modelos aditivos generalizados; Teoria de *splines*; Métodos MCMC; Teste de hipóteses; Análise de Regressão.

Sumário

Agradecimentos	xi
1 Introdução	1
1.1 Motivação	1
1.2 Funções Base	2
1.2.1 Bases Polinomiais	2
1.2.2 Funções <i>Splines</i>	3
1.2.3 Bases B- <i>splines</i>	4
1.3 Métodos para Estimação de Curvas	6
1.3.1 Regressão com <i>Splines</i>	6
1.3.2 Suavização por <i>Splines</i>	6
1.3.3 Regressão Penalizada com <i>Splines</i>	7
1.3.4 H- <i>splines</i>	7
1.4 Generalização do Método H- <i>splines</i>	8
1.5 Organização do Trabalho	9
2 Estimação Bayesiana para Regressão Múltipla H-<i>Splines</i>	11
2.1 Introdução	11
2.2 Regressão Penalizada para Modelo Aditivos	12
2.3 Posicionamento dos Nós	13
2.4 Abordagem Bayesiana para o Método H- <i>splines</i>	15
2.5 Dados Simulados	20

2.5.1	Exemplo 1	21
2.5.2	Exemplo 2	23
2.5.3	Exemplo 3	27
3	Evidência Bayesiana para Modelos de Regressão Múltipla H-<i>Splines</i>	30
3.1	Introdução	30
3.2	Teste de Hipóteses Bayesiano para Igualdade de Curvas	31
3.3	Dados Simulados	34
4	Estimação Sequencial Adaptativa para Regressão Múltipla H-<i>Splines</i>	37
4.1	Introdução	37
4.2	Estimação Adaptativa Múltipla	38
4.3	Seleção da Distância $d(\cdot, \cdot)$	40
4.4	Dados Simulados	42
4.4.1	Exemplo 1 (Continuação)	42
4.4.2	Exemplo 2 (Continuação)	43
4.4.3	Exemplo 3 (Continuação)	45
5	Teste de Hipóteses para Regressão Múltipla H-<i>Splines</i>	48
5.1	Introdução	48
5.2	Teste DQI para Igualdade de Curvas	49
5.3	Simulações	51
6	Variáveis Resposta com Restrições	53
6.1	Introdução	53
6.2	Modelos Aditivos Generalizados Livres de Distribuição	54
6.3	Dados Simulados: Exemplo 4	57
6.3.1	Exemplo 4.a: Resposta sem Restrições	58
6.3.2	Exemplo 4.b: Resposta Binária	59
6.3.3	Exemplo 4.c: Resposta em Contagem	61

7	Aplicação a Dados Reais	64
7.1	Descrição dos Dados	64
7.2	Abordagem Bayesiana	66
7.2.1	Teste Bayesiano	67
7.3	Abordagem Sequencial	68
7.3.1	Teste DQI	69
7.4	Análise dos Resultados	70
8	Considerações Finais	72
8.1	Conclusões	72
8.2	Propostas para Trabalhos Futuros	74
	Referências	75
A	Forma Matricial para $\int [f''(t)]^2 dt$	78
B	Identificabilidade do Modelo Aditivo	79
I	Licença	81
I.1	Sobre a licença dessa obra	81

Agradecimentos

Em primeiro lugar agradeço a Deus por mais essa realização em minha vida.

À minha querida namorada Lorena Knupp, que é fonte de inspiração, serenidade e luz em minha vida.

Aos meus pais, que não pouparam esforços para que eu alcançasse este objetivo.

Aos meus irmãos e sobrinhos pelo apoio e incentivo durante todos os anos de estudo.

Ao professor Dr. Ronaldo Dias pela orientação e pelas idéias durante todo este trabalho.

Aos colegas do Departamento de Estatística da UFES pelo apoio e amizade.

Finalmente, à CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela assistência financeira.

À todos, meu muito obrigado.

Saulo Almeida Morellato

Capítulo 1

Introdução

1.1 Motivação

Existem diversas situações em que se deseja fazer inferência usando modelos não lineares do tipo

$$Y = f(X) + \epsilon, \tag{1.1.1}$$

em que $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$ e sendo o erro de X . Geralmente, a questão que guia todo o processo de estimação do modelo é: qual a forma de f ? Em alguns casos f é totalmente desconhecida; em outros tem-se dúvida se a forma f_0 é adequada para representar f ; ou ainda se podemos substituir f_0 por outro modelo f_0^* mais parcimonioso sem perder na qualidade do ajuste.

O principal objetivo deste trabalho é tentar responder apropriadamente estas questões. A idéia é generalizar, para o caso com várias covariáveis, usando a regressão não paramétrica via *splines* híbridos (*h-splines*) proposta por Dias (1999) e estendida para uma abordagem bayesiana por Dias e Gamerman (2002). Esta generalização ajudará à responder a primeira questão, estimando uma forma para f . Além disso, são criados testes de hipóteses que ajudarão a responder se f_0 ou f_0^* é mais adequada para representar f .

Antes de propor um método para estimar uma forma para f , é necessário fazer uma rápida revisão de alguns métodos já existentes. As seções 1.2 e 1.3 descrevem métodos usuais de aproximar uma função f por uma combinação linear de funções base conhecidas.

1.2 Funções Base

Um procedimento não paramétrico usual é aproximar f através de uma combinação linear de K funções base conhecidas b_1, \dots, b_K , ou seja,

$$f(x) \approx f_K(x) = \sum_{j=1}^K \theta_j b_j(x), \quad (1.2.1)$$

em que os coeficientes $\theta_1, \dots, \theta_K$ são valores a serem estimados.

É importante que as funções base apresentem características que se relacionem com aquelas encontradas nas funções que serão aproximadas. Teoricamente, uma base deveria ser escolhida por produzir uma excelente aproximação usando ao mesmo tempo um número K pequeno de funções base. Isso não somente implica menos computação, mas também os coeficientes em si podem ser usados para descrever os dados de forma interessante. Consequentemente, existem bases que não são apropriadas para certas aplicações. Não existe algo como uma base universal que seja boa para todos os casos.

Bases comumente utilizadas são as polinomiais, os *splines* e os *B-splines*, que são apresentados nas seções 1.2.1, 1.2.2 e 1.2.3, respectivamente. Uma outra base de funções bastante conhecida é a obtida pelas séries de Fourier,

$$f(x) = \theta_0 + \theta_1 \sin(vx) + \theta_2 \cos(vx) + \theta_3 \sin(2vx) + \theta_4 \cos(2vx) + \dots,$$

definida por bases $b_0(x) = 1$, $b_{2r-1}(x) = \sin(rvx)$ e $b_{2r}(x) = \cos(rvx)$. Estas bases são periódicas, com período $2\pi/v$ determinado pelo parâmetro v .

1.2.1 Bases Polinomiais

Uma alternativa é aproximar a função f como uma combinação linear das funções bases

$$b_j(x) = (x - v)^j, \quad j = 0, 1, \dots, K.$$

O principal problema com esta base é que os polinômios não podem exibir muitas características locais sem fazer uso de um K grande. Mais ainda, os polinômios tendem a ajustar bem o centro dos dados, mas seu comportamento nas caudas não é muito bom. Pode-se dizer que as bases de

Fourier e as bases polinomiais têm sido muito utilizadas em trabalhos aplicados. Porém, a falta de capacidade delas em descrever características locais levou ao desenvolvimento dos polinômios *splines*. Tais funções serão descritas nas próximas seções.

1.2.2 Funções *Splines*

Uma alternativa para as bases polinomiais são *splines* polinomiais, que oferecem maior flexibilidade e têm a capacidade de capturar a mudança de comportamento locais. A fim de obter estas funções, primeiro o intervalo $[a, b]$ da função a ser estimada é particionado em k subintervalos $[\xi_{i-1}, \xi_i]$, $1 \leq i \leq k$, em que

$$a < \xi_0 < \dots < \xi_k < b.$$

Em seguida, um polinômio p_i é usado para aproximação em cada intervalo $[\xi_{i-1}, \xi_i]$, $i = 1, \dots, k$. Esse procedimento produz uma função de aproximação polinomial por partes $s(\cdot)$, ou seja, $s(x) = p_i(x)$ em $[\xi_{i-1}, \xi_i]$, $i = 1, \dots, k$. Os valores ξ_0, \dots, ξ_k são chamados de nós (*knots*), sendo ξ_0 e ξ_k os nós externos e os demais ξ_1, \dots, ξ_{k-1} os nós internos.

No caso geral, as partes de polinômio $p_i(x)$ são constituídas independentemente umas das outras e, portanto, não formam uma função contínua $s(x)$ em $[a, b]$. Isso não pode ser aceito se alguém deseja, particularmente, aproximar uma função suave. Portanto, é necessário que as partes do polinômio sejam unidas suavemente nos nós internos ξ_1, \dots, ξ_{k-1} e também que sejam deriváveis um certo número de vezes. Como resultado, obtém-se uma função polinomial por partes, suave, chamada função *spline*.

Um *spline* de ordem m (ordem=grau+1) com $k - 1$ nós internos em ξ_1, \dots, ξ_{k-1} é qualquer função da forma

$$s(x) = \sum_{i=0}^{m-1} c_i x^i + \sum_{j=1}^{k-1} d_j (x - \xi_j)_+^{m-1} \quad (1.2.2)$$

em que os coeficientes c_0, \dots, c_{m-1} e d_1, \dots, d_{k-1} são números reais e, dada uma função u , a função de potência truncada de grau r é definida como

$$u_+^r = \begin{cases} u^r, & \text{se } u \geq 0 \\ 0, & \text{se } u < 0 \end{cases}$$

Assim, pode-se concluir que qualquer função *spline* é uma combinação linear de $m + k$ funções base. De acordo com a equação (1.2.2), as funções base para um conjunto de nós interiores $\{\xi_1 \dots, \xi_{k-1}\}$ são $\{1, x, x^2, \dots, x^{m-1}, (x - \xi_1)_+^{m-1}, \dots, (x - \xi_{k-1})_+^{m-1}\}$.

Na próxima seção é apresentado um tipo de *spline* que tem a importante propriedade computacional de ter suporte compacto, ou seja, ele é não nulo (de fato positivo) num intervalo pequeno e zero fora desse intervalo.

1.2.3 Bases B-splines

Os B-splines são constituídos de pedaços de polinômios unidos em certos valores chamados nós. Para os nós ξ_0 , ξ_1 e ξ_2 um B-spline de grau 1 consiste de dois pedaços lineares, um pedaço de ξ_0 a ξ_1 , e outro de ξ_1 a ξ_2 . À esquerda de ξ_0 e à direita de ξ_2 esse B-spline é zero (ver Figura 1.1 (a)). Para construir um conjunto mais amplo de B-splines basta introduzir mais nós. Na Figura 1.1, em (b), tem-se todos os B-splines possíveis de grau 1 no intervalo $[0, 1]$ com os nós em $\{0; 0,25; 0,5; 0,75; 1\}$.

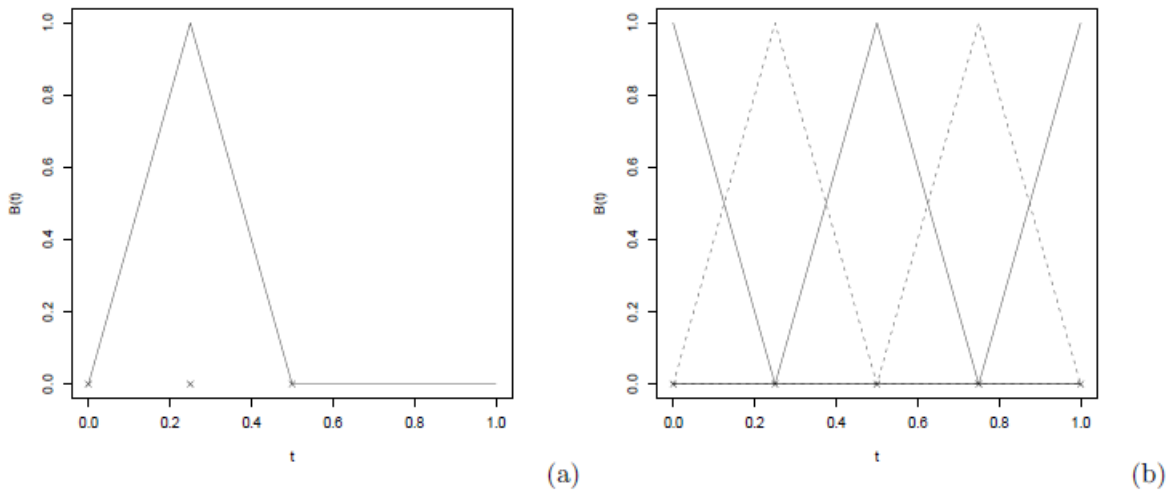


Figura 1.1: (a) B-spline de grau 1 isolado com 3 nós em “x”. (b) B-splines de grau 1 com nós posicionados em “x”

Por sua vez, a Figura 1.2 em (a) tem-se um B-spline cúbico (ordem=4) que consiste de quatro pedaços de polinômios cúbicos, unidos em três nós internos. Nos pontos de união não apenas as

ordens dos pedaços de polinômios se encaixam, também são iguais as suas primeiras e segundas derivadas (mas não as terceiras derivadas). Na Figura 1.2 em (b) tem-se todos os B-splines possíveis de grau 3 no intervalo $[0, 1]$, com nós em $\{0; 0, 2; 0, 4; 0, 6; 0, 8; 1\}$. Os B-splines cúbicos são amplamente utilizados em regressão não paramétrica, uma vez que uma combinação linear deles resulta em uma curva suave.

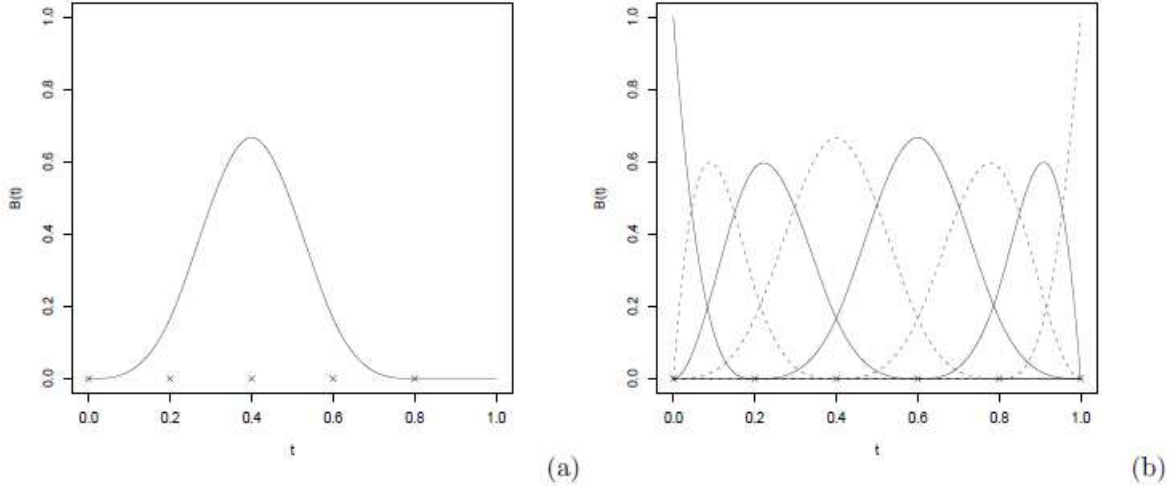


Figura 1.2: (a) B-spline cúbico isolado com 5 nós em “x”. (b) B-splines cúbico com nós posicionados em “x”

O i -ésimo B-spline de ordem m pode ser obtido recursivamente por

$$B_{i,m}(x) = \frac{x - \xi_i}{\xi_{i+m-1} - \xi_i} B_{i,m-1}(x) + \frac{\xi_{i+m} - x}{\xi_{i+m} - \xi_{i+1}} B_{i+1,m-1}(x)$$

sendo

$$B_{i,1}(x) = \begin{cases} 1, & \text{se } \xi_i \leq x \leq \xi_{i+1} \\ 0, & \text{caso contrário} \end{cases}$$

Assim como foi feito para os B-splines, uma relação de recorrência também pode ser utilizada para o cálculo de suas derivadas. Para mais detalhes sobre bases polinomiais, splines e B-splines vide Souza e Dias (2008). Durante todo este trabalho são consideradas funções base B-splines.

1.3 Métodos para Estimação de Curvas

Vários métodos são sugeridos para estimar de modo não paramétrico uma curva de regressão f usando *splines*. Alguns dos mais conhecidos são a suavização por *splines*, a regressão com *splines* e a regressão penalizada com *splines* (*smoothing splines*, *regression splines* e *penalized regression splines*, respectivamente). A seguir é feita uma breve apresentação destes métodos. Na sequência descreve-se o método *h-splines*, o qual deseja-se generalizar. Em todos os métodos a função f é aproximada por uma combinação linear de funções base, como descrito em (1.2.1). O que diferencia os métodos é a escolha do número de bases a serem usadas, K , além da maneira de estimar os coeficientes $\theta_1, \dots, \theta_K$.

1.3.1 Regressão com *Splines*

Na regressão com *splines*, os coeficientes $\theta_1, \dots, \theta_K$ são escolhidos de tal forma a minimizar

$$\sum_{i=1}^n [y_i - f_K(x_i)]^2.$$

O grau de suavização que será aplicado aos dados é determinado pelo número de funções base K . Uma escolha comum para as funções base é um conjunto de *B-splines* cúbicos. As principais dificuldades de se trabalhar com este método é a escolha das posições dos nós na construção das bases e a escolha do número de bases.

1.3.2 Suavização por *Splines*

Em suavização por *splines* os coeficientes $\theta_1, \dots, \theta_K$ são obtidos minimizando o critério

$$\sum_{i=1}^n [y_i - f_K(x_i)]^2 + \lambda \int [f_K''(t)]^2 dt. \quad (1.3.1)$$

Observe que os limites de integração em (1.3.1) foram omitidos. Sempre que isso ocorrer neste trabalho, considere uma integral sobre um intervalo que cobre toda a variável em questão. Sabe-se que a função \hat{f} que minimiza este critério é necessariamente um *spline* natural cúbico com nós em x_i , vide por exemplo, Green e Silverman (1994), Wahba (1981) e Craven e Wahba (1978). O primeiro termo mede a proximidade com os dados, enquanto o segundo penaliza a curvatura da

função. É comum definir a curvatura total da curva como sendo a integral da sua segunda derivada ao quadrado. Isso mede a não suavidade da curva. O parâmetro λ é chamado de parâmetro de suavização. Valores grandes de λ implicam em curvas estimadas mais suaves, enquanto valores pequenos implicam em curvas com mais curvatura. Observe que o número de coeficientes pode ser tão grande quanto o número de observações. Com isso, a computação se torna mais difícil do que no caso da regressão com *splines* em que estima-se K coeficientes.

1.3.3 Regressão Penalizada com *Splines*

O método de regressão penalizada com *splines* é uma combinação dos dois métodos anteriores. Neste método também utiliza-se o critério (1.3.1) na estimação de f , assim como na suavização por *splines*, mas o número de funções base não é tão grande quanto n . Trabalha-se com $K < n$, assim como na regressão com *splines*. Deste modo, aproveita-se o melhor de cada método: a quantidade reduzida de coeficientes da regressão com *splines* e o controle da suavização através de λ , característica da suavização por *splines*.

Muitos pacotes em R como `gamlss`, `mgcv` e `DPpackage` usam a regressão penalizada com *splines* para estimar f . Neles, o número de funções base K é predeterminado de acordo com o número de observações n ou pode ser determinado diretamente pelo usuário. O problema é que nem sempre o valor escolhido para K é o mais adequado. Para saber mais sobre os pacotes citados, vide Stasinopoulos e Rigby (2007), Wood (2001) e Jara et al. (2011).

1.3.4 H-*splines*

O método h-*splines* é semelhante à regressão penalizada com *splines*, com a diferença que o número de funções base K é uma quantidade a ser estimada pelo método.

Como já foi mencionado, a base desta tese são os trabalhos de Dias (1999) e Dias e Gamerman (2002), que tratam de regressão não paramétrica via h-*splines*.

Dias (1999) apresenta um procedimento sequencial adaptativo para obter o melhor número de bases e assim estimar f . Nesta abordagem inicia-se com o menor número de funções base possível ($K = 4$ para B-*splines* cúbicos) e acrescentam-se bases até que um critério de parada seja satisfeito.

Para cada K tem-se uma nova estimativa de λ . O critério é baseado na distância de Hellinger (afinidade) entre duas curvas estimadas consecutivas. Este método será apresentado com mais detalhes no Capítulo 4.

Dias e Gamerman (2002) apresentam uma abordagem bayesiana que usa o método MCMC com saltos reversíveis (RJ-MCMC) para possibilitar que várias estimativas de f sejam geradas usando diferentes valores de K atribuindo-lhe uma distribuição *a priori*, ou seja, as funções são estimadas com diferentes números de funções base. A generalização desta abordagem é dada no Capítulo 2. O procedimento RJ-MCMC construído é baseado no trabalho de Green (1995).

Observe que o método *h-splines* define o número de funções base adequado para ajustar um conjunto de dados, evitando valores muito altos ou muito baixos para K . Uma quantidade pequena de funções base pode não ser suficiente para capturar a relação estrutural entre a variável resposta Y e a covariável X . Por outro lado, um valor muito grande de K pode não trazer ganhos ao ajuste, e ainda aumentar o esforço computacional. Além disso, tanto a abordagem adaptativo-sequencial quanto a bayesiana aparentemente podem ser generalizadas para p covariáveis de modo direto. Por estes motivos escolheu-se o método *h-splines* para fazer as generalizações.

1.4 Generalização do Método H-*splines*

Para a generalização do método *h-splines* para o caso com várias covariáveis, considere modelos do tipo

$$Y = f(X_1, \dots, X_p) + \epsilon \quad (1.4.1)$$

sendo $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$ e o erro independente de X_1, \dots, X_p . Aqui, a generalização será feita considerando modelos aditivos, ou seja, será considerado que (1.4.1) pode ser expresso (ou aproximado) por

$$Y = \alpha + \sum_{j=1}^p f_j(X_j) + \epsilon,$$

sendo f_1, \dots, f_p funções arbitrárias univariadas, uma para cada preditor.

A estimação por modelos aditivos fornece uma aproximação simples e direta para a função f dada em (1.4.1). Esse é o principal motivo para usar este tipo de modelo. Observe que modelos

aditivos não consideram interações entre os preditores. Existem métodos de suavização baseados em *splines* que estimam modelos com interações entre as covariáveis, como por exemplo

$$Y = f_1(X_1) + f_2(X_2, X_4) + f_3(X_3, X_5, X_6) + \epsilon.$$

Estes métodos ganham em flexibilidade em relação aos modelos aditivos, mas perdem em uma série de pontos que são citados a seguir. Primeiro, a dificuldade em determinar quais interações utilizar na modelagem. Muitas vezes é inviável considerar todas as interações dois a dois, três a três, quatro a quatro, etc. Segundo, a construção de bases p -dimensionais envolvem produtos tensoriais, o que aumenta em muito o esforço computacional. Uma expansão de bases de dimensão dois, digamos para X_1 e X_2 , resultaria em uma matriz com n linhas e $K_1 \times K_2$ colunas. Para dimensão três seria uma matriz com n e o número de colunas seria $K_1 \times K_2 \times K_3$, e assim por diante. Terceiro, a interpretabilidade dos modelos aditivos é bem simples. Uma vez ajustado o modelo, pode-se examinar o efeito de cada preditor separadamente. Nos modelos aditivos isso pode ser feito diretamente, enquanto que para modelos com interação, seria necessário condicionar a covariável de interesse a todas as demais, o que pode ser muito trabalhoso para grandes dimensões. Quarto, para modelos aditivos, a generalização de métodos de estimação univariados é mais direta, uma vez que tratamos as covariáveis separadamente. Embora os modelos aditivos sejam mais simples, o modelo é quase sempre uma aproximação da verdadeira superfície de regressão, mas espera-se que seja uma aproximação útil. Quando ajustamos um modelo de regressão linear, geralmente não acreditamos que este modelo seja correto. Acreditamos que será uma boa aproximação de primeira ordem para a superfície verdadeira, e que podemos descobrir os preditores importantes e seus efeitos sobre a variável resposta. Considere os modelos aditivos como sendo uma aproximação mais geral.

1.5 Organização do Trabalho

A generalização do método *h-splines* para as abordagens bayesiana e sequencial é apresentada nos Capítulos 2 e 4, respectivamente. Em estudos de simulação os resultados obtidos pelo método sequencial são comparados aos resultados dos pacotes `gamlss` e `mgcv` em R, que utilizam o método

de regressão penalizada com *splines*. No contexto bayesiano, os resultados do método proposto são confrontados com resultados do pacote **DPpackage** em R, que utiliza uma versão bayesiana para regressão penalizada com *splines*.

Os Capítulos 3 e 5 apresentam testes de hipóteses baseados nas abordagens bayesiana e sequencial, respectivamente. O teste bayesiano é baseado no teste de significância totalmente bayesiano (FBST) de Pereira e Stern (1999), enquanto o teste referente à abordagem sequencial é baseado na distância quadrática integrada, que é uma medida de distância entre curvas.

O Capítulo 6 propõe uma abordagem para tratar variáveis resposta com restrições, como por exemplo variáveis binárias ou contagens. Hastie e Tibshirani (1990) propõem a classe de modelos aditivos generalizados (GAM) para tratar variáveis da família exponencial de distribuições. No Capítulo 6 é proposta uma abordagem para modelar este tipo de dados de modo não paramétrico, denominada modelos aditivos generalizados livres de distribuição (DFGAM). O modelo DFGAM não assume uma distribuição específica para a variável resposta. Serão realizadas algumas simulações para comparar as funções estimadas para o modelo DFGAM com as estimativas obtidas pelos pacotes em R citados anteriormente, que consideram o modelo GAM.

O Capítulo 7 descreve um estudo de caso que trata de um conjunto de dados referente ao trabalho de Tanaka e Nishii (2009) sobre o desmatamento no leste da Ásia. No trabalho os autores consideram como principais causas de desmatamento em uma área (variável dependente) a população humana no local (preditor 1) e o declive do terreno (preditor 2). Vários modelos são considerados pelos autores. Através dos procedimentos apresentados neste trabalho, obteremos as estimativas não paramétricas e selecionaremos o modelo mais indicado.

Capítulo 2

Estimação Bayesiana para Regressão Múltipla H-*Splines*

Neste capítulo, é apresentada uma generalização do método *h-splines* no contexto bayesiano. Inicialmente é feita uma rápida revisão de trabalhos que tratam de modelos aditivos usando uma abordagem bayesiana. Em seguida é apresentado um procedimento RJ-MCMC para modelos aditivos, baseado no método *h-splines*. Por último é feito um estudo de simulação que compara os resultados obtidos pelo procedimento proposto com os conseguidos pelo pacote `DPpackage` em R.

2.1 Introdução

Em seu trabalho, Dias e Gamerman (2002) comentam que desde Craven e Wahba (1978), vários métodos são sugeridos para estimar, de modo não paramétrico, uma curva de regressão f usando *splines*. Kimeldorf e Wahba (1970) e Wahba (1983) deram uma atrativa interpretação bayesiana a uma \hat{f} estimada de uma curva desconhecida f . Eles mostraram que \hat{f} pode ser vista como uma estimativa de Bayes para f com respeito a uma certa distribuição *a priori* sobre a classe de todas as funções suaves. Segundo Wahba (1983) a abordagem bayesiana permite estimar não somente a função desconhecida, como também os limites de confiança bayesianos.

Uma abordagem não paramétrica para modelos aditivos foi proposta por Hastie e Tibshirani (1990). Nela, as f_j são funções univariadas que podem ser estimadas usando o algoritmo de

retroajustamento (*backfitting*) proposto pelos autores. DiMatteo, Genovese e Kass (2001) criaram um procedimento MCMC de saltos reversíveis no contexto univariado, que enfatiza a adição e remoção de nós para a construção dos *splines*. Denison, Mallick e Smith (1998) apresentam uma abordagem semelhante, e ainda estendem para o caso com várias covariáveis, usando modelos aditivos. Porém, Denison, Mallick e Smith (1998) e DiMatteo, Genovese e Kass (2001) construíram o método MCMC com saltos reversíveis sem considerar a penalização à não suavidade. Dias e Gamerman (2002) desenvolveram um procedimento de saltos reversíveis usando *h-splines*, ou seja, além de controlar a curvatura selecionando o melhor número de funções base, ainda penalizavam a não suavidade através do parâmetro λ .

Neste capítulo estendemos o trabalho de Dias e Gamerman (2002) para o caso de várias covariáveis. Neste contexto foram utilizados modelos aditivos.

2.2 Regressão Penalizada para Modelo Aditivos

Para estabelecer a notação que será usada no trabalho, considere um conjunto de dados (\mathbf{y}, \mathbf{x}) , com $\mathbf{x} = (\mathbf{x}_1^t, \dots, \mathbf{x}_p^t)^t$ em que $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^t$ são as observações referentes à covariável X_j para $j = 1, \dots, p$ e $\mathbf{y} = (y_1, \dots, y_n)^t$ é o vetor de observações da variável resposta. Para este conjunto de dados, o modelo de regressão dado por (1.1.1) pode ser escrito como

$$\mathbf{y} = f(\mathbf{x}_1, \dots, \mathbf{x}_p) + \boldsymbol{\epsilon}, \quad (2.2.1)$$

sendo $f(\mathbf{x}) = [f(x_{11}, \dots, x_{1p}), \dots, f(x_{n1}, \dots, x_{np})]^t$ um vetor de tamanho n , $\boldsymbol{\epsilon}$ um vetor de erros com vetor de médias igual a $\mathbf{0}_n$ e matriz de covariâncias dada por $\sigma^2 \mathbf{I}_n$, em que \mathbf{I}_n é uma matriz identidade $n \times n$ e $\mathbf{0}_n$ é um vetor nulo de tamanho n . Como mencionado anteriormente, adotaremos o modelo aditivo. Com isso, para $i = 1, \dots, n$,

$$f(x_{i1}, \dots, x_{ip}) = \alpha + \sum_{j=1}^p f_j(x_{ij}),$$

com $f_j(x_{ij}) = \sum_{l=1}^{K_j} \theta_{lj} b_l(x_{ij})$, sendo b_l 's bases *B-splines* cúbicas conhecidas. Assim, reescrevendo (2.2.1) como modelo aditivo tem-se

$$\mathbf{y} = \boldsymbol{\alpha} + \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\theta}_j + \boldsymbol{\epsilon}, \quad (2.2.2)$$

em que $\boldsymbol{\alpha} = \alpha \mathbf{1}_n$, com $\mathbf{1}_n$ sendo um vetor $n \times 1$ com todas as entradas iguais a 1, \mathbf{X}_j é uma matrix $n \times K_j$ contendo os valores das K_j funções base calculadas em \mathbf{x}_j e $\boldsymbol{\theta}_j = (\theta_{j1}, \dots, \theta_{jK_j})^t$ para $j = 1, \dots, p$.

Na abordagem usual de regressão penalizada com *splines* para modelos aditivos as estimativas são obtidas encontrando α, f_1, \dots, f_p que minimizem

$$\sum_{i=1}^n \left[y_i - \alpha - \sum_{j=1}^p f_j(x_{ij}) \right]^2 + \sum_{j=1}^p \lambda_j \int [f_j''(t)]^2 dt,$$

ou em notação matricial, obter $\boldsymbol{\theta}$ que minimize

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \sum_{j=1}^p \lambda_j \boldsymbol{\theta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\theta}_j, \quad (2.2.3)$$

em que $\boldsymbol{\theta} = (\alpha, \boldsymbol{\theta}_1^t, \dots, \boldsymbol{\theta}_p^t)^t$, $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_1, \dots, \mathbf{X}_p)$ é uma matriz $n \times (1 + K_1 + \dots + K_p)$ e $\boldsymbol{\Omega}_j$ tem dimensão $K_j \times K_j$. A demonstração de que $\int [f_j''(t)]^2 dt$ pode ser escrito na forma matricial como $\boldsymbol{\theta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\theta}_j$ é dada no Apêndice A. Entretanto, é simples entender que a matriz $\boldsymbol{\Omega}_j$ carrega informação sobre a curvatura de f_j .

A soma de quadrados penalizada dada em (2.2.3) pode ser simplificada para a forma

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \boldsymbol{\theta}^t \boldsymbol{\Lambda} \boldsymbol{\theta},$$

em que $\boldsymbol{\Lambda} = \text{diag}(0, \lambda_1 \boldsymbol{\Omega}_1, \dots, \lambda_p \boldsymbol{\Omega}_p)$ é uma matriz diagonal em blocos.

Geralmente, o modelo dado por (2.2.2) é não identificável, a menos que cada \mathbf{f}_j seja sujeito a uma restrição de centralidade. O modo como esta restrição é inserida na estimação do modelo pode ser visto no Apêndice B.

No método da regressão penalizada com *splines* os valores K_1, \dots, K_p são fixados. A seguir é apresentada uma abordagem bayesiana para o método *h-splines*, em que estes valores não são fixos.

2.3 Posicionamento dos Nós

Antes de descrever o procedimento RJ-MCMC, deve-se discutir um fator importante que é a posição dos nós. Denison, Mallick e Smith (1998) e DiMatteo, Genovese e Kass (2001) utilizaram

diferentes metodologias para alocar os nós. O método adotado aqui é semelhante ao de Denison, Mallick e Smith (1998).

Suponha que a função f_j , referente à covariável X_j , está sendo estimada usando K_j funções bases. Se usarmos B-splines cúbicos, teremos $K_j - 2$ nós $\xi_1, \dots, \xi_{K_j-2}$. Os nós das extremidades são fixos, sendo $\xi_1 = \min(\mathbf{x}_j)$ e $\xi_{K_j-2} = \max(\mathbf{x}_j)$. A ideia é escolher os demais $K_j - 4$ nós (nós internos) usando diretamente os dados, ou seja, as estatísticas de ordem. Portanto, os nós restantes devem ser valores sorteados entre os X_j observados. Deste modo teremos nós somente sobre os dados.

Nos passos de nascimento e morte do procedimento RJ-MCMC devemos adicionar ou remover nós internos. No passo nascimento, escolhe-se um dos dados para se tornar o novo nó. Como são n observações e $K_j - 2$ nós já escolhidos, deve-se escolher uma das $n - (K_j - 2)$ observações disponíveis de maneira uniforme. No passo morte, um dos $K_j - 4$ nós internos é removido, também escolhido de modo uniforme. Obviamente, tudo isso deve ser feito garantindo que os nós estejam em ordem crescente.

Após adicionar (ou remover) um nó tem-se a adição (ou remoção) de uma função base, e consequentemente, a adição (ou remoção) de um coeficiente em $\boldsymbol{\theta}_j$. Em que posição, em $\boldsymbol{\theta}_j$, entrará (ou sairá) o coeficiente? A resposta depende da posição do nó adicionado (ou removido). Se o nó interno que foi adicionado (removido) está na posição ω , então o coeficiente a ser adicionado (removido) será o da posição $\omega + 1$. Exceto se a posição do nó for a segunda ou penúltima, casos em que o coeficiente a ser adicionado ou removido será o segundo ou penúltimo, respectivamente.

Vale lembrar que para B-splines cúbicos, cada nó interno é utilizado na construção de cinco bases. Portanto, na adição ou remoção de um nó, cinco bases são afetadas, e consequentemente cinco coeficientes. Desse modo, não apenas o coeficiente da posição $\omega + 1$ é afetado na adição ou remoção. A proposta do parágrafo anterior, que determina qual coeficiente deve ser incluído ou excluído é baseada em resultados empíricos vindos de simulações, que indicaram que o coeficiente da posição $\omega + 1$ é o mais afetado.

De posse dessas informações passamos à construção do procedimento RJ-MCMC.

2.4 Abordagem Bayesiana para o Método H-splines

Para a abordagem bayesiana, considere o modelo de regressão dado por (2.2.1), porém aqui $\epsilon \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$.

A função verossimilhança para o vetor de parâmetros $(\boldsymbol{\theta}^t, \mathbf{K}^t, \sigma^2)$ com $\mathbf{K} = (K_1, \dots, K_p)^t$ é

$$L(\boldsymbol{\theta}, \mathbf{K}, \sigma^2 | \mathbf{y}, \mathbf{x}) \propto \sigma^{-n/2} \exp \left[-\frac{Q(\boldsymbol{\theta})}{2\sigma^2} \right]$$

com $Q(\boldsymbol{\theta}) = \left\| \mathbf{y} - \boldsymbol{\alpha} - \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\theta}_j \right\|^2$.

A abordagem bayesiana usual para um modelo de regressão não paramétrica é considerar uma distribuição *a priori* para f (vide Green e Silverman (1994)) como

$$p(f_j) \propto \exp \left\{ -\frac{\lambda_j}{2} \int [f_j''(t)]^2 dt \right\} \quad j = 1, \dots, p.$$

Esta é a forma encontrada na abordagem bayesiana para justificar uma penalização à função verossimilhança.

A distribuição *a priori* escolhida para o intercepto é uma normal, ou seja,

$$\alpha \sim N(0, \tau^{-1}).$$

Similar a Dias e Gamerman (2002), considere as seguintes distribuições *a priori* condicionais independentes

$$p(\boldsymbol{\theta}_j | K_j, \lambda_j) \propto \exp \left(-\frac{\lambda_j}{2} \boldsymbol{\theta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\theta}_j \right),$$

para $j = 1, \dots, p$. Ou seja, estamos assumindo distribuições *a priori* independentes tais como

$$(\boldsymbol{\theta}_j | K_j, \lambda_j) \sim N(\mathbf{0}_{K_j}, \lambda_j^{-1} \boldsymbol{\Omega}_j^-),$$

em que $\boldsymbol{\Omega}_j^-$ é a inversa generalizada de $\boldsymbol{\Omega}_j$ e a distribuição *a priori* conjunta dada por

$$p(\boldsymbol{\theta}, \mathbf{K}, \boldsymbol{\lambda}) = p(\alpha) \times \prod_{j=1}^p p(\boldsymbol{\theta}_j | \lambda_j, K_j) p(\lambda_j | K_j) p(K_j)$$

sendo $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^t$, com

$$p(\lambda_j | K_j) = \psi(K_j) \exp[-\lambda_j \psi(K_j)],$$

Dias e Gamerman (2002) sugerem $\psi(K) = K^{-b} \exp(-cK)$. Além disso,

$$p(K_j) = \frac{\exp(-a_j) a_j^{K_j} / K_j!}{1 - \exp(-a_j)(1 - q^*)}, \quad K_j = 1, \dots, K_j^*,$$

sendo K_j^* o número máximo de funções base e $q^* = \sum_{l=K_j^*+1}^{\infty} a_j^l / l!$. Para σ^2 tem-se uma distribuição gama inversa $IG(u, v)$.

A distribuição *a posteriori* é escrita como

$$\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{K}, \sigma^2 | \mathbf{y}, \mathbf{x}) \propto L(\boldsymbol{\theta}, \mathbf{K}, \sigma^2 | \mathbf{y}, \mathbf{x}) \times p(\sigma^2) \times p(\alpha) \times \prod_{j=1}^p p(\boldsymbol{\theta}_j | \lambda_j, K_j) p(\lambda_j | K_j) p(K_j).$$

Como a dimensão de $\boldsymbol{\theta}_j$ varia com o valor de K_j , é proposto a seguir um algoritmo MCMC de saltos reversíveis para amostrar da distribuição *a posteriori*.

1. Amostrando σ^2 :

Note que a distribuição condicional *a posteriori* é

$$\begin{aligned} p(\sigma^2 | \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{K}, \mathbf{y}, \mathbf{x}) &\propto L(\boldsymbol{\theta}, \mathbf{K}, \sigma^2 | \mathbf{y}, \mathbf{x}) p(\sigma^2) \\ &= (\sigma^2)^{-n/2} \exp \left[\frac{-Q(\boldsymbol{\theta})}{2\sigma^2} \right] (\sigma^2)^{-(u+1)} \exp \left(-\frac{v}{\sigma^2} \right). \end{aligned}$$

Consequentemente,

$$(\sigma^2 | \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{K}, \mathbf{y}, \mathbf{x}) \sim IG \left(u + \frac{n}{2}, v + \frac{Q(\boldsymbol{\theta})}{2} \right).$$

2. Amostrando $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^t$:

$$\begin{aligned} p(\boldsymbol{\lambda} | \boldsymbol{\theta}, \mathbf{K}, \sigma^2, \mathbf{y}, \mathbf{x}) &\propto \prod_{j=1}^p p(\boldsymbol{\theta}_j | \lambda_j, K_j) p(\lambda_j | K_j) \\ &= \prod_{j=1}^p \lambda_j^{K_j/2} \exp \left(-\frac{\boldsymbol{\theta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\theta}_j}{2} \lambda_j \right) \exp [-\psi(K_j) \lambda_j] \end{aligned}$$

Assim, a distribuição de $(\boldsymbol{\lambda} | \boldsymbol{\theta}, \mathbf{K}, \sigma^2, \mathbf{y}, \mathbf{x})$ é o produto de p densidades de variáveis gama independentes, ou seja, para $j = 1, \dots, p$

$$(\lambda_j | \boldsymbol{\theta}_j, K_j, \sigma^2, \mathbf{y}, \mathbf{x}) \sim G \left(\frac{K_j}{2} + 1, \psi(K_j) + \frac{\boldsymbol{\theta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\theta}_j}{2} \right).$$

3. Amostrando $\boldsymbol{\theta} = (\alpha, \boldsymbol{\theta}_1^t, \dots, \boldsymbol{\theta}_p^t)^t$:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \mathbf{K}, \sigma^2, \mathbf{y}, \mathbf{x}) &\propto L(\boldsymbol{\theta}, \mathbf{K}, \sigma^2|\mathbf{y}, \mathbf{x})p(\sigma^2)p(\alpha)\prod_{j=1}^p p(\boldsymbol{\theta}_j|\lambda_j, K_j) \\ &= \exp\left[-\frac{Q(\boldsymbol{\theta})}{2\sigma^2}\right]\exp\left(-\frac{\tau}{2}\alpha^2\right)\exp\left(-\frac{1}{2}\sum_{j=1}^p \lambda_j \boldsymbol{\theta}_j^t \boldsymbol{\Omega}_j \boldsymbol{\theta}_j\right) \\ &= \exp\left[-\frac{Q(\boldsymbol{\theta})}{2\sigma^2}\right]\exp\left(-\frac{\boldsymbol{\theta}^t \boldsymbol{\Lambda} \boldsymbol{\theta}}{2}\right), \end{aligned}$$

sendo $\boldsymbol{\Lambda} = \text{diag}(\tau, \lambda_1 \boldsymbol{\Omega}_1, \dots, \lambda_p \boldsymbol{\Omega}_p)$ com $\boldsymbol{\Omega}_j$ de dimensão $K_j \times K_j$. Dessa maneira,

$$(\boldsymbol{\theta}|\boldsymbol{\lambda}, \mathbf{K}, \sigma^2, \mathbf{y}, \mathbf{x}) \sim N\left(\left(\frac{\mathbf{X}^t \mathbf{X}}{\sigma^2} + \boldsymbol{\Lambda}\right)^{-1} \frac{\mathbf{X}^t \mathbf{y}}{\sigma^2}, \left(\frac{\mathbf{X}^t \mathbf{X}}{\sigma^2} + \boldsymbol{\Lambda}\right)^{-1}\right),$$

em que $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_p)$ é uma matriz $n \times (1 + K_1 + \dots + K_p)$.

4. Amostrando \mathbf{K} (nascimento/morte):

Neste passo as amostragens são feitas separadamente para cada $j = 1, \dots, p$. Assim, será amostrado $(\boldsymbol{\theta}_j, K_j)$ e o índice j , em alguns momentos, será removido por uma questão de notação. Defina

$$\begin{aligned} b_K &= P(K \mapsto K+1) = c \min\left\{1, \frac{p(K+1)}{p(K)}\right\}, \\ d_K &= P(K \mapsto K-1) = c \min\left\{1, \frac{p(K-1)}{p(K)}\right\} \text{ e} \\ \eta_K &= P(K \mapsto K) = 1 - (b_K + d_K) \end{aligned}$$

como sendo, respectivamente, as probabilidades de nascimento, morte e de permanência no mesmo estado na cadeia. Baseados em simulações, Dias e Gamerman (2002) sugerem $c = 0,4$. A próxima etapa é, usando b_K , d_K e η_K , escolher um entre os movimentos. Para isso deve ser gerada uma distribuição trinomial $\text{trinomial}(1, b_K, d_K, \eta_K)$.

4.1 Movimento $K \mapsto K+1$:

Defina $\boldsymbol{\theta}_j^{K_j}$ como sendo o vetor $\boldsymbol{\theta}_j$ $K_j \times 1$ para $j = 1, \dots, p$.

Proposta para $\boldsymbol{\theta}_j^{K_j+1} = \left(\boldsymbol{\theta}_{j,(1:\omega)}^{K_j}, \theta^*, \boldsymbol{\theta}_{j,(\omega+1:K_j)}^{K_j}\right)^t$:

(a) Escolha ao acaso um valor de \mathbf{x}_j entre os $n - (K_j - 2)$ disponíveis para ser seu novo nó;

(b) Verifique a posição ω que este irá assumir e crie $K_j + 1$ funções base para X_j , obtendo \mathbf{X}_j^{novo} ;

(c) Amostre $\theta^* \sim N(\mu, \lambda_j^{-1} \mathbf{\Omega}_{j,\omega+1,\omega+1}^-)$, em que $\mathbf{\Omega}_{j,\omega+1,\omega+1}$ é o elemento $(\omega + 1, \omega + 1)$ da matriz $\mathbf{\Omega}_j$ de dimensão $(K_j + 1) \times (K_j + 1)$. O valor de μ será discutido mais adiante.

O valor proposto $(\boldsymbol{\theta}_j^{K_j+1}, K_j + 1)$ será aceito com probabilidade

$$\min \left\{ 1, \frac{\pi(\boldsymbol{\theta}_j^{K_j+1}, \lambda_j, K_j + 1, \sigma^2)}{\pi(\boldsymbol{\theta}_j^{K_j}, \lambda_j, K_j, \sigma^2)} \frac{q[(\boldsymbol{\theta}_j^{K_j+1}, K_j + 1) \mapsto (\boldsymbol{\theta}_j^{K_j}, K_j)]}{q[(\boldsymbol{\theta}_j^{K_j}, K_j) \mapsto (\boldsymbol{\theta}_j^{K_j+1}, K_j + 1)]} J \right\}, \quad (2.4.1)$$

em que $J = \left| \frac{\partial \boldsymbol{\theta}_j^{K_j+1}}{\partial (\boldsymbol{\theta}_{j,(1:\omega)}^{K_j}, \theta^*, \boldsymbol{\theta}_{j,(\omega+1:K_j)}^{K_j})^t} \right| = 1$. Observe que

$$\begin{aligned} q[(\boldsymbol{\theta}_j^{K_j}, K_j) \mapsto (\boldsymbol{\theta}_j^{K_j+1}, K_j + 1)] &= \frac{b_K}{n - (K_j - 2)} \frac{\sqrt{\lambda_j \mathbf{\Omega}_{\omega+1,\omega+1}}}{\sqrt{2\pi}} \\ &\times \exp \left[-\frac{\lambda_j \mathbf{\Omega}_{j,\omega+1,\omega+1}}{2} (\theta^* - \mu)^2 \right] \end{aligned}$$

e

$$q[(\boldsymbol{\theta}_j^{K_j+1}, K_j + 1) \mapsto (\boldsymbol{\theta}_j^{K_j}, K_j)] = \frac{d_{K+1}}{(K_j + 1) - 4}.$$

Note que

$$\begin{aligned} \frac{\pi(\boldsymbol{\theta}_j^{K_j+1}, \lambda_j, K_j + 1, \sigma^2)}{\pi(\boldsymbol{\theta}_j^{K_j}, \lambda_j, K_j, \sigma^2)} &= \frac{p(\mathbf{y}|\boldsymbol{\theta}_j^{K_j+1}, K_j + 1, \sigma^2)}{p(\mathbf{y}|\boldsymbol{\theta}_j^{K_j}, K_j, \sigma^2)} \frac{p(\boldsymbol{\theta}_j^{K_j+1}|K_j + 1, \lambda_j)}{p(\boldsymbol{\theta}_j^{K_j}|K_j, \lambda_j)} \\ &\times \frac{p(\lambda_j|K_j + 1)}{p(\lambda_j|K_j)} \frac{p(K_j + 1)}{p(K_j)}, \\ \frac{p(\mathbf{y}|\boldsymbol{\theta}_j^{K_j+1}, K_j + 1, \sigma^2)}{p(\mathbf{y}|\boldsymbol{\theta}_j^{K_j}, K_j, \sigma^2)} &= \exp \left[\frac{1}{2\sigma^2} \left(\|\mathbf{y}^{(j)} - \mathbf{X}_j \boldsymbol{\theta}_j^{K_j}\|^2 - \|\mathbf{y}^{(j)} - \mathbf{X}_j^{novo} \boldsymbol{\theta}_j^{K_j+1}\|^2 \right) \right], \\ \frac{p(\boldsymbol{\theta}_j^{K_j+1}|K_j + 1, \lambda_j)}{p(\boldsymbol{\theta}_j^{K_j}|K_j, \lambda_j)} &= \exp \left\{ \frac{\lambda_j}{2} \left[(\boldsymbol{\theta}_j^{K_j})^t \Lambda_j \boldsymbol{\theta}_j^{K_j} - (\boldsymbol{\theta}_j^{K_j+1})^t \Lambda_j \boldsymbol{\theta}_j^{K_j+1} \right] \right\}, \\ \frac{p(\lambda_j|K_j + 1)}{p(\lambda_j|K_j)} &= \frac{\psi(K_j + 1)}{\psi(K_j)} \exp \{ \lambda_j [\psi(K_j) - \psi(K_j + 1)] \} \text{ e} \\ \frac{p(K_j + 1)}{p(K_j)} &= \frac{a_j}{K_j + 1} \end{aligned}$$

sendo $\mathbf{y}^{(j)} = \mathbf{y} - \boldsymbol{\alpha} - \sum_{l \neq j} \mathbf{X}_l \boldsymbol{\theta}_l$.

4.2 Movimento $K \mapsto K - 1$:

Proposta para $\boldsymbol{\theta}_j^{K_j-1}$:

- (a) Escolha ao acaso um dos $K_j - 4$ nós internos para ser removido e verifique sua posição ω ;
- (b) Com os nós restantes, crie $K_j - 1$ funções base para X_j e proponha $\boldsymbol{\theta}_j^{K_j-1} = \left(\boldsymbol{\theta}_{j,(1:\omega)}^{K_j}, \boldsymbol{\theta}_{j,(\omega+2:K_j)}^{K_j} \right)^t$.

A probabilidade de aceitação de $K \mapsto K - 1$ é similar à expressão (2.4.1), porém com as razões invertidas.

4.3 Movimento $K \mapsto K$:

Mantenha $\boldsymbol{\theta}_j$ amostrado no passo 3.

5. Retorne ao passo 1 para amostrar um novo valor da distribuição *a posteriori*.

No passo nascimento da cadeia, o candidato a novo coeficiente em $\boldsymbol{\theta}_j$ é obtido fazendo $\theta^* \sim N(\mu, \lambda_j^{-1} \Omega_{j,\omega+1,\omega+1}^-)$. O valor escolhido para μ pode variar. Em simulações, inicialmente adotou-se $\mu = 0$, mas em certas funções simuladas o número de nascimentos aceitos foi muito reduzido e o número de funções bases acabava sendo abaixo do necessário. Dessa forma, decidiu-se adotar

$$\mu = \arg \min_{x \in R} \left\| \mathbf{y} - \boldsymbol{\alpha} - \mathbf{X}_j^{\text{nov}} \boldsymbol{\theta}_j^{K_j+1} - \sum_{l \neq j} \mathbf{X}_l \boldsymbol{\theta}_l \right\|^2, \quad (2.4.2)$$

sendo $\boldsymbol{\theta}_j^{K_j+1} = \left(\boldsymbol{\theta}_{j,(1:\omega)}^{K_j}, x, \boldsymbol{\theta}_{j,(\omega+1:K_j)}^{K_j} \right)^t$. Como todos os outros coeficientes são determinados, definimos μ como sendo o valor x que, condicionado aos demais, minimiza a soma dos quadrados dos resíduos. Nas simulações, esta escolha para μ trouxe ganhos. O número de nascimentos aceitos cresceu consideravelmente, equilibrando-se com o de mortes. Por esse motivo, recomendamos μ dado por (2.4.2), mas outros valores podem ser usados.

O algoritmo apresentado gera uma amostra de curvas *a posteriori*. Graças à habilidade do procedimento MCMC de saltos reversíveis de saltar por diferentes dimensões, as diversas curvas são geradas utilizando diferentes valores para $\mathbf{K} = (K_1, \dots, K_p)^t$, satisfazendo o propósito do método *h-splines*, que é encontrar a melhor combinação entre \mathbf{K} e $\boldsymbol{\lambda}$.

Para obter a estimativa final da superfície f lembre que, para cada covariável X_j , tem-se uma amostra *a posteriori* de curvas que estimam f_j , digamos $\hat{f}_j^{(1)}, \hat{f}_j^{(2)}, \dots, \hat{f}_j^{(m)}$. Deste modo, a estimativa para f_j é a média *a posteriori*, dada por

$$\hat{f}_j = \frac{1}{m} \sum_{l=1}^m \hat{f}_j^{(l)}.$$

Então, a estimativa para f será

$$\hat{f} = \hat{\alpha} + \sum_{j=1}^p \hat{f}_j, \quad (2.4.3)$$

sendo $\hat{\alpha}$ a média *a posteriori* de α . Neste trabalho, denominaremos intervalos ou limites bayesianos os intervalos empíricos obtidos a partir da amostra de superfícies *a posteriori*. Neste trabalho são usados os quantis empíricos *a posteriori* de 2,5% e 97,5%, que são denotados respectivamente por LI 2,5% e LS 97,5%. Estes limites bayesianos podem se referir tanto às curvas quanto aos parâmetros.

Com relação à convergência do método, Dias e Gamerman (2002) sugerem um descarte de 1000 amostras de curvas para funções com pouca curvatura e um descarte de 5000 curvas para funções mais estruturadas (grande curvatura). Nos estudos de simulação deste trabalho são geradas 5000 amostras de curvas através do algoritmo, sendo as 4000 primeiras descartadas e as 1000 seguintes são consideradas como vindas da distribuição *a posteriori*.

O procedimento RJ-MCMC para estimação de modelos aditivos proposto nesta seção será referido como método *h-splines* bayesiano (BHS).

2.5 Dados Simulados

O algoritmo foi desenhado para contemplar qualquer dimensão finita, mas foi aplicado em algumas funções com duas covariáveis. O uso de apenas duas covariáveis se dá para que as superfícies estimadas possam ser comparadas visualmente com as verdadeiras. As funções usadas para testar o algoritmo em cada exemplo são as seguintes:

- Exemplo 1: $f_1(x) = \sin(x^2 - x + 2)$ e $f_2(x) = \exp(-2x^2)$,
- Exemplo 2: $f_1(x) = \exp(-0,5x^2) \cos(4\pi x)$ e $f_2(x) = x^3$ e

- Exemplo 3: $f(x_1, x_2) = (x_1^2 + 3x_2^2) \exp(-x_1^2 - x_2^2)$.

O exemplo 1 é um caso inicial para verificar o desempenho do algoritmo proposto. No exemplo 2 a função f_1 é uma função mais estruturada, com uma maior curvatura total, o que em geral dificulta a estimação. Além disso, as covariáveis neste exemplo são correlacionadas. No exemplo 3 temos um modelo com interação entre as covariáveis.

Em todos os exemplos, os hiperparâmetros τ , u e v são escolhidos de tal forma que tenhamos distribuições *a priori* vagas, com variâncias extremamente grandes. As distribuições *a priori* para K_1 e K_2 são distribuições Poisson truncadas em 40, com média igual a 15. Além disso, os valores iniciais dos parâmetros foram $K_1 = K_2 = 15$, $\lambda_1 = \lambda_2 = 0,01$, $\boldsymbol{\theta} = (\mathbf{X}^t \mathbf{X} + \boldsymbol{\Lambda})^{-1} \mathbf{X}^t \mathbf{y}$ e $\sigma^2 = n^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$. Escolheu-se ainda $\psi_1 = K_1$ e $\psi_2 = K_2$. O algoritmo descrito foi implementado em R.

Os resultados do método BHS são comparados com os resultados do pacote **DPpackage** em R. Esse pacote utiliza uma abordagem bayesiana para regressão com *splines* que usa métodos MCMC para estimar modelos aditivos. O método referente ao pacote **DPpackage** será denominado DP. As distribuições *a priori* usadas para o método DP também são vagas, similares às utilizadas no método BHS.

2.5.1 Exemplo 1

A geração dos dados segue os seguintes passos:

1. Gere $X_1 \sim U(0, 4)$ e $X_2 \sim U(-1, 3)$;
2. Calcule $f_1(X_1) = \sin(X_1^2 - X_1 + 2)$, $f_2(X_2) = \exp(-2X_2^2)$ e $y = f_1(X_1) + f_2(X_2) + \epsilon$, com $\epsilon \sim N(0; 0,05)$;
3. Repita os passos (1) e (2) até obter $n = 50$ observações.

Com os dados em mãos aplica-se o procedimento BHS gerando uma amostra *a posteriori* de 1000 curvas. A Figura 2.1 mostra, em cada linha, uma amostra de curvas *a posteriori* para f_j (relativas às X_j) e em seguida a curva média \hat{f}_j juntamente com a estimativa de DP e a curva f_j verdadeira.

A estimativas \hat{f}_j obtidas pelo método proposto estão muito próximas às verdadeiras funções f_j . O mesmo vale para as estimativas obtidas por DP.

Vale lembrar que nessa amostra *a posteriori* existem curvas com diferentes quantidades de funções base, ou seja, f_j é estimada usando vários valores para K_j . E mesmo que duas curvas sejam amostradas com um mesmo valor de K_j , elas podem ter diferentes nós. As modas *a posteriori* para K_1 e K_2 foram respectivamente 18 e 16. Informações sobre a distribuição *a posteriori* de λ_1 , λ_2 , σ^2 e α são dadas na Tabela 2.1.

Tabela 2.1: Médias *a posteriori* e limites bayesianos para λ_1 , λ_2 , σ^2 e α estimados para o exemplo 1.

Parâmetro	LI 2,5%	Média	LS 97,5%
α	0,5382	0,5510	0,5635
σ^2	0,0011	0,0020	0,0034
λ_1	0,0098	0,0223	0,0392
λ_2	0,0920	0,2705	0,5520

Observe que f_1 necessitou de um número um pouco maior de funções bases, isso reflete o fato de f_1 ser uma função que, em geral, possui maior curvatura que f_2 , como esperado quanto maior a curvatura mais funções base são usadas. O comportamento de λ_1 e λ_2 dados na Tabela 2.1 também pode ser explicado pela curvatura. Espera-se que quanto maior a curvatura, menor o valor do parâmetro de suavização.

A estimativa da superfície é obtida fazendo a soma do intercepto com as curvas médias, $\hat{f} = \hat{\alpha} + \hat{f}_1 + \hat{f}_2$. A Figura 2.1 compara a superfície verdadeira em (a) com a estimada por BSH em (b). Visualmente, a superfície estimada parece estar próxima da verdadeira, evidenciando o bom desempenho da abordagem bayesiana para regressão *h-splines* não-paramétrica. O conceito de visualmente próximo é muito subjetivo, por este motivo no Capítulo 4 são apresentadas medidas de distância e afinidade entre curvas. Essas medidas são usadas para verificar o quão próxima da função verdadeira está a curva estimada.

O uso de muitas bases na estimação de uma função que é suave, pode resultar em uma sub-suavização. A Figura 2.1 não favorece a visualização de que as curvas estimadas pelo método DP possuem muitas rugosidades. Mas isso pode ser verificado na Figura 2.3, que mostra a superfície estimada pelo método. A falta de suavidade ocorre porque o pacote `DPpackage` utilizou 53 funções

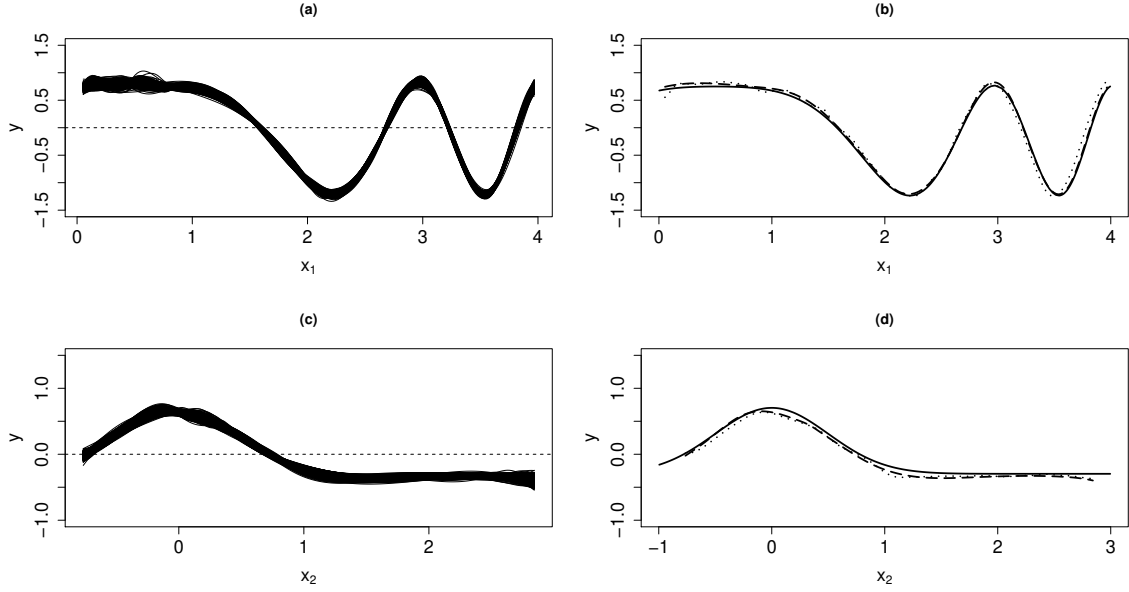


Figura 2.1: Amostras *a posteriori* de f_1 e f_2 do exemplo 1 em (a) e (c), respectivamente. Em (b) e (d), as linha contínuas são as curvas verdadeiras, as tracejadas são estimativas de BHS e as pontilhadas DP.

base para estimar f_1 e 53 para f_2 . Como esperado, constata-se então que a utilização de mais funções base que o necessário aumenta o esforço computacional e ainda pode prejudicar a estimação.

2.5.2 Exemplo 2

A obtenção dos dados é feita da seguinte forma:

1. Gere $X_1 \sim U(0, \pi)$ e calcule $X_2 = 2 - 4X_1/\pi + \epsilon_{X_1}$, com $\epsilon_{X_1} \sim N(0; 0, 5)$;
2. Calcule $f_1(X_1) = \exp(-0, 5X_1^2) \cos(4\pi X_1)$, $f_2(X_2) = X_2^3$ e $y = f_1(X_1) + f_2(X_2) + \epsilon$, com $\epsilon \sim N(0; 0, 05)$;
3. Repita os passos (1) e (2) até obter $n = 50$ observações.

Observe que aqui as covariáveis são correlacionadas. Isso pode trazer problemas na estimação. Além disso, o fato de a função f_1 ser bastante estruturada (pouco suave) também pode trazer dificuldades.

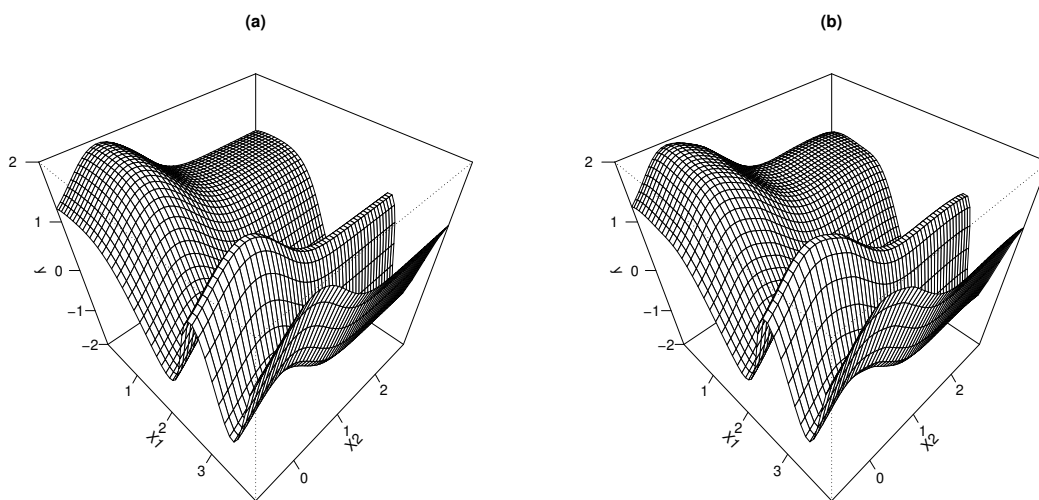


Figura 2.2: Em (a) a superfície apresentada no exemplo 1 e em (b) a superfície estimada por BHS.

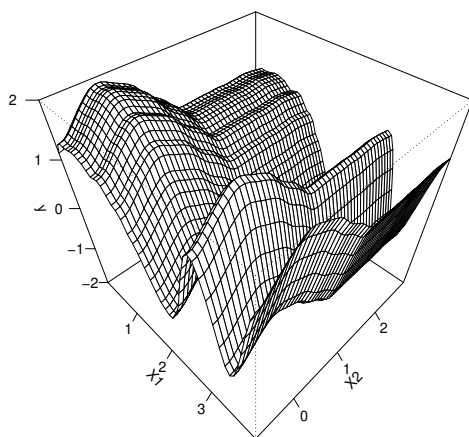


Figura 2.3: Superfície estimada pelo método DP para função do exemplo 1.

A Figura 2.4 apresenta f_1 e f_2 juntamente com as estimativas dos métodos BSH e DP, assim como a amostra de curvas *a posteriori*. Neste exemplo, as estimativas do método DP não se mantem tão próximas das funções f_1 e f_2 quanto as do BHS. Claramente o método BHS conseguiu

uma estimativa que consegue acompanhar bem a estrutura da curva original. Apesar de utilizar um número maior de funções base, a estimativa referente ao procedimento DP não conseguiu retratar essa estrutura. Provavelmente os problemas na estimação se deram pela correlação existente entre as covariáveis e pela grande curvatura apresentada pela função f_1 .

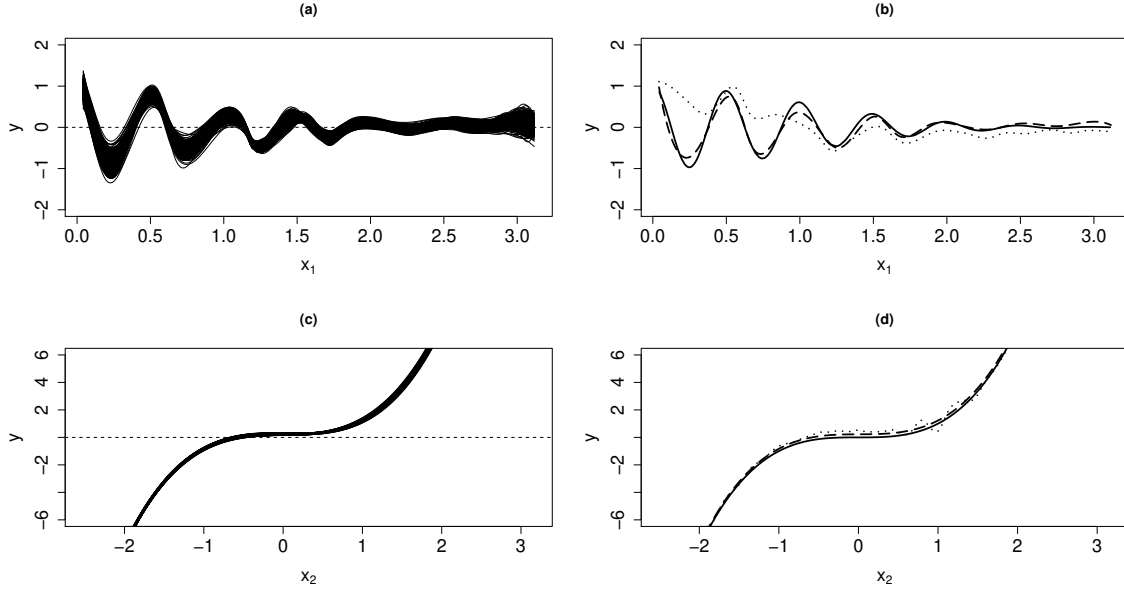


Figura 2.4: Amostras *a posteriori* de f_1 e f_2 do exemplo 2 em (a) e (c), respectivamente. Em (b) e (d), as linhas contínuas são as curvas verdadeiras, as tracejadas são estimativas de BHS e as pontilhadas DP.

A Tabela 2.2 mostra as médias *a posteriori* dos parâmetros λ_1 , λ_2 , σ^2 e α , acompanhadas pelos limites bayesianos de 95%. Observe na Tabela 2.2 que os valores de λ_1 e λ_2 novamente estão de acordo com o esperado, valores pequenos para funções estruturadas e valores altos para funções suaves.

Tabela 2.2: Médias *a posteriori* e limites bayesianos para λ_1 , λ_2 , σ^2 e α estimados para o exemplo 2.

Parâmetro	LI 2.5%	Média	LS 97.5%
α	-0,2217	-0,2022	-0,1839
σ^2	0,0023	0,0042	0,0078
λ_1	0,0010	0,0021	0,0040
λ_2	0,0031	0,0128	0,0298

Para estimar f_1 foram necessárias muitas funções base. A moda *a posteriori* para K_1 foi igual a 19. Para f_2 , que é uma função mais suave, a moda *a posteriori* foi de apenas 6 bases. Novamente, a estimação utilizando o pacote **DPpackage** usou 53 bases para cada função do modelo aditivo. Lembremos que o uso de muitas bases na estimação de uma função que é suave como f_2 , pode resultar em uma sub-suavização.

A Figura 2.5 apresenta o gráfico com curvas de nível para a superfície verdadeira, para a estimativa BHS e para a estimativa DP. Assim como foi observado na Figura 2.4, as curvas de níveis da Figura 2.5 indicam que a superfície estimada pelo método BHS aproxima-se mais da superfície verdadeira. O método DP não conseguiu capturar a estrutura da parte central das curvas de nível.

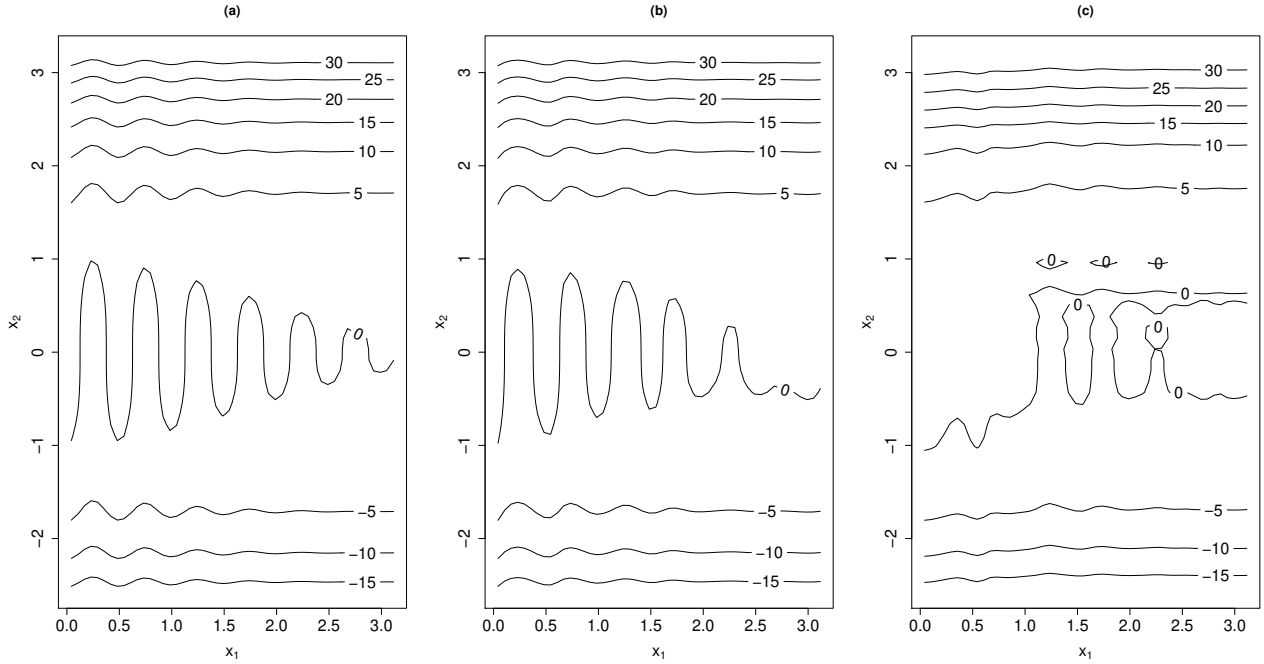


Figura 2.5: Em (a) as curvas de nível para a função do exemplo 2, em (b) a estimativa obtida por BHS e em (c) para DP.

2.5.3 Exemplo 3

Nos dois primeiros exemplos tratou-se de modelos aditivos. Neste exemplo é considerado um modelo com interação. Os dados são gerados conforme os seguintes passos:

1. Gere $X_1 \sim U(-2, 2)$ e $X_2 \sim U(-2, 2)$;
2. Calcule $f(X_1, X_2) = (X_1^2 + 3X_2^2) \exp(-X_1^2 - X_2^2)$ e $y = f(X_1, X_2) + \epsilon$, com $\epsilon \sim N(0; 0,05)$;
3. Repita os passos (1) e (2) até obter $n = 50$ observações.

Por não considerar interações, o modelo aditivo não deve conseguir capturar toda a estrutura da função a ser estimada. Entretanto, espera-se obter uma boa aproximação.

A amostra de curvas *a posteriori*, assim como as estimativas obtidas pelos métodos BHS e DP são apresentadas na Figura 2.6. Novamente observamos algumas rugosidades nas estimativas do procedimento DP. A Figura 2.6 não apresenta a função verdadeira, pois esta não pode ser separada em $f_1(X_1)$ e $f_2(X_2)$. Informações da distribuição *a posteriori* de α , λ_1 , λ_2 e σ^2 são dadas na Tabela 2.3.

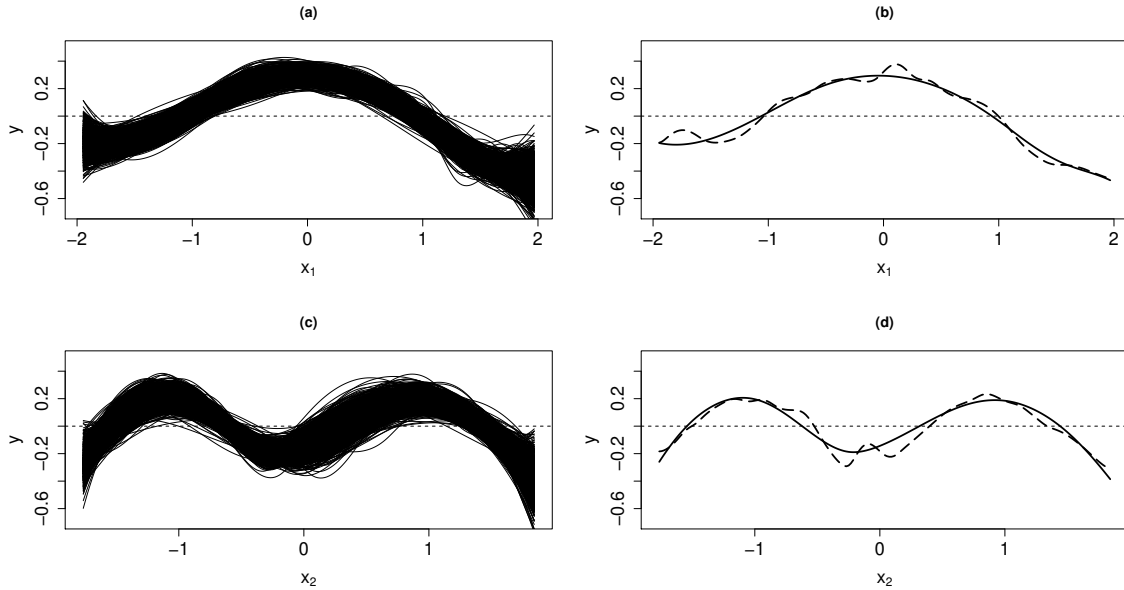


Figura 2.6: Amostras *a posteriori* de f_1 e f_2 do exemplo 3 em (a) e (c), respectivamente. Em (b) e (d), as linha contínuas são as estimativas do método BHS e as tracejadas são do DP.

Tabela 2.3: Médias *a posteriori* e limites bayesianos para λ_1 , λ_2 , σ^2 e α estimados para o exemplo 3.

Parâmetro	LI 2.5%	Média	LS 97.5%
α	0,3547	0,3939	0,4346
σ^2	0,0117	0,0192	0,0299
λ_1	0,1327	0,4932	1,1420
λ_2	0,0920	0,3248	0,7280

As estimativas suaves mostradas na Figura 2.6 foram obtidas com poucas funções base. A moda *a posteriori* para ambos K_1 e K_2 foi igual a 7. A Figura 2.7 mostra a superfície verdadeira e a estimada pelo método BHS. Comparando visualmente as superfícies pode-se argumentar que é uma aproximação razoável, tendo em vista todas as dificuldades já citadas no Capítulo 1 sobre a utilização de modelos com interação.

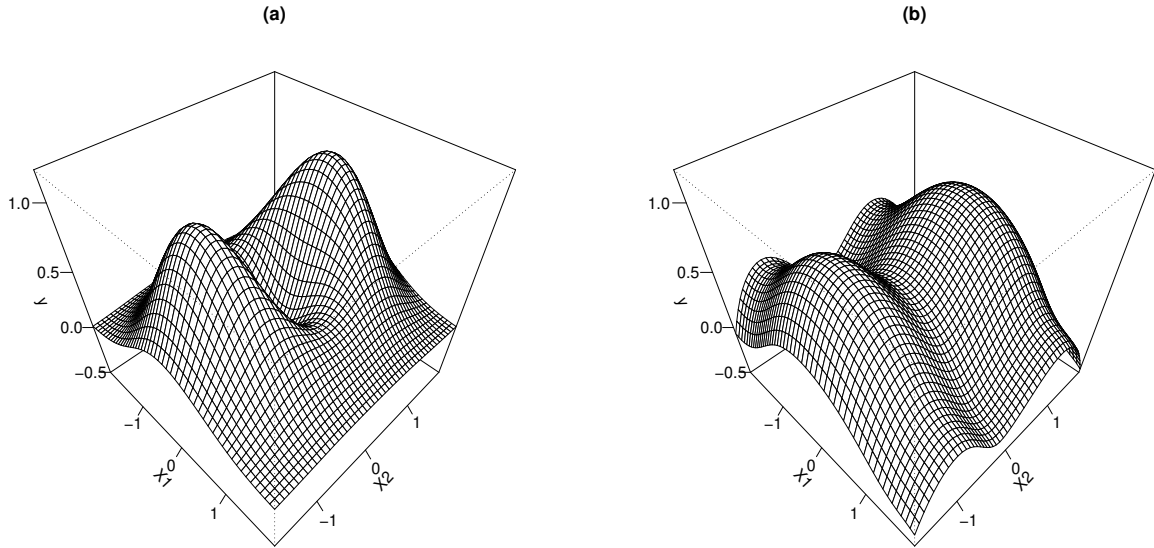


Figura 2.7: Em (a) a superfície referente à função do exemplo 3 e em (b) a estimativa obtida por BHS.

Nos exemplos apresentados pôde-se observar como a escolha adequada para o número de funções base afeta a estimação de uma função. Neles vimos as consequências do excesso de bases usadas para estimar uma função. Em outros exemplos deste trabalho veremos o que ocorre quando a

quantidade de bases é menor do que a adequada. Isso reforça a importância do uso de um método baseado em *h-splines* para escolha dos valores K_1, \dots, K_p .

Capítulo 3

Evidência Bayesiana para Modelos de Regressão Múltipla H-*Splines*

Neste capítulo é proposto um teste de hipóteses baseado na medida de evidência bayesiana de Pereira e Stern (1999), o FBST. Em seguida é feito um estudo de simulação para observar o comportamento da medida de evidência proposta.

3.1 Introdução

Várias situações podem ser descritas por modelos não lineares em que a variável resposta Y é uma função f das covariáveis X_1, X_2, \dots, X_p mais um erro. Muitas vezes tem-se dúvida se a forma mais adequada para descrever a função f seria f_0 ou f_0^* . Neste caso, seria necessário um procedimento para escolher um dentre os candidatos, um teste de hipóteses. Este teste pode até mesmo apontar que ambos, f_0 e f_0^* , são adequados para descrever f , deixando a escolha a cargo do pesquisador.

A literatura apresenta alguns critérios para seleção de modelos tais como AIC , BIC e DIC , além de testes bayesianos. Outros critérios como medidas de distância e afinidade entre curvas poderiam ser usados. A próxima seção descreve um teste de hipóteses baseado na medida de evidência bayesiana de Pereira e Stern (1999), o FBST. O principal desafio é adaptar de modo adequado o FBST para o contexto de testes com curvas.

3.2 Teste de Hipóteses Bayesiano para Igualdade de Curvas

Suponha que se deseja testar a hipótese nula

$$H_0 : f = f_0$$

contra a hipótese alternativa

$$H_1 : f \neq f_0.$$

A igualdade apresentada em H_0 é uma igualdade em quase toda parte, exceto para um conjunto de pontos de medida nula. O procedimento BHS apresentado no Capítulo 2 fornece uma distribuição *a posteriori* para f . Desse modo, pode-se obter a evidência bayesiana do FBST a favor de H_0 . Pereira e Stern (1999) definiram a seguinte medida de evidência em favor de uma hipótese precisa.

Definição 3.2.1. Considere um modelo estatístico, isto é, uma quintupla $(\mathcal{X}, \mathcal{A}, \mathbf{F}, \Theta, \pi)$, em que \mathcal{X} é o espaço amostral, \mathcal{A} é a σ -álgebra conveniente de subconjuntos de \mathcal{X} , \mathbf{F} é uma classe de distribuições de probabilidade em \mathcal{A} indexadas no espaço paramétrico Θ e π é uma função densidade a priori em Θ . Suponha que um subconjunto Θ_0 de Θ tendo medida de Lebesgue nula é de interesse. Seja $\pi(\theta|\mathbf{x})$ uma função densidade a posteriori de θ , dada a observação amostral \mathbf{x} , e $T(\mathbf{x}) = \{\theta \in \Theta : \pi(\theta|\mathbf{x}) > \sup_{\Theta_0} \pi(\theta|\mathbf{x})\}$. A medida de evidência de Pereira-Stern é definida como

$$EV(\Theta_0, \mathbf{x}) = 1 - P(\theta \in T(\mathbf{x})|\mathbf{x})$$

e um teste de Pereira-Stern é não rejeitar Θ_0 sempre que o valor de $EV(\Theta_0, \mathbf{x})$ seja “grande”.

A medida de evidência de Pereira-Stern considera todos os pontos que são menos “prováveis” do que algum ponto em Θ_0 . De acordo com esse procedimento, um valor “grande” da evidência significa que o subconjunto Θ_0 representa uma região do espaço paramétrico de alta densidade *a posteriori*. Portanto, os dados favorecem a hipótese nula. Por outro lado, um valor pequeno da evidência levaria à rejeição da hipótese nula.

Assim como descrito anteriormente na criação dos intervalos bayesianos, cada ponto amostral x_i pode ser avaliado em cada uma das estimativas da superfície *a posteriori* $\hat{f}^{(1)}, \dots, \hat{f}^{(m)}$.

Desse modo, considere apenas o ponto x_i . Para estimar $f(x_i)$ temos a amostra *a posteriori* $\hat{f}^{(1)}(x_i), \dots, \hat{f}^{(m)}(x_i)$, da qual pode-se obter uma função densidade. Essa função densidade pode ser estimada por métodos que utilizam *kernel* ou *splines*. Com essa densidade *a posteriori* é possível construir a medida de evidência de Pereira-Stern para $f_0(x_i)$, ou seja, $EV(f_0(x_i)|\mathbf{y}, \mathbf{x})$. A proposta é que a medida de evidência a favor de f_0 seja dada por

$$EV(f_0|\mathbf{y}, \mathbf{x}) = n^{-1} \sum_{i=1}^n EV(f_0(x_i)|\mathbf{y}, \mathbf{x}). \quad (3.2.1)$$

Portanto, a evidência a favor de uma função f_0 será a média das evidências em cada ponto da amostra.

A Figura 3.1 mostra um exemplo ilustrativo com uma covariável. Em (a) está uma amostra de mil curvas *a posteriori* obtida pelo procedimento BHS, além da curva tracejada f_0 a ser testada. Marcou-se neste gráfico três valores observados da covariável X , -1,63, 0,73 e 1,56. Para $X = -1,63$, tem-se a amostra *a posteriori* $\hat{f}^{(1)}(-1,63), \dots, \hat{f}^{(1000)}(-1,63)$, com a qual contrói-se a função densidade em (b). A área em cinza mostra a medida de evidência a favor de $f_0(-1,63)$, que considera todos os pontos que são menos “prováveis” do que $f_0(-1,63)$. Em (c) e (d) tem-se o mesmo para as observações 0,73 e 1,56 da covariável estudada. Considerando apenas estas três observações tem-se:

$$EV(f_0(-1,63)|\mathbf{y}, \mathbf{x}) = 0,344,$$

$$EV(f_0(0,73)|\mathbf{y}, \mathbf{x}) = 0,149 \text{ e}$$

$$EV(f_0(1,56)|\mathbf{y}, \mathbf{x}) = 0,752.$$

Se forem consideradas todas as observações pode-se obter $EV(f_0(x_1)|\mathbf{y}, \mathbf{x}), \dots, EV(f_0(x_n)|\mathbf{y}, \mathbf{x})$, e conseqüentemente $EV(f_0|\mathbf{y}, \mathbf{x})$. Neste exemplo específico a evidência a favor de f_0 , que é a média das evidências para cada observação, é de $EV(f_0|\mathbf{y}, \mathbf{x}) = 0,365$.

Para a regressão múltipla considere $\mathbf{X} = (X_1, \dots, X_p)^t$ como sendo as covariáveis e $\mathbf{x}_i = (x_{1i}, \dots, x_{pi})$ como a i -ésima observação das mesmas. Neste cenário, a reta $\mathbf{X} = \mathbf{x}_i$ não interceptará curvas como visto na Figura 3.1(a), mas sim superfícies *a posteriori*. A partir daí o procedimento é o mesmo descrito no exemplo ilustrativo.

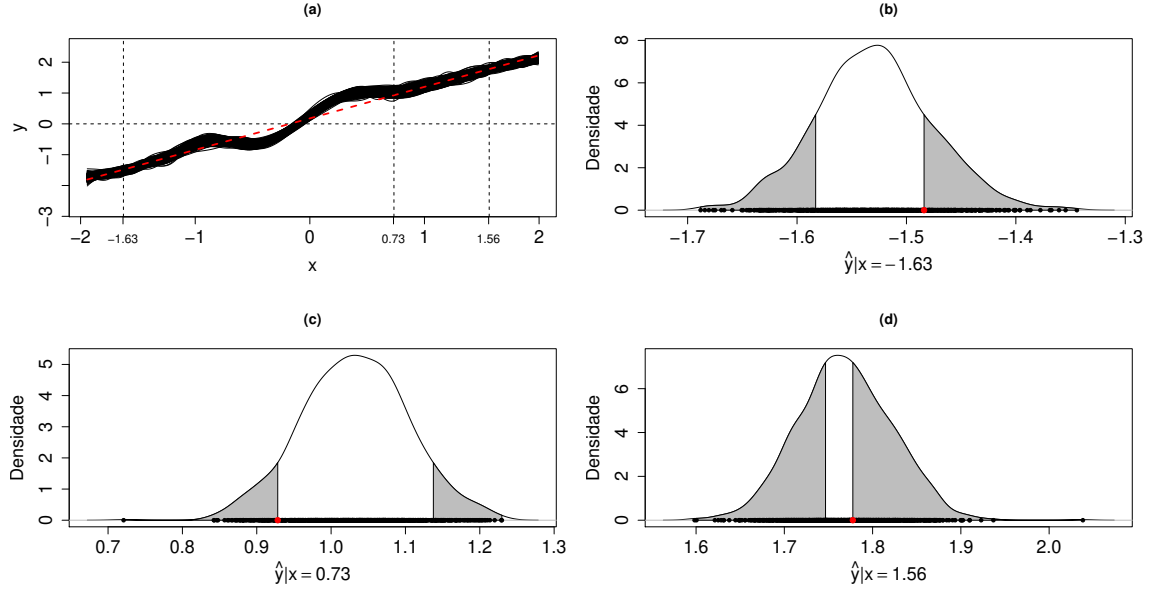


Figura 3.1: Em (a) amostra de curvas *a posteriori* juntamente da curva f_0 tracejada; em (b) a densidade de $\hat{f}^{(1)}(-1.63), \dots, \hat{f}^{(1000)}(-1.63)$, com $EV(f_0(-1.63|y, x))$ em cinza; e em (c) e (d) o mesmo para as observações 0.73 e 1.56 da covariável.

Após o cálculo da medida de evidência, é necessário construir uma regra de decisão, ou seja determinar um valor e_c de modo que rejeita-se H_0 se $EV(f_0|\mathbf{y}, \mathbf{x}) < e_c$ e aceita-se H_0 se $EV(f_0|\mathbf{y}, \mathbf{x}) \geq e_c$.

O valor e_c depende da função perda e pode assumir valores diferentes, isso porque existem variações da função perda com interpretações diferentes. Madruga, Esteves e Wechsler (2001) definem o problema da seguinte forma: Considere $D = \{d_0, d_1\}$ o espaço de decisões usual em um problema estatístico de teste de hipóteses, com d_0 representado a decisão de aceitar H_0 e d_1 a de rejeitar H_0 , e seja a função perda definida por $L : D \times \Theta_0 \rightarrow \mathbb{R}^+$ tal que $L(\text{Rejeitar } H_0, \theta) = a \{1 - I[\theta \in T(\mathbf{x})]\}$ e $L(\text{Aceitar } H_0, \theta) = b + cI[\theta \in T(\mathbf{x})]$, com a, b e $c > 0$. Para esta função perda, Madruga, Esteves e Wechsler (2001) mostram que o valor de corte é $e_c = (b + c)/(a + c)$.

Na prática, a escolha dos valores de a, b e c , necessários para a tomada de decisão não é simples e envolve a opinião do pesquisador sobre o erro mais (ou menos) danoso na sua decisão. Nos casos em que a evidência obtida é muito próxima de zero (ou de 1), a decisão natural é rejeitar (ou

aceitar) a hipótese H_0 . Nas demais situações, pode-se estabelecer o nível de significância do teste, como é feito nos testes clássicos, e buscar a validação do resultado obtido.

Neste trabalho não será estabelecido um ponto de corte, pois os objetivos aqui são, na Seção 3.3, estudar o comportamento da medida de evidência em diferentes cenários; e no Capítulo 7 usar essa medida de evidência para escolher um entre vários modelos candidatos para descrever um conjunto de dados reais.

3.3 Dados Simulados

Nesta seção é examinado, em um estudo de simulação, o comportamento da medida de evidência proposta. O modelo usado neste estudo é

$$y_i = f(x_{i1}, x_{i2}) + \epsilon_i = x_{i1} + x_{i2} + c \sin(2x_{i2}) + \epsilon_i, \quad i = 1, \dots, n, \quad (3.3.1)$$

com $\epsilon_1, \dots, \epsilon_n$ independentes, sendo $\epsilon_i \sim N(0, 0.01)$, com $-6 < x_{i1} < 6$ e $-6 < x_{i2} < 6$. A relação entre a variável resposta e as covariáveis pode ser linear ($c = 0$), ou não linear ($c \neq 0$). Dependendo do valor usado para c , o termo não linear da função ($\sin(2x_{i2})$) fica mais aparente. A Figura 3.2 mostra a superfície f para diferentes valores de c .

As hipóteses consideradas são

$$H_0 : f(X_1, X_2) = X_1 + X_2$$

e

$$H_1 : f(X_1, X_2) \neq X_1 + X_2.$$

São considerados vários cenários: $n = 30, 50, 100$ e $c = 0, 0.5, 1, 1.5$. Ou seja, é estudado o comportamento da evidência bayesiana quando são alterados o tamanho amostral e a magnitude de c . Observe que à medida que c cresce, a função f verdadeira se distancia da função testada $f(X_1, X_2) = X_1 + X_2$. Deste modo espera-se que

- quanto maior o valor de c , menor seja a evidência a favor de H_0 ,
- para $c = 0$, $EV(f_0|\mathbf{y}, \mathbf{x})$ cresça quando n cresce e

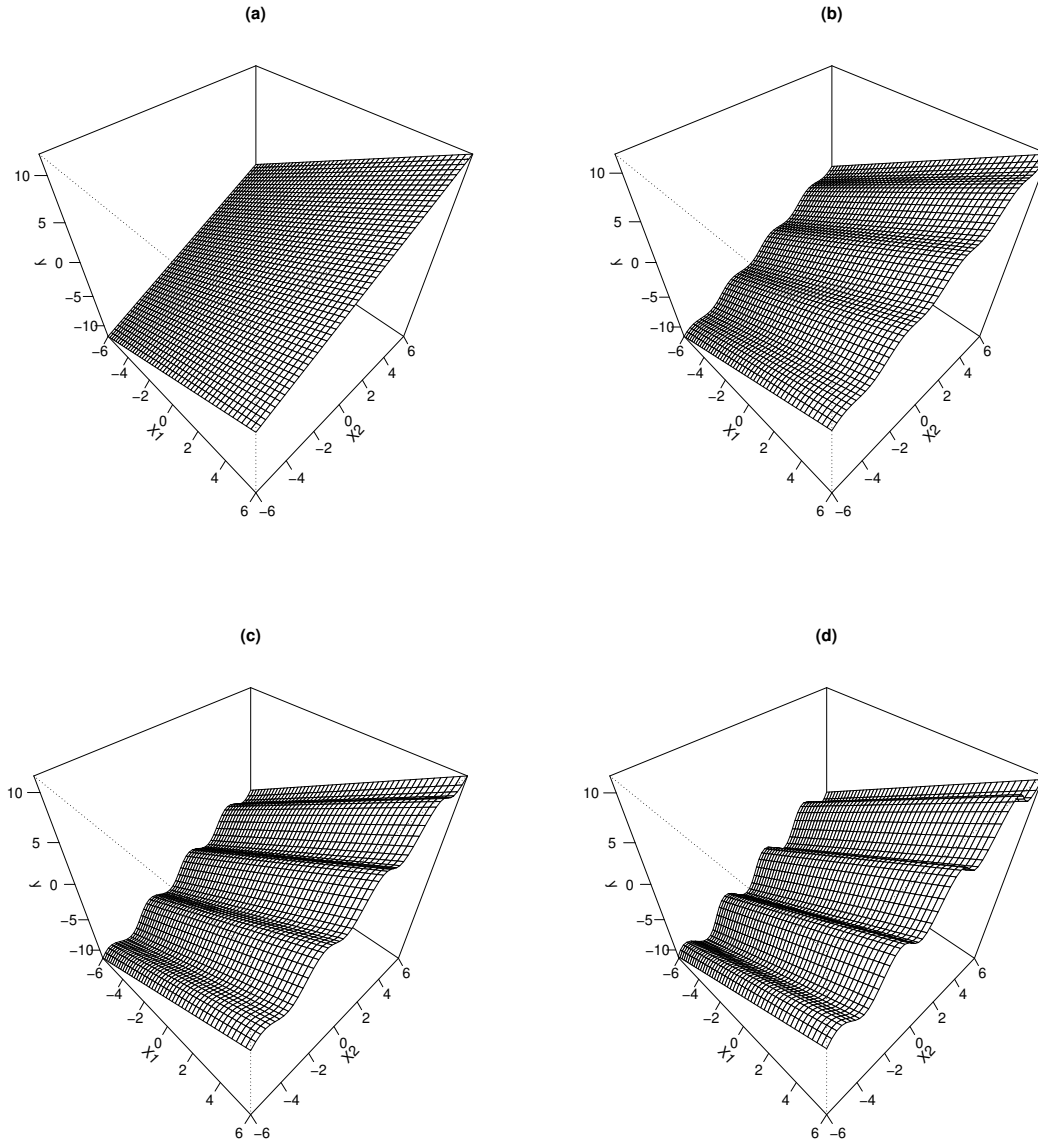


Figura 3.2: Superfície $Y = f(X_1, X_2)$ para diferentes valores de c . (a) $c = 0$, (b) $c = 0,5$, (c) $c = 1$ e (d) $c = 1,5$.

- para $c \neq 0$, $EV(f_0|\mathbf{y}, \mathbf{x})$ decresça quando n cresce.

A Tabela 3.1 mostra o valor de $EV(f_0|\mathbf{y}, \mathbf{x})$ obtida para cada cenário. Observe que a Tabela 3.1 descreve aproximadamente o que se esperava do comportamento de uma medida de evidência favorável a H_0 . Para amostras com $n = 30, 50$ e 100 , somente a partir de $c = 1$ obteve-se um valor “baixo” para $EV(f_0|\mathbf{y}, \mathbf{x})$. Ou seja, para amostras pequenas, precisa-se de um c maior

para diferenciar f de f_0 . Para $n = 200$, a evidência a favor de H_0 foi baixa para todo $c > 0$. Aparentemente a proposta de calcular, usando (3.2.1), a evidência bayesiana para testar a igualdade de funções é razoável. No Capítulo 7 essa proposta é utilizada em um conjunto de dados reais com objetivo de escolher qual forma funcional é mais adequada para descrevê-lo.

Tabela 3.1: Valores de $EV(f_0|\mathbf{y}, \mathbf{x})$ obtidos para diferentes valores de n e c .

n	$c = 0$	$c = 0,5$	$c = 1$	$c = 1,5$
30	0,5397	0,2456	0,0982	0,0806
50	0,4797	0,1469	0,0977	0,0704
100	0,7618	0,2107	0,1045	0,0379
200	0,8170	0,0581	0,0279	0,0163

Capítulo 4

Estimação Sequencial Adaptativa para Regressão Múltipla H-*Splines*

Neste capítulo, é apresentada uma generalização do método h-*splines* proposto por Dias (1999). Primeiramente é apresentada a versão original do método, que considera apenas uma covariável. Em seguida, é apresentado o algoritmo que estende o método para o caso em que tem-se várias covariáveis. Por último é feito um estudo de simulação que compara os resultados obtidos pelo procedimento proposto com os conseguidos pelos pacotes `gamlss` e `mgcv` em R.

4.1 Introdução

Assim como no Capítulo 2, o objetivo é estimar uma função f . O procedimento usado é a regressão h-*splines* não paramétrica múltipla. A principal dificuldade aqui é como definir o número de funções base mais adequado. Neste capítulo é apresentado um procedimento sequencial adaptativo para obter o melhor número de bases para cada covariável e assim estimar f .

Considere o caso em que temos apenas uma covariável. Como já foi mencionado na Seção 2.2, na abordagem usual da regressão penalizada com *splines* as estimativas são obtidas encontrando θ que minimize

$$(\mathbf{y} - \mathbf{X}\theta)^t (\mathbf{y} - \mathbf{X}\theta) + \lambda \theta^t \Omega \theta.$$

Equivalentemente, $\boldsymbol{\theta}$ é obtido como uma solução do sistema linear $(\mathbf{X}^t\mathbf{X} + \lambda\boldsymbol{\Omega})\boldsymbol{\theta} = \mathbf{X}^t\mathbf{y}$. Como há uma única covariável, \mathbf{X} é uma matriz $n \times K$ contendo as funções base para $\mathbf{x} = (x_1, \dots, x_n)^t$ e $\boldsymbol{\Omega}$ é a matriz de penalizações. Ambos K e λ controlam o equilíbrio entre suavização e fidelidade aos dados. Dias (1999) propôs um algoritmo em que se inicia com o menor número de funções base possível e acrescenta-se bases até que um critério de parada seja satisfeito. O algoritmo proposto é o seguinte:

1. Escolha K_0 como número inicial de funções base e fixe λ_0 ;
2. Obtenha $\hat{\boldsymbol{\theta}}$ solucionando $(\mathbf{X}^t\mathbf{X} + \lambda_0\boldsymbol{\Omega})\boldsymbol{\theta} = \mathbf{X}^t\mathbf{y}$;
3. Encontre $\hat{\lambda}$ que minimize

$$GCV(\lambda) = \frac{n(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})}{\{n - \text{tr}[\mathbf{H}(\lambda)]\}^2}, \quad (4.1.1)$$

com $\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{X}^t$;

4. Calcule $f_{K_0, \hat{\lambda}} = \mathbf{H}(\hat{\lambda})\mathbf{y}$;
5. Acrescente uma função base ao número de funções base e repita os passos de (2) a (4) para obter $f_{K_0+1, \hat{\lambda}}$;
6. Para um número real $\delta > 0$, se a distância $d(f_{K_0, \hat{\lambda}}, f_{K_0+1, \hat{\lambda}}) < \delta$, pare o procedimento. O valor δ pode ser determinado empiricamente de acordo com uma particular distância $d(\cdot, \cdot)$.

Dias (1999) mostrou que esta abordagem teve uma boa performance para estimar vários tipos de curvas.

A principal motivação deste capítulo é estender este algoritmo para o caso em que temos várias covariáveis. Como já foi mencionado, a generalização é feita considerando modelos aditivos.

4.2 Estimação Adaptativa Múltipla

Assim como na abordagem bayesiana apresentada no Capítulo 2, consideramos modelos aditivos

$$f = \boldsymbol{\alpha} + \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\theta}_j = \mathbf{X}\boldsymbol{\theta},$$

em que novamente $\boldsymbol{\theta} = (\alpha, \theta_1, \dots, \theta_p)^t$, $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_1, \dots, \mathbf{X}_p)$ e \mathbf{X}_j é uma matrix $n \times K_j$ contendo os valores das K_j funções base calculadas em \mathbf{x}_j .

A estimação é feita minimizando

$$L_p(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{\Lambda}\boldsymbol{\theta},$$

sendo $\boldsymbol{\Lambda} = \text{diag}(0, \lambda_1\boldsymbol{\Omega}_1, \dots, \lambda_p\boldsymbol{\Omega}_p)$. Derivando $L_p(\boldsymbol{\theta})$ em relação a $\boldsymbol{\theta}$ e igualando a $\mathbf{0}_p$ temos

$$\frac{\partial L_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\mathbf{X}^t\mathbf{y} + \mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\theta}} + \boldsymbol{\Lambda}\hat{\boldsymbol{\theta}} = \mathbf{0}_p \quad \Rightarrow \quad (\mathbf{X}^t\mathbf{X} + \boldsymbol{\Lambda})\hat{\boldsymbol{\theta}} = \mathbf{X}^t\mathbf{y}.$$

$$\Rightarrow \hat{\boldsymbol{\theta}} = (\mathbf{X}^t\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^t\mathbf{y} \quad (4.2.1)$$

Dessa forma, o problema consiste em encontrar o número de funções base mais adequado para cada covariável. A idéia é criar um algoritmo que trabalhe como a proposta de Dias (1999), ou seja, inicie com o número mínimo de bases e vá acrescentando bases até que um critério de parada seja satisfeito.

O procedimento sequencial adaptativo múltiplo é dado por:

1. Escolha $\mathbf{K} = (K_1, \dots, K_p)^t$ como as quantidades iniciais de funções base;

2. Encontre $\hat{\boldsymbol{\lambda}}$ que minimize

$$GCV(\lambda) = \frac{n[\mathbf{y} - \mathbf{H}(\boldsymbol{\lambda})\mathbf{y}]^t[\mathbf{y} - \mathbf{H}(\boldsymbol{\lambda})\mathbf{y}]}{\{n - \text{tr}[\mathbf{H}(\boldsymbol{\lambda})]\}^2}, \quad (4.2.2)$$

com $\mathbf{H}(\boldsymbol{\lambda}) = \mathbf{X}(\mathbf{X}^t\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^t$;

3. Com $\hat{\boldsymbol{\lambda}}$ obtenha $\hat{\boldsymbol{\Lambda}}$ e em seguida $\hat{\boldsymbol{\theta}} = (\mathbf{X}^t\mathbf{X} + \hat{\boldsymbol{\Lambda}})^{-1}\mathbf{X}^t\mathbf{y}$;

4. Calcule $f^{\mathbf{K}, \hat{\boldsymbol{\lambda}}} = \mathbf{X}\hat{\boldsymbol{\theta}}$, além dos $f_j^{K_j} = \mathbf{X}_j\hat{\boldsymbol{\theta}}_j$, para $j = 1, \dots, p$.

5. Acrescente uma base ao número de funções base de cada covariável e repita os passos de (2) a (3) para obter $f_1^{K_1+1}, \dots, f_p^{K_p+1}$;

6. Para um número real $\delta > 0$ e para $j = 1, \dots, p$, verifique se as distâncias $d_j(f_j^{K_j}, f_j^{K_j+1}) < \delta$. Caso existam j 's tais que $d_j > \delta$, acrescente um função base apenas

para estas covariáveis e repita os passos (2) e (3). Faça isso até que $d_j < \delta$, $\forall j$. O valor δ pode ser determinado empiricamente de acordo com uma particular distância $d(\cdot, \cdot)$.

A minimização da validação cruzada generalizada (GCV) do passo 2 pode ser feita através do método de Newton, usando a abordagem de Gu e Wahba (1991) ou a de Wood (2004).

Denominaremos o procedimento proposto por método sequencial h-splines (SHS).

4.3 Seleção da Distância $d(\cdot, \cdot)$

Nesta seção vamos apresentar algumas candidatas para a distância $d(\cdot, \cdot)$ usada no algoritmo sequencial proposto. Antes, devemos definir a norma \mathcal{L}_2 e o conjunto $\mathcal{W}_2^2[a, b]$. A norma \mathcal{L}_2 de uma função f é definida como

$$\sqrt{\int f^2(u)du},$$

enquanto $\mathcal{W}_2^2[a, b]$ é o conjunto formado pelas funções definidas no intervalo $[a, b]$, cuja primeira derivada é absolutamente contínua e a norma \mathcal{L}_2 da segunda derivada é finita.

Como critério de parada para o algoritmo, Dias (1999) sugeriu $d(\cdot, \cdot)$ como sendo uma medida de afinidade entre duas funções baseada na distância de Hellinger. Para uma função qualquer g em $\mathcal{W}_2^2[a, b]$, defina

$$t_g(s) = \frac{g^2(s)}{\int g^2(u)du}.$$

Então, $t_g \geq 0$ e $\int t_g(u)du = 1$. Para quaisquer funções $f, g \in \mathcal{W}_2^2[a, b]$, o quadrado da distância de Hellinger é dado por

$$H^2(f, g) = \int \left(\sqrt{t_f(u)} - \sqrt{t_g(u)} \right)^2 du = 2[1 - \rho(f, g)],$$

em que

$$\rho(f, g) = \int \sqrt{t_f(u)t_g(u)}du = \int \sqrt{\frac{f^2(u)}{\int f^2(v)dv} \frac{g^2(u)}{\int g^2(v)dv}}du = \int \frac{|f(u)g(u)|}{\sqrt{\int f^2(v)dv \int g^2(v)dv}}du$$

é a afinidade de Hellinger entre f e g . Não é difícil verificar que $0 \leq \rho(f, g) \leq 1$, $\forall f, g \in \mathcal{W}_2^2[a, b]$.

Note que $H^2(f, g)$ é mínima quando $\rho(f, g) = 1$.

Com essa medida em mãos pode-se definir um critério de parada para os procedimentos descritos. Por exemplo, pode-se definir que quando $\rho(f^K, f^{K+1}) > \delta$ o procedimento deve ser parado.

Uma outra medida de afinidade entre curvas é a medida do cosseno do ângulo ϕ formado entre f e g , dada por

$$\cos(\phi_{f,g}) = \frac{\int f(u)g(u)du}{\sqrt{\int f^2(u)du \int g^2(u)du}}.$$

A região de variação desta medida é $-1 \leq \cos \phi_{f,g} \leq 1$. Para manter esta medida variando entre 0 e 1, assim como a afinidade de Hellinger, o cosseno deveria ser dado por

$$|\cos(\phi_{f,g})| = \frac{|\int f(u)g(u)du|}{\sqrt{\int f^2(u)du \int g^2(u)du}}. \quad (4.3.1)$$

Neste trabalho, quando nos referirmos ao cosseno do ângulo formado entre duas curvas, sempre será considerada a equação (4.3.1). Esta é a medida de afinidade usada no trabalho como critério de parada no método SHS.

Quando discretizamos as curvas, ou seja, quando as curvas são representadas por vetores de valores, o cosseno é dado por

$$|\cos(\phi_{f,g})| = \frac{|\sum_{i=1}^n f_i g_i|}{\sqrt{\sum_{i=1}^n f_i^2 \sum_{i=1}^n g_i^2}}.$$

Dias (1999) explorou simulações em que buscava uma distribuição para a afinidade entre as curvas verdadeiras e as estimadas via o método adaptativo *h-splines*. Esta ideia será utilizada no Capítulo 5 a fim de construir um teste de hipóteses. Porém, a medida de afinidade de Hellinger é substituída pelo cosseno.

Várias medidas de distâncias bem conhecidas também podem ser usadas como critério de parada, como por exemplo

- distância L_1 , dada por $L_1(f, g) = \int |f(u) - g(u)| du$,
- distância quadrática integrada, $DQI(f, g) = \int [f(u) - g(u)]^2 du$,
- distância de Hellinger, definida por $H(f, g) = \left\{ \int [\sqrt{t_f(u)} - \sqrt{t_g(u)}]^2 du \right\}^{1/2}$ e
- distância de Kullbak-Leibler, $KL(f, g) = \int \{\log[f(u)] - \log[g(u)]\} \log[f(u)] du$.

A distância quadrática integrada ainda será considerada neste trabalho. No Capítulo 5 a *DQI* é utilizada na construção de um teste de hipóteses.

O próximo passo é verificar o desempenho do procedimento SHS na estimação de uma série de superfícies. Os resultados do método são comparados com os resultados dos pacotes `gamlss` e `mgcv` em R. O pacote `gamlss` contém funções para o ajuste de modelos aditivos generalizados para parâmetros de locação, escala e forma. Funções no pacote `mgcv` ajustam modelos aditivos generalizados e têm como destaque a eficiência na seleção do parâmetro de suavização, que é feita utilizando algoritmos muito estáveis. Ambos utilizam regressão penalizada com *splines* na estimação de modelos aditivos. Os métodos de estimação usados pelos pacotes `gamlss` e `mgcv` serão denominados GAMLSS e MGCV, respectivamente.

4.4 Dados Simulados

Dois exemplos de superfícies são considerados nesta seção, que são os exemplos 1 e 2 do Capítulo 2. O objetivo é acompanhar o desempenho do método de estimação proposto (SHS) e, além disso, confrontar os resultados com os obtidos pelos procedimentos GAMLSS e MGCV.

4.4.1 Exemplo 1 (Continuação)

O mesmo conjunto de dados gerados no exemplo 1 do Capítulo 2 é usado aqui. Os métodos SHS, GAMLSS e MGCV ajustaram o modelo aditivo para este conjunto de dados. A Figura 4.1 apresenta as curvas verdadeiras traçadas juntamente com as estimativas. Visualmente, os três métodos de estimação obtiveram bons resultados, com estimativas muito próximas às curvas verdadeiras. Entretanto, agora além de realizar uma inspeção visual, podemos analisar as distâncias e/ou afinidades entre as estimativas e a curva verdadeira. A Tabela 4.1 mostra, além dos valores de K_1 e K_2 , os cossenos e *DQI* entre as superfícies estimadas e a verdadeira.

Analizando a Tabela 4.1, vemos que as quantidades de funções base K_1 e K_2 obtidas pelo SHS são pouco diferentes dos valores mais frequentes obtidos pelo procedimento bayesiano BHS, que foram 18 e 16. Os valores de K_1 e K_2 dos outros métodos são valores padrão nos respectivos

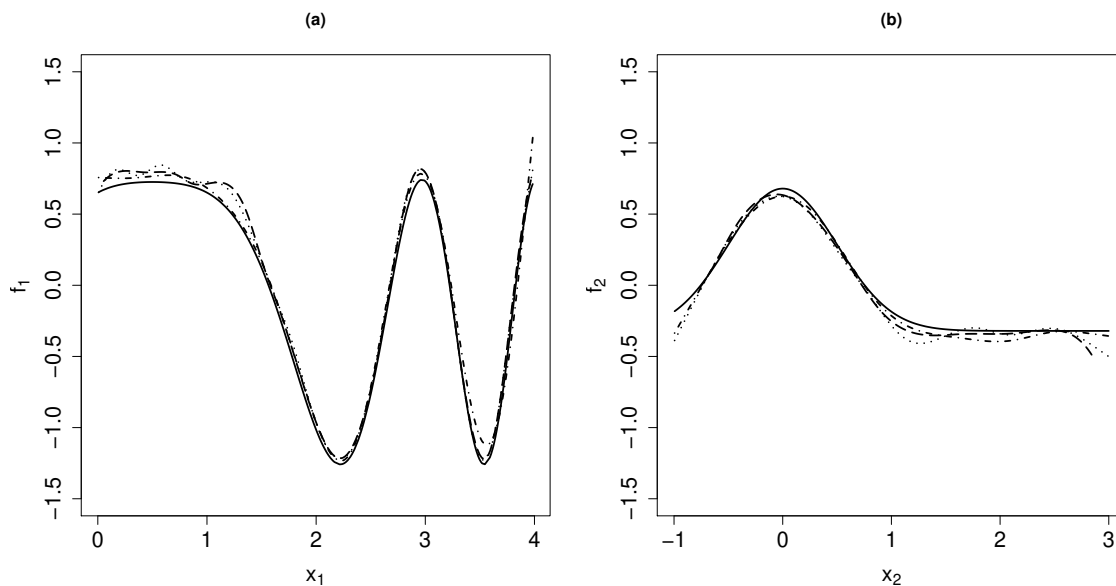


Figura 4.1: Para o exemplo 1, em (a) tem-se a função f_1 e em (b) f_2 . Em ambos a linha contínua representa a curva original, a linha tracejada a estimativa SHS, a linha pontilhada o GAMLSS e a linha de traços e pontos o método MGCV.

Tabela 4.1: Comparação entre os métodos SHS, MGCV e GAMLSS para o número de funções base e as medidas $|\cos(\phi_{f,\hat{f}})|$ e $DQI(f, \hat{f})$, sendo f referente ao exemplo 1.

Procedimento	K_1	K_2	$ \cos(\phi_{f,\hat{f}}) $	$DQI(f, \hat{f})$
SHS	19	12	0,9996	0,0409
GAMLSS	13	13	0,9993	0,0622
MGCV	9	9	0,9989	0,1031

pacotes. Considerando o cosseno e a distância quadrática integrada como medidas de adequação, pode-se observar que os procedimentos utilizados têm desempenhos muito semelhantes.

4.4.2 Exemplo 2 (Continuação)

O mesmo conjunto de dados gerado no exemplo 2 do Capítulo 2 é usado aqui. Lembremos que este exemplo apresenta mais dificuldades para estimação do que o exemplo 1. As covariáveis são correlacionadas e a função f_1 é bastante estruturada. Os métodos SHS, GAMLSS e MGCV ajustaram o modelo aditivo para este conjunto de dados. A Figura 4.2 apresenta as curvas verda-

deiras traçadas juntamente com as estimativas. É importante observar como os métodos MGCV e GAMLSS não conseguiram capturar a estrutura da função f_1 em suas estimativas. Por outro lado, a curva estimada pelo método SHS se mantém sempre próxima à verdadeira curva.

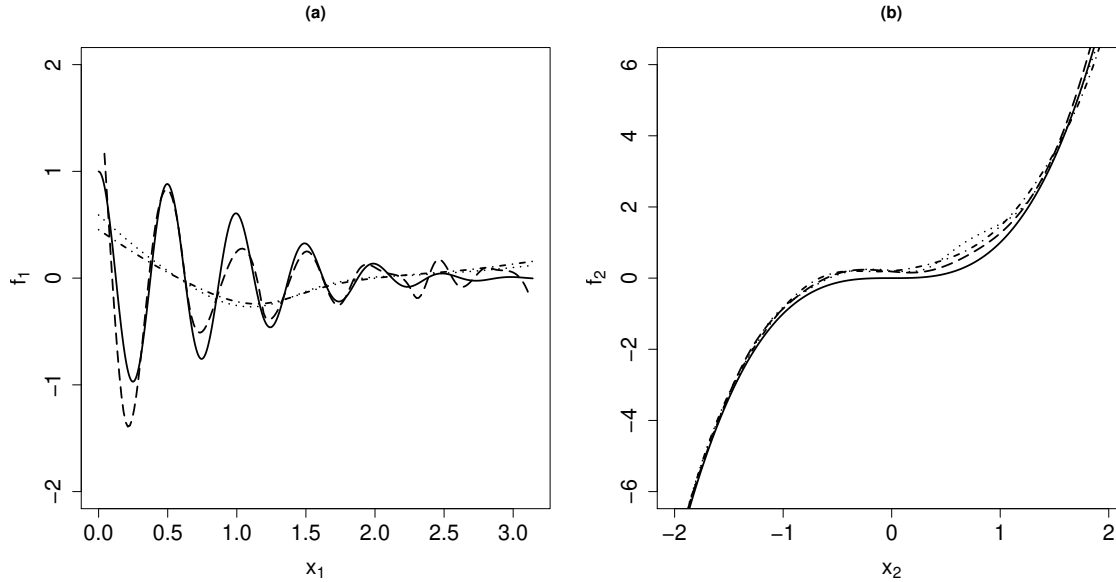


Figura 4.2: Para o exemplo 1, em (a) tem-se a função f_1 e em (b) f_2 . Em ambos a linha contínua representa a curva original, a linha tracejada a estimativa SHS, a linha pontilhada o GAMLSS e a linha de traços e pontos o método MGCV.

A Tabela 4.2 apresenta os valores de K_1 e K_2 , os cossenos e DQI entre as superfícies estimadas e a verdadeira. Novamente, o número de funções base utilizados pelos métodos GAMLSS e MGCV são os valores padrão de seus respectivos pacotes em R. O método SHS estimou f_1 utilizando 25 funções base e usou apenas 12 para f_2 . Essa coerência, de utilizar muitas bases para funções com maior curvatura e poucas bases para curvas mais suaves é uma característica forte do método SHS. Essa escolha do número adequado de funções base ajudou a obter melhores estimativas. Mais uma vez, consideremos o cosseno e a distância quadrática integrada como medidas de adequação. A Tabela 4.2 mostra que não há muita diferença entre os métodos quando analisamos a medida cosseno. Entretanto, existe uma diferença considerável na medida DQI obtida por cada método. O método SHS apresenta uma distância menor que os demais em relação à verdadeira curva, reforçando o que já apontava a inspeção visual.

Tabela 4.2: Comparação entre os métodos SHS, MGCV e GAMLSS para o número de funções base e as medidas $\cos(\phi_{f,\hat{f}})$ e $DQI(f, \hat{f})$, sendo f referente ao exemplo 2

Procedimento	K_1	K_2	$\cos(\phi_{f,\hat{f}})$	$DQI(f, \hat{f})$
SHS	25	12	0,9999	0,2267
GAMLSS	13	13	0,9996	1,7555
MGCV	9	9	0,9995	2,1368

Uma possível explicação para que os métodos MGCV e GAMLSS não tenham conseguido capturar toda a estrutura da função f_1 é que ambos utilizaram poucas funções base. A quantidade de bases usadas por estes métodos foi suficiente para estimar f_2 , mas não para f_1 . O método SHS usou $K_1 = 25$ e conseguiu reproduzir boa parte da estrutura de f_1 . Em situações de curvas mais suaves como no exemplo 1, as estimativas dos métodos são bastante semelhantes. Entretanto, para casos como o exemplo 2, a adaptabilidade do procedimento SHS é um diferencial.

A principal vantagem do método SHS em relação aos concorrentes é que este fornece de modo automático o número de funções base que deve ser utilizado para cada covariável e em seguida estima os parâmetros do modelo. Em geral, nos pacotes `gamlss` e `mgcv` o usuário deve fazer um estudo mais aprofundado para verificar se as quantidades de bases usadas são adequadas. E caso não sejam, pode-se perder uma grande quantidade de tempo até encontrar a melhor combinação de K_1, K_2, \dots, K_p .

Para este exemplo, ao utilizar $K_1 = 25$ e $K_2 = 12$ como indicado pelo método SHS, as medidas $\cos(\phi_{f,\hat{f}})$ e $DQI(f, \hat{f})$ do procedimento MGCV passam de 0,9995 e 2,1368 para 0,9999 e 0,2403, respectivamente. Para o procedimento GAMLSS, não há alteração significativa nas medidas mesmo após a mudança do número de bases. Observe que mesmo com todos os métodos utilizando $K_1 = 25$ e $K_2 = 12$ o procedimento SHS tem um desempenho superior.

4.4.3 Exemplo 3 (Continuação)

Os dados do exemplo 3 do Capítulo 2, que considera um modelo com interação, são usados aqui. Novamente, o objetivo é verificar se podemos conseguir uma boa aproximação utilizando modelos aditivos. A Figura 4.3 mostra que as estimativas obtidas com os métodos SHS, GAMLSS

e MGCV foram próximas e visualmente similares à estimativa do método BHS obtida no Capítulo 2. A Tabela 4.3 reforça a afirmação de que não há muita diferença entre os métodos, pois as medidas de cosseno e DQI dos mesmos são muito próximas. Novamente, pode-se argumentar que é uma aproximação razoável, tendo em vista todas as dificuldades já citadas no Capítulo 1 sobre a utilização de modelos com interação.

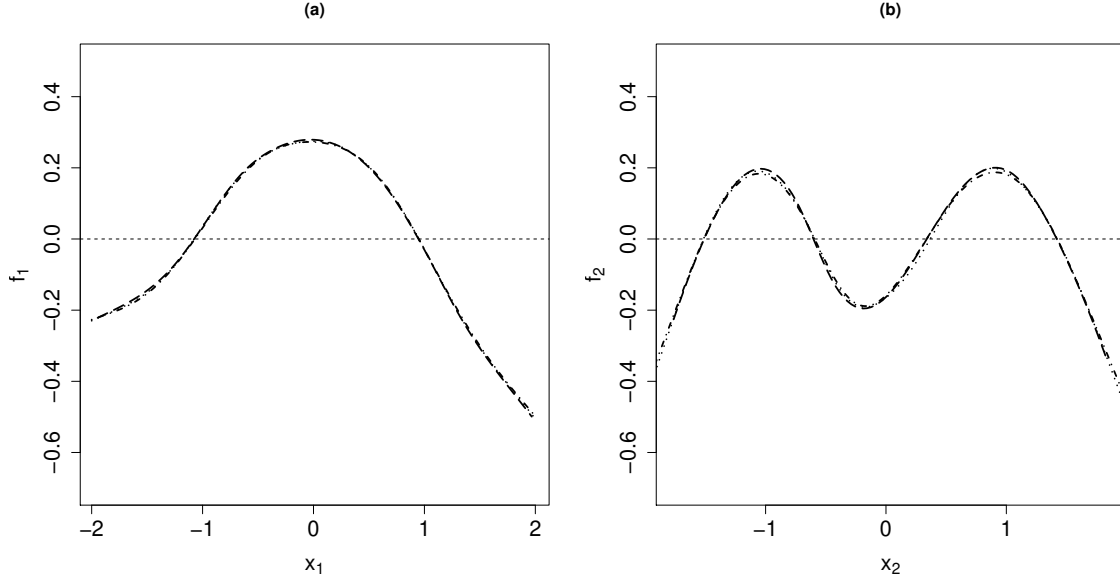


Figura 4.3: Para o exemplo 3, em (a) tem-se a função f_1 e em (b) f_2 . Em ambos a linha tracejada representa a estimativa SHS, a linha pontilhada o GAMLSS e a linha de traços e pontos o método MGCV.

Tabela 4.3: Comparação entre os métodos SHS, MGCV e GAMLSS para o número de funções base e as medidas $\cos(\phi_{f,\hat{f}})$ e $DQI(f, \hat{f})$, sendo f referente ao exemplo 3.

Procedimento	K_1	K_2	$\cos(\phi_{f,\hat{f}})$	$DQI(f, \hat{f})$
SHS	11	11	0,9749	0,5507
GAMLSS	13	13	0,9741	0,5677
MGCV	9	9	0,9742	0,5675

Como no Capítulo 2, foram traçadas a curva original e a estimativa obtida pelo método BHS. A Figura 4.4 apresenta as curvas de nível para a superfície verdadeira e para a estimativa SHS. Comparando os gráficos nota-se que a superfície estimada pelo método SHS aproxima-se razoavelmente da superfície verdadeira.

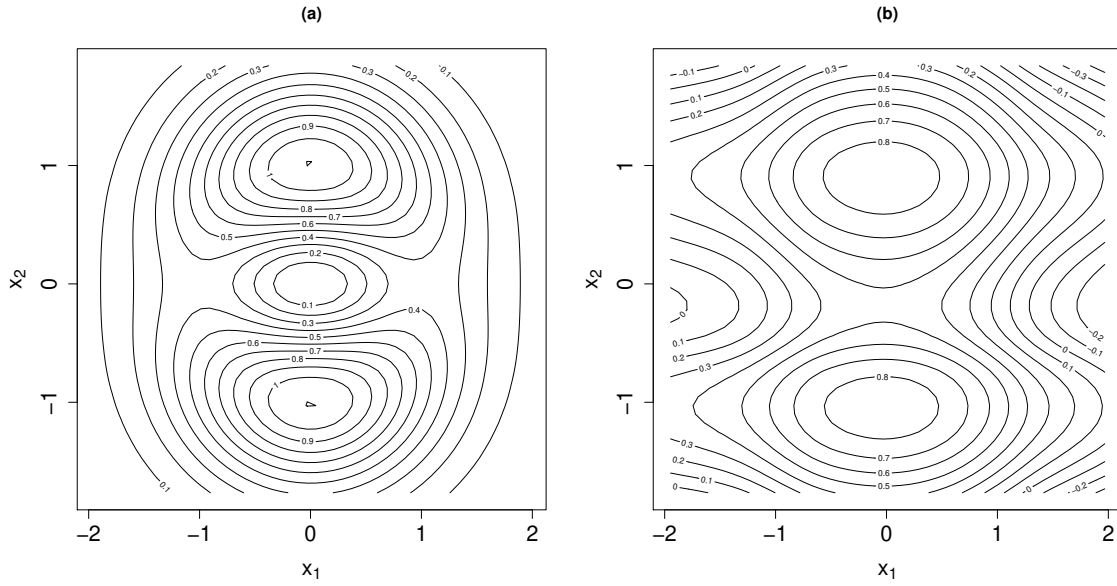


Figura 4.4: Em (a) as curvas de nível para a função do exemplo 3 e em (b) a estimativa obtida por SHS.

O método proposto neste capítulo será generalizado no Capítulo 6. Essa generalização tratará de casos em que a variável resposta Y possui restrições do tipo ser positiva ($Y \geq 0$) ou ser binária ($Y \in \{0, 1\}$).

Capítulo 5

Teste de Hipóteses para Regressão Múltipla H-*Splines*

Neste capítulo é proposto um teste de hipóteses baseado na distância quadrática integrada, que é uma medida de distância entre curvas. Em seguida é feito um estudo de simulação para estimar de modo empírico o poder do teste em diversos cenários.

5.1 Introdução

Como já citado no Capítulo 3, um teste de hipóteses pode ser útil para verificar se uma forma funcional é adequada para modelar um certo conjunto de dados, ou para fazer uma seleção da forma funcional mais adequada. Neste capítulo serão consideradas as mesmas hipóteses do Capítulo 3. Entretanto, aqui a abordagem é frequentista.

A idéia é que o teste de hipóteses deva se basear em uma distância (ou afinidade) entre a função f e uma estimativa desta, \hat{f} . Em seu trabalho, após estudos de simulação, Dias (1999) sugere a distribuição beta para a afinidade de Hellinger citada no Capítulo 4, $\rho(f, \hat{f})$, sendo \hat{f} obtida através do procedimento sequencial adaptativo. Aqui considera-se \hat{f} obtida segundo o método SHS.

5.2 Teste DQI para Igualdade de Curvas

Como já foi mencionado, Dias (1999) explorou simulações em que buscava uma distribuição para a afinidade entre as curvas verdadeiras e as estimadas via o método adaptativo *h-splines*. Suas simulações apontavam que possivelmente a afinidade de Hellinger tem uma distribuição beta. Partindo desta idéia, decimos obter a distribuição de uma distância entre a curva verdadeira e a estimada, considerando esta distância como sendo a estatística do teste. A seguir apresentamos o desenvolvimento para encontrar a distribuição da distância $DQI(f, \hat{f})$, apresentada no Capítulo 4.

Primeiramente, considere que

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon} \quad (5.2.1)$$

sendo $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$. Para encontrar a distribuição do cosseno deve-se primeiramente encontrar a distribuição de \hat{f} . Sabe-se que $\hat{\mathbf{f}} = \hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, com

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{X} + \boldsymbol{\Lambda})^{-1} \mathbf{X}^t.$$

Partindo dos pressupostos do modelo temos

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \Rightarrow \mathbf{y} \sim N(\mathbf{f}, \sigma^2 \mathbf{I}) \Rightarrow \hat{\mathbf{f}} \sim N(\mathbf{H}\mathbf{f}, \sigma^2 \mathbf{H}\mathbf{H}^t). \quad (5.2.2)$$

Provost e Rudiuk (1996) mostraram que se considerarmos $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com $\text{posto}(\boldsymbol{\Sigma}) = n$ e $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^t$, então a variável aleatória dada pela forma quadrática $q = \mathbf{X}^t \mathbf{A} \mathbf{X}$, sendo \mathbf{A} uma matriz simétrica, tem a mesma distribuição que a variável aleatória

$$W = \sum_{i=1}^n d_i Q_i$$

em que d_i são os autovalores de $\mathbf{A}\boldsymbol{\Sigma}$ e Q_1, \dots, Q_n são variáveis aleatórias independentes com distribuição $\chi_1^2(\delta_i^2)$, sendo os parâmetros de não-centralidade das mesmas dados por $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^t = \mathbf{P}^t \mathbf{L}^{-1} \boldsymbol{\mu}$, com \mathbf{P} sendo a matriz ortogonal dos autovetores de $\mathbf{A}\boldsymbol{\Sigma}$.

Provost e Rudiuk (1996) ainda apresentaram a função densidade e a função distribuição de W , mas o cálculo de valores como $P(W < w)$ é impraticável, pois tanto a função densidade quanto a distribuição dependem de somas infinitas. Vários autores construíram diferentes aproximações para $P(W < w)$, e algumas delas estão disponíveis em R no pacote **CompQuadForm**. Para saber mais sobre este pacote vide Duchesne e Micheaux (2010).

Deste modo, uma alternativa razoável seria trabalhar com uma estatística que meça a distância entre duas curvas e que seja uma forma quadrática do tipo q . Assim, consideremos a distância quadrática integrada (DQI) entre as curvas f e g dada por

$$DQI(f, g) = \int [f(u) - g(u)]^2 du.$$

No contexto discreto, em que as curvas são representadas por vetores de valores, a DQI é definida por

$$DQI(f, g) = \sum_{i=1}^n (f_i - g_i)^2 = (\mathbf{f} - \mathbf{g})^t (\mathbf{f} - \mathbf{g}).$$

De (5.2.2) pode-se deduzir que

$$(\mathbf{f} - \hat{\mathbf{f}}) \sim N \left[(\mathbf{I} - \mathbf{H})\mathbf{f}, \sigma^2 \mathbf{H}\mathbf{H}^t \right],$$

e que

$$DQI(f, \hat{f}) = (\mathbf{f} - \hat{\mathbf{f}})^t (\mathbf{f} - \hat{\mathbf{f}}) \sim \sum_{i=1}^n d_i Q_i \equiv W.$$

sendo os parâmetros de não-centralidade dados por $\delta_i = \left[\mathbf{P}^t \mathbf{L}^{-1} (\mathbf{I} - \mathbf{H}) \mathbf{f} \right]_i$, em que a matriz de covariâncias de $\hat{\mathbf{f}}$ pode ser escrita como $\Sigma_{\hat{f}} = \sigma^2 \mathbf{H}\mathbf{H}^t = \mathbf{L}\mathbf{L}^t$, \mathbf{P} é uma matriz ortogonal dos autovetores de $\Sigma_{\hat{f}}$ e d_i é o i -ésimo autovalor desta.

Considerando novamente o teste da hipótese $H_0 : f = f_0$, a estatística $DQI(f_0, \hat{f})$ tem a mesma distribuição que W se H_0 for verdadeiro. Portanto, se H_0 for falsa espera-se que a estatística de teste seja “grande”, o que indica que a estimativa não paramétrica não é próxima de f_0 , que deve implicar em

$$DQI(f_0, \hat{f}) > W_\alpha,$$

em que W_α é o quantil de interesse para o nível de significância α , que pode ser estimado utilizando o pacote **CompQuadForm**.

Observe que a distribuição da estatística do teste depende do valor de σ^2 , mas esse valor em geral é desconhecido. Espera-se que o uso do estimador

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \hat{\mathbf{y}})^t (\mathbf{y} - \hat{\mathbf{y}})}{n - \text{tr}(\mathbf{H})}$$

não afete fortemente o valor- p obtido no teste. Nas simulações este estimador é utilizado e o poder do teste é estimado em diferentes cenários.

5.3 Simulações

Nesta seção é examinado, em um estudo de simulação, o poder estimado do teste proposto (teste DQI). Assim como na simulação para o teste FBST do Capítulo 3, o modelo usado neste estudo é

$$y_i = x_{i1} + x_{i2} + c \sin(2x_{i2}) + \epsilon_i, \quad i = 1, \dots, n,$$

em que $\epsilon \sim N(0; 0, 01)$. No Capítulo 3 já foi mostrado como a constante c afeta a forma da função. Dependendo do valor usado para c , a parte não linear da função fica mais aparente.

Novamente, as hipóteses consideradas são

$$H_0 : f(X_1, X_2) = X_1 + X_2$$

e

$$H_1 : f(X_1, X_2) \neq X_1 + X_2.$$

São considerados os mesmos cenários: $n = 30, 50, 100$ e $c = 0, 0, 5, 1, 1, 5$. Em cada cenário, primeiramente é gerado um conjunto de n valores de uma distribuição uniforme no intervalo $[-6, 6]$ para covariável X_1 e o mesmo, de forma independente para X_2 . Em seguida, os valores de Y são obtidos segundo a função $y_i = x_{i1} + x_{i2} + c \sin(2x_{i2}) + \epsilon_i$. Daí os valores das estatísticas de teste são obtidos. Este processo é replicado 200 vezes para cada cenário (os valores das covariáveis são fixos para cada replicação). Então, define-se a proporção de vezes em que H_0 é rejeitada como sendo o poder estimado do teste.

A Tabela 5.1 mostra o poder estimado do teste para cada cenário considerando um nível de significância de 5%. Entre parênteses está o poder estimado no caso de usarmos $\hat{\sigma}^2$ no cálculo da estatística de teste. Os valores fora dos parênteses são referentes ao uso de σ na obtenção da estatística.

Observe que quando $c \neq 0$, mesmo para amostras pequenas, o poder estimado do teste é bem alto. Por outro lado, para $c = 0$ deveríamos ter algo em torno 5% de rejeições, sendo este o nível de significância do teste. Nos cenários em que usamos σ , o nível de significância foi próximo ao esperado, exceto para $n = 200$. Entretanto, quando utilizou-se a estimativa $\hat{\sigma}^2$, o número de rejeições foi maior que o esperado, mas aproxima-se de 5% à medida que o tamanho

Tabela 5.1: Taxa de rejeição de H_0 para o teste DQI de igualdade de curvas

n	$c = 0$	$c = 0,5$	$c = 1$	$c = 1,5$
30	0,040 (0,175)	1,000 (1,000)	1,000 (1,000)	1,000 (1,000)
50	0,045 (0,125)	1,000 (1,000)	1,000 (1,000)	1,000 (1,000)
100	0,055 (0,090)	1,000 (1,000)	1,000 (1,000)	1,000 (1,000)
200	0,080 (0,070)	1,000 (1,000)	1,000 (1,000)	1,000 (1,000)

da amostra cresce. Isso indica que para amostras pequenas devemos ter cautela ao realizar o teste DQI , principalmente nos casos em que a hipótese H_0 for rejeitada. Para uma investigação mais adequada o número de replicações deveria ser maior e outros valores para o nível de significância deveriam ser considerados.

Capítulo 6

Variáveis Resposta com Restrições

Neste capítulo é proposta uma abordagem para tratar variáveis resposta com restrições, como por exemplo variáveis binárias ou contagens. Esta abordagem não assume uma distribuição específica para a variável resposta, diferente dos modelos aditivos generalizados (GAM) propostos por Hastie e Tibshirani (1990) que tratam de variáveis da família exponencial de distribuições. São realizados estudos de simulação para comparar os resultados obtidos pela abordagem proposta com os conseguidos pelos pacotes `gamlss` e `mgcv` em R, que consideram a abordagem GAM.

6.1 Introdução

Em várias situações a variável resposta em estudo não pode ser tratada pela modelagem descrita até aqui. Alguns exemplos são:

- quando a resposta é restrita entre dois valores, $Y \in (a, b)$;
- quando a resposta assume apenas valores não negativos, $Y \geq 0$; ou
- quando a resposta é do tipo binário, $Y \in \{0, 1\}$.

No primeiro caso, em geral, tem-se como resposta uma fração ou proporção que está entre 0 e 1. Muito frequentemente trata-se este tipo de variável como vinda de uma distribuição beta. No

caso de respostas binárias utiliza-se a distribuição Bernoulli. E para Y não negativa adota-se a distribuição Poisson, gama, log-normal, ou outra distribuição com suporte não negativo.

A estrutura básica dos modelos aditivos generalizados (GAM) proposta por Hastie e Tibshirani (1990) é

$$r(\mu) = f(X_1, \dots, X_p) = \alpha + \sum_{j=1}^p f_j(X_j)$$

sendo $\mu_i = E(Y_i|X_1, \dots, X_p)$, r uma função de ligação adequada e com Y seguindo uma distribuição pertencente à família exponencial de distribuições. A estimação dos parâmetros do modelo é feita maximizando a função log-verossimilhança penalizada

$$l_p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) = l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) - \frac{1}{2} \sum_{j=1}^p \lambda_j \int [f_j''(t)]^2 dt,$$

sendo $l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$ a função log-verossimilhança do modelo para Y .

6.2 Modelos Aditivos Generalizados Livres de Distribuição

A escolha de uma distribuição para a variável resposta pode não ser simples. Além disso, faz com que a abordagem não seja totalmente não paramétrica. Vale lembrar que um dos principais objetivos deste trabalho é justamente fornecer uma ferramenta (o teste de hipóteses) para escolha de uma forma funcional e/ou distribuição mais adequada. Assim, propomos uma estratégia para estimar μ de modo que não seja necessário atribuir uma distribuição para a variável resposta. Denominaremos esta abordagem proposta por modelo aditivo generalizado livre de distribuição (DFGAM).

De modo geral, assim como nos modelos GAM de Hastie e Tibshirani (1990), a abordagem DFGAM considera

$$\boldsymbol{\mu} = r^{-1} \left(\alpha + \sum_{j=1}^p f_j(x_j) \right) = r^{-1}(\mathbf{X}\boldsymbol{\theta}),$$

em que novamente $\boldsymbol{\theta} = (\alpha, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)^t$, $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_p)$ e \mathbf{X}_j é uma matrix $n \times K_j$ contendo os valores das K_j funções base calculadas em \mathbf{x}_j . Para respostas entre 0 e 1 ou para respostas

dicotômicas usa-se o logito como função de ligação, ou seja,

$$\boldsymbol{\mu} = \frac{\exp(\mathbf{X}\boldsymbol{\theta})}{1 + \exp(\mathbf{X}\boldsymbol{\theta})}.$$

Para $Y \geq 0$ adota-se a ligação logarítmica

$$\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\theta}).$$

A estimação dos parâmetros é feita minimizando-se

$$L_p(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^t(\mathbf{y} - \boldsymbol{\mu}) + \frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{\Lambda}\boldsymbol{\theta}.$$

Observe que a estimação é feita em uma abordagem livre de distribuições. Outras funções de ligação podem ser utilizadas. Para uma variável resposta sem restrições, a função de ligação adotada é a identidade, ou seja, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\theta}$. Observe que dessa maneira, $L_p(\boldsymbol{\theta})$ é exatamente a mesma utilizada no método SHS descrito no Capítulo 4. Sendo assim, pode-se afirmar que no Capítulo 4 foi considerado o modelo DFGAM com função de ligação identidade. Neste trabalho, a estimação do modelo DFGAM sempre é feita utilizando o procedimento SHS. Deste modo, toda vez que for mencionada a utilização do método SHS estará implícito que foi considerado o modelo DFGAM.

A estimação de $\boldsymbol{\theta}$ pode ser feita através do método escore de Fisher. Para tanto, obtém-se

$$\frac{\partial L_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}(\mathbf{y} - \boldsymbol{\mu}) + \boldsymbol{\Lambda}\boldsymbol{\theta} = -\mathbf{X}^t\mathbf{V}_1(\mathbf{y} - \boldsymbol{\mu}) + \boldsymbol{\Lambda}\boldsymbol{\theta}$$

e

$$E \left[\frac{\partial^2 L_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \right] = \mathbf{X}^t\mathbf{V}_2\mathbf{X} + \boldsymbol{\Lambda}. \quad (6.2.1)$$

A Tabela 6.1 mostra as formas das matrizes \mathbf{V}_1 e \mathbf{V}_2 para as funções ligação logística e logarítmica.

Tabela 6.1: Formas das matrizes \mathbf{V}_1 e \mathbf{V}_2 para diferentes restrições em Y e diferentes funções de ligação.

Resposta	Ligação	\mathbf{V}_1	\mathbf{V}_2
$Y \geq 0$	log	$\text{diag}(\mu_i)$	$\text{diag}(\mu_i^2)$
$Y \in (0, 1)$ ou $Y \in \{0, 1\}$	logito	$\text{diag}[\mu_i(1 - \mu_i)]$	$\text{diag} \left\{ [\mu_i(1 - \mu_i)]^2 \right\}$

Dessa maneira, para uma certa iteração m no processo iterativo do método escore de Fisher, a estimativa de $\boldsymbol{\theta}$ da iteração seguinte é dada por

$$\begin{aligned}\boldsymbol{\theta}^{(m+1)} &= \boldsymbol{\theta}^{(m)} - E \left[\frac{\partial^2 L_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \right]^{-1} \frac{\partial L_p}{\partial \boldsymbol{\theta}} \\ &= \boldsymbol{\theta}^{(m)} + (\mathbf{X}^t \mathbf{V}_2^{(m)} \mathbf{X} + \boldsymbol{\Lambda})^{-1} [\mathbf{X}^t \mathbf{V}_1^{(m)} (\mathbf{y} - \boldsymbol{\mu}^{(m)}) - \boldsymbol{\Lambda} \boldsymbol{\theta}^{(m)}].\end{aligned}\quad (6.2.2)$$

Assim como no modelo GAM, no modelo DFGAM podemos reescrever o método escore de Fisher como sendo um procedimento de mínimos quadrados iterativamente reponderados (IRLS). A equação (6.2.2) é reescrita como

$$\boldsymbol{\theta}^{(m+1)} = (\mathbf{X}^t \mathbf{V}_2^{(m)} \mathbf{X} + \boldsymbol{\Lambda})^{-1} \mathbf{X}^t \mathbf{V}_2^{(m)} \left[\mathbf{X} \boldsymbol{\theta}^{(m)} + (\mathbf{V}_2^{(m)})^{-1} \mathbf{V}_1^{(m)} (\mathbf{y} - \boldsymbol{\mu}^{(m)}) \right]$$

No procedimento IRLS, dada a estimativa corrente $\boldsymbol{\theta}^{(m)}$ é criada uma pseudovariável resposta

$$\mathbf{z} = \mathbf{X} \boldsymbol{\theta}^{(m)} + (\mathbf{V}_2^{(m)})^{-1} \mathbf{V}_1^{(m)} (\mathbf{y} - \boldsymbol{\mu}^{(m)}),$$

e a nova estimativa é obtida ajustando-se o modelo para resposta \mathbf{z} usando $\mathbf{V}_2^{(m)}$ como matriz de pesos, ou seja,

$$\boldsymbol{\theta}^{(m+1)} = (\mathbf{X}^t \mathbf{V}_2^{(m)} \mathbf{X} + \boldsymbol{\Lambda})^{-1} \mathbf{X}^t \mathbf{V}_2^{(m)} \mathbf{z}.$$

E assim, iterativamente constrói-se novamente novas pseudovariáveis que são ajustadas usando pesos renovados. O procedimento é repetido até que um critério de parada seja satisfeito. O interessante do procedimento IRLS é que pode-se tomar

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{V}_2 \mathbf{X} + \boldsymbol{\Lambda})^{-1} \mathbf{X}^t \mathbf{V}_2$$

obtida da última iteração como sendo a matriz *hat*. A matriz \mathbf{H} é importante pois fornece os graus de liberdade do ajuste, além de ser utilizada na construção de estatísticas para testes de hipótese.

O procedimento de estimação é uma versão modificada do algoritmo sequencial descrito no Capítulo 4. O algoritmo para estimação dos parâmetros do modelo DFGAM é dado por

1. Escolha $\mathbf{K} = (K_1, \dots, K_p)^t$ como as quantidades iniciais de funções base e fixe $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)$.

Além disso, fixe $\boldsymbol{\theta}$ e conseqüentemente $\boldsymbol{\mu}, \mathbf{V}_1$ e \mathbf{V}_2 .

2. Encontre $\hat{\boldsymbol{\lambda}}$ que minimize

$$GCV(\hat{\boldsymbol{\lambda}}) = \frac{n[\mathbf{z} - \mathbf{H}(\boldsymbol{\lambda})\mathbf{z}]^t[\mathbf{z} - \mathbf{H}(\boldsymbol{\lambda})\mathbf{z}]}{n - \text{tr}[\mathbf{H}(\boldsymbol{\lambda})]^2}, \quad (6.2.3)$$

com $\mathbf{H}(\boldsymbol{\lambda}) = \mathbf{X}(\mathbf{X}^t\mathbf{V}_2\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}^t\mathbf{V}_2$.

3. Obtenha $\hat{\boldsymbol{\theta}}$ utilizando o método escore de Fisher descrito na Seção 6.2. Em seguida calcule $f^{\mathbf{K},\hat{\boldsymbol{\lambda}}} = \mathbf{X}\hat{\boldsymbol{\theta}}$, além dos $f_j^{K_j} = \mathbf{X}_j\hat{\boldsymbol{\theta}}_j$, para $j = 1, \dots, p$.
4. Acrescente uma base ao número de funções base de cada covariável e repita os passos de (2) a (3) para obter $f_1^{K_1+1}, \dots, f_p^{K_p+1}$.
5. Para um número real $\delta > 0$ e para $j = 1, \dots, p$, verifique se as distâncias $d_j(f_j^{K_j}, f_j^{K_j+1}) < \delta$. Caso existam j 's tais que $d_j > \delta$, acrescente uma função base apenas para estas covariáveis e repita os passos (2) e (3). Faça isso até que $d_j < \delta, \forall j$.

6.3 Dados Simulados: Exemplo 4

Para verificar o desempenho do método SHS na estimação do modelo DFGAM é usado um exemplo com dados simulados. Neste exemplo, cinco funções são consideradas para as covariáveis, são elas

- $f_1(x) = 0,4 \sin(\pi x)$,
- $f_2(x) = 0,2 \exp(2x)$,
- $f_3(x) = 0,04x^{11}[10(1-x)]^6 + 2(10x)^3(1-x)^{10}$,
- $f_4(x) = 0x$ e
- $f_5(x) = x \sin(20\pi x)$.

o exemplo 4 é dividido em três partes, exemplo 4.a, 4.b e 4.c. Em cada parte é usada uma restrição diferente para variável resposta. Serão considerados modelos com quatro covariáveis. O desempenho do método SHS, que estima o modelo DFGAM, é comparado com os resultados obtidos

nos procedimentos GAMLSS e MGCV, que utilizam a função de log-verossimilhança penalizada na estimação.

Até aqui, as estimativas obtidas foram comparadas usando as medidas DQI e cosseno. Geralmente estas medidas são referentes à f e \hat{f} . Entretanto, neste capítulo serão referentes à μ e $\hat{\mu}$, ou seja, as medidas serão $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$. Isso se deve ao fato de usarmos funções de ligação. Nestes casos, frequentemente estamos interessados em μ e não em f .

6.3.1 Exemplo 4.a: Resposta sem Restrições

Nesta parte do exemplo, os dados são obtidos da seguinte maneira:

1. Gerar $X_j \sim U(0, 1)$, $j = 1, \dots, 4$;
2. Calcular $\mu = f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4)$;
3. Gerar $Y \sim N(\mu; 0, 3)$;
4. Repetir o processo até obter $n = 50$ observações.

Como o objetivo é trabalhar com uma variável resposta sem restrições, uma distribuição normal foi considerada.

A Figura 6.1 mostra as curvas f_1 , f_2 , f_3 e f_4 referentes a este exemplo. Juntamente com as curvas verdadeiras estão traçadas as estimativas obtidas pelos métodos SHS, MGCV e GAMLSS. Observando a Figura 6.1 pode-se verificar que as três estimativas estão próximas às curvas verdadeiras.

A Tabela 6.2 mostra as medidas $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$ para cada método. Observe que o método SHS obteve um desempenho semelhante aos demais. O procedimento proposto parece ser competitivo neste cenário 4.a em que variável resposta não possui restrições e as funções a serem estimadas são suaves. A principal diferença deste exemplo em relação aos usados no Capítulo 4 é o uso de quatro covariáveis. Em 4.b e 4.c serão usadas variáveis resposta com restrições.

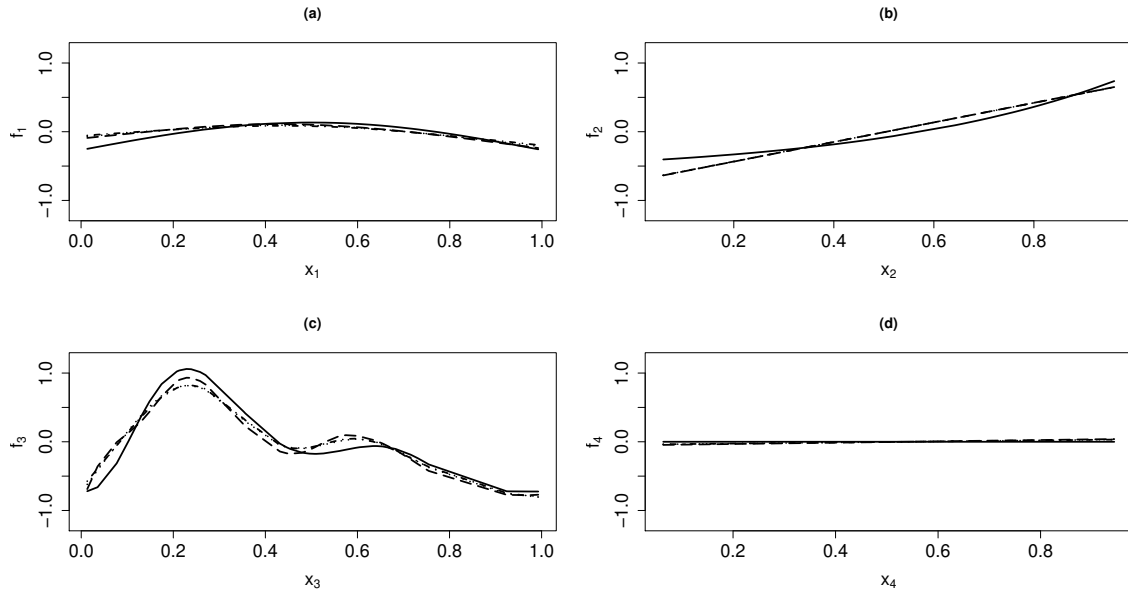


Figura 6.1: Funções f_1 , f_2 , f_3 e f_4 do exemplo 4.a traçadas em (a), (b), (c) e (d), respectivamente. Linha contínua para a função verdadeira, tracejada para o método SHS, pontilhada para GAMLSS e tracejada e pontilhada para MGCV.

Tabela 6.2: Comparação entre os métodos SHS, MGCV e GAMLSS para o número de funções base e as medidas $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$. Exemplo 4.a.

Procedimento	K_1	K_2	K_3	K_4	$\cos(\phi_{\mu, \hat{\mu}})$	$DQI(\mu, \hat{\mu})$
SHS	6	6	12	6	0,9958	1,2771
GAMLSS	13	13	13	13	0,9958	1,2903
MGCV	9	9	9	9	0,9957	1,3275

6.3.2 Exemplo 4.b: Resposta Binária

Neste exemplo considerou-se a distribuição Bernoulli para gerar a variável resposta. O algoritmo para gerar o conjunto de dados é decrito por

1. Gerar $X_j \sim U(0, 1)$, $j = 1 \dots, 4$;
2. Calcular $\mu = \text{logito}^{-1}\{1.7[-5 + f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4)]\}$, sendo que para qualquer valor real a define-se $\text{logito}^{-1}(a) = \exp(a)/[1 + \exp(a)]$;
3. Gerar $Y \sim \text{Bernoulli}(\mu)$;

4. Repetir o processo até obter $n = 400$ observações.

Observe que as funções utilizadas foram as mesmas do exemplo 4.a, e o número de observações passou para $n = 400$.

A Figura 6.2 mostra as curvas originais traçadas juntamente com as estimadas. Pode-se verificar que as três estimativas são próximas e estão próximas às curvas verdadeiras. Não parece haver muita diferença entre as estimativas dos diferentes métodos.

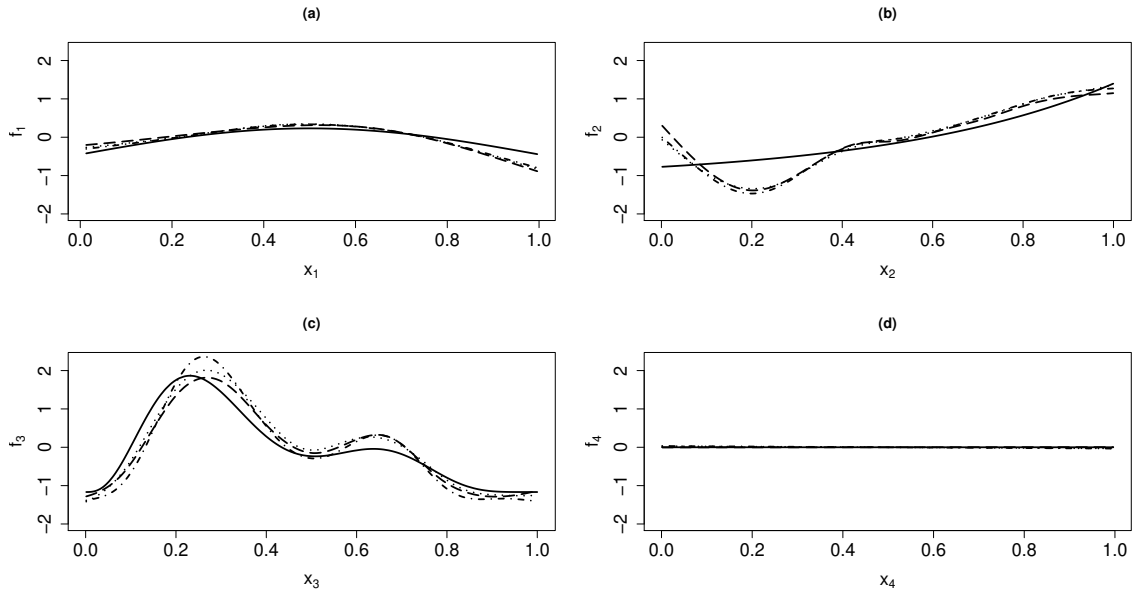


Figura 6.2: Funções f_1 , f_2 , f_3 e f_4 do exemplo 4.b traçadas em (a), (b), (c) e (d), respectivamente. Linha contínua para verdadeira, tracejada para o método SHS, pontilhada para GAMLSS e tracejada e pontilhada para MGCV.

A Tabela 6.3 apresenta as medidas $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$ para os métodos SHS, GAMLSS e MGCV. Analisando a tabela, podemos afirmar que os métodos obtiveram desempenhos semelhantes. Dessa forma, constatamos que o método SHS é competitivo em relação aos demais, mesmo sem utilizar a função de verossimilhança no processo de estimação, como fazem os métodos MGCV e GAMLSS.

Em 4.c é adicionada uma série de detalhes que devem trazer problemas para os procedimentos de estimação. O objetivo é analisar o desempenho do método proposto em um cenário menos favorável.

Tabela 6.3: Comparação entre os métodos SHS, MGCV e GAMLSS para o número de funções base e as medidas $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$. Exemplo 4.b.

Procedimento	K_1	K_2	K_3	K_4	$\cos(\phi_{\mu, \hat{\mu}})$	$DQI(\mu, \hat{\mu})$
SHS	6	13	12	6	0,9913	6,4280
GAMLSS	23	23	23	23	0,9919	6,7729
MGCV	9	9	9	9	0,9906	7,4346

6.3.3 Exemplo 4.c: Resposta em Contagem

Neste exemplo considerou-se a distribuição Poisson para gerar a variável resposta. Os dados são obtidos segundo o seguinte procedimento.

1. Gerar $X_1 \sim U(0; 1)$, $X_2 = 0,7(1 - X_1) + \epsilon_{X_2}$, $X_4 \sim U(0; 1)$ e $X_3 = 0,8(1 - X_4) + \epsilon_{X_3}$, com $\epsilon_{X_2} \sim U(0; 0,3)$ e $\epsilon_{X_3} \sim U(0; 0,2)$;
2. Calcular $\mu = \exp[f_1(X_1) + f_2(X_2) + f_3(X_3) + f_5(X_4)]$;
3. Gerar $Y \sim Poisson(\mu)$;
4. Repetir o processo até obter $n = 400$ observações.

A principal diferença desta parte do exemplo 4 para as demais é que as covariáveis X_1 e X_2 são correlacionadas, assim como X_3 e X_4 . Além disso, a função f_5 é utilizada com a covariável X_4 . Esta é uma função muito estruturada, em geral difícil de estimar.

A Figura 6.3 apresenta as funções utilizadas no exemplo 4.c juntamente com suas respectivas estimativas, obtidas pelos métodos SHS, MGCV e GAMLSS. Os procedimentos GAMLSS e MGCV mais uma vez não conseguiram reproduzir a função com uma curvatura total mais elevada que as demais. Como já foi mencionado antes, isso provavelmente ocorre devido aos número insuficiente de bases usadas para estimar f_5 .

A Tabela 6.4 traz as medidas de comparação assim como o número de funções base usadas para estimar as curvas. Observe que o método SHS utiliza poucas bases para estimar f_1 e f_2 que são curvas mais suaves, uma quantidade moderada para estimar f_3 e muitas funções base para obter a estimativa de f_5 . O ganho obtido com essa ponderação na escolha das bases pode ser

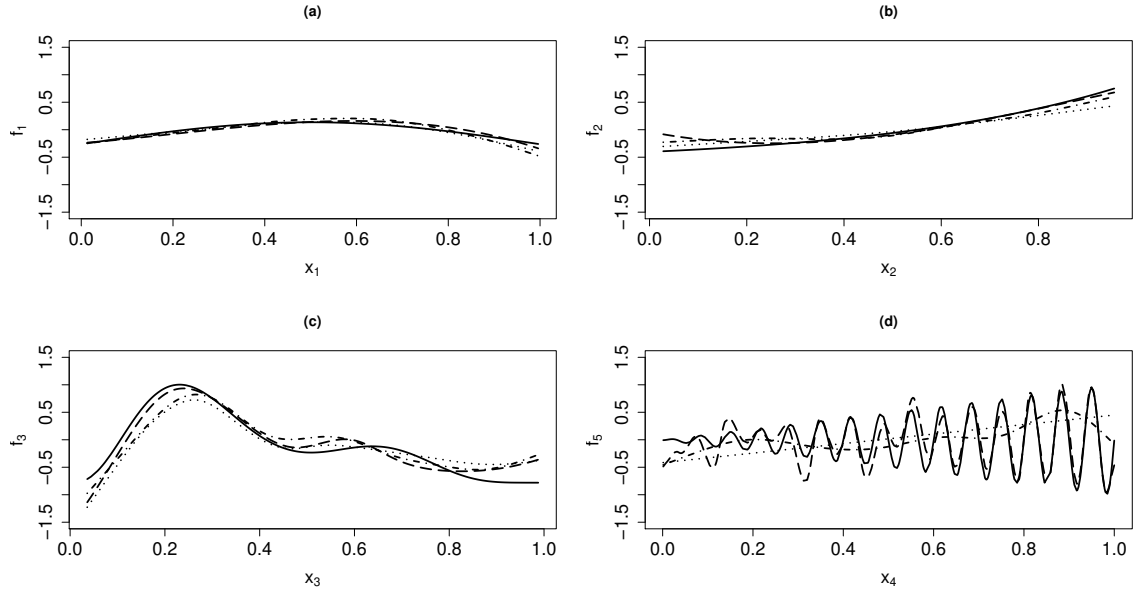


Figura 6.3: Funções f_1 , f_2 , f_3 e f_5 do exemplo 4.c traçadas em (a), (b), (c) e (d) respectivamente. Linha contínua para verdadeira, tracejada para o método SHS, pontilhada para GAMLSS e tracejada e pontilhada para MGCV.

Tabela 6.4: Comparação entre os métodos SHS, MGCV e GAMLSS para o número de funções base e as medidas $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$. Exemplo 4.c.

Procedimento	K_1	K_2	K_3	K_4	$\cos(\phi_{\mu, \hat{\mu}})$	$DQI(\mu, \hat{\mu})$
SHS	4	4	10	55	0,9915	730,26
GAMLSS	23	23	23	23	0,8901	8104,74
MGCV	9	9	9	9	0,8994	7454,85

visto nas medidas de comparação mostradas na Tabela 6.4. Nas duas medidas o método SHS é consideravelmente melhor que os demais.

Assim como no Capítulo 4, os valores $K_1=4$, $K_2=4$, $K_3=10$ e $K_4=55$ indicados pelo método SHS são usados nos procedimentos MGCV e GAMLSS. A Tabela 6.5 mostra os resultados obtidos e faz uma segunda comparação. Observe que os procedimentos GAMLSS e MGCV melhoram muito seus desempenhos, o que mostra que a escolha de valores adequados para o número de bases é um fator muito importante na estimação do modelo. Mas apesar de melhorar seus desempenhos, os resultados obtidos pelos métodos GAMLSS e MGCV ainda estão abaixo dos obtidos pelo SHS.

Tabela 6.5: Comparação de medidas $\cos(\phi_{\mu, \hat{\mu}})$ e $DQI(\mu, \hat{\mu})$ obtidas pelos métodos SHS, MGCV e GAMLSS, com todos utilizando $K_1 = 4$, $K_2 = 4$, $K_3 = 10$ e $K_4 = 55$. Exemplo 4.c.

Procedimento	$\cos(\phi_{\mu, \hat{\mu}})$	$DQI(\mu, \hat{\mu})$
SHS	0,9915	730,26
GAMLSS	0,9906	887,59
MGCV	0,9900	935,09

Reforçamos que o método SHS fornece de modo automático as quantidades de funções base a serem utilizadas para cada covariável. Nos pacotes em R citados neste trabalho a obtenção de valores adequados para K_1, K_2, \dots, K_p pode demandar muito tempo, sobretudo quando o número de covariáveis é grande.

Reforço ainda que, ao contrário dos demais métodos, o procedimento SHS utiliza o modelo DFGAM, obtendo suas estimativas sem considerar uma função de verossimilhança no processo de estimação. Em muitas situações a distribuição de Y não é conhecida. Quando isso ocorre a estimativa dos métodos MGCV e GAMLSS (e outros que considerem a função de verossimilhança) pode ser severamente prejudicada em caso de uma escolha errada para a distribuição de Y .

Capítulo 7

Aplicação a Dados Reais

Para ilustrar a aplicação dos métodos de estimação e dos testes apresentados, é usado um conjunto de dados que se refere ao trabalho de Tanaka e Nishii (2009) sobre o desmatamento no leste da Ásia. No trabalho os autores consideram como principais causas de desmatamento em uma área (variável dependente) a população humana no local (preditor 1) e o declive do terreno (preditor 2). Vários modelos são considerados pelos autores. São tratados vários conjuntos de dados referentes ao leste da Ásia, entretanto dar-se-á atenção especial aos dados referentes a Hiroshima (Hiroshima data). São 8538 sítios observados em Hiroshima, dados cedidos pela prefeitura local, cada sítio tem área de 1 km². O objetivo deste capítulo é estimar a relação entre a variável resposta e os preditores utilizando os métodos BHS e SHS. Além disso, os teste propostos neste trabalho são usados para escolher um entre os diversos modelos são considerados pelos autores.

7.1 Descrição dos Dados

Considere $N = N(s)$ a população e $R = R(s)$ o declive do relevo no sítio s . Este declive de relevo é a diferença entre as altitudes máxima e mínima do sítio. Considere ainda $F = F(s)$ como sendo a proporção do local coberta por floresta. É considerado um modelo aditivo da forma

$$F = g(N) + h(R) + \epsilon. \quad (7.1.1)$$

Para contornar problemas encontrados na estimação, os autores transformaram a variável F em Y de forma que

$$Y = \log \left(\frac{F + 0,5}{1 - F + 0,5} \right).$$

Assim, o modelo estudado foi

$$Y = g(N) + h(R) + \epsilon$$

sendo $\epsilon \sim N(0, \sigma^2)$, e $g(\cdot)$ e $h(\cdot)$ funções não lineares dadas na Tabela 7.1.

Tabela 7.1: Formas funcionais para o modelo de regressão

$g_1(N)$	=	$-\beta \log(N + 1)$
$g_2(N)$	=	$-\frac{\beta_1}{1 + \exp\{\beta_2 - \beta_3 \log(N + 1)\}} + \frac{\beta_1}{1 + \exp(\beta_2)}$
$g_3(N)$	=	$\beta_1 \{1 - \exp[\beta_2 \log(N + 1)]\}$
$h_1(R)$	=	$I(R > \gamma_1) \gamma_2 \log(R - \gamma_1 + 1)$
$h_2(R)$	=	$I(R > \gamma_1) \gamma_2 \log \left(\frac{R}{\gamma_1} \right)$
$h_3(R)$	=	$\gamma_3 \exp(-\gamma_1 e^{-\gamma_2 R}) - \gamma_3 \exp(-\gamma_1)$
$h_4(R)$	=	$\frac{\gamma_1}{1 + \exp(\gamma_2 - \gamma_3 R)} - \frac{\gamma_1}{1 + \exp(\gamma_2)}$

No trabalho de Tanaka e Nishii (2009) foram avaliadas todas as possíveis combinações de modelos da Tabela 7.1, isto é,

$$Y = \alpha + g_k(N) + h_l(R) + \epsilon$$

em que $k = 1, 2, 3$, $l = 1, 2, 3, 4$ e α é o intercepto. Foram considerados modelos com erros independentes e com dependência espacial. Neste trabalho são considerados apenas modelos com erros independentes.

Para dados de Hiroshima, segundo o critério AIC, os modelos com melhor desempenho são $g_2 + h_3$, $g_2 + h_4$ e $g_2 + h_2$, respectivamente. A Tabela 7.2 traz o desempenho de todos os modelos segundo o critério AIC. É somada a constante 35857,64 ao valor do critério para facilitar a visualização.

Observe que os modelos que apresentam pior desempenho são os que consideram g_1 , dessa forma, estes serão descartados. Assim, os modelos que são comparados aqui, usando os critérios de seleção apresentados são $g_2 + h_1$, $g_2 + h_2$, $g_2 + h_3$, $g_2 + h_4$, $g_3 + h_1$, $g_3 + h_2$, $g_3 + h_3$ e $g_3 + h_4$.

Tabela 7.2: Valores $AIC + 35857,64$ para modelos $Y = \beta_0 + g(N) + h(R)$.

Modelos	h_1	h_2	h_3	h_4
g_1	634,78	709,87	595,11	622,03
g_2	104,14	53,44	0,00	21,26
g_3	221,15	233,43	161,73	187,56

Os modelos destacados são comparados segundo os métodos descritos nos Capítulos 3 e 5.

7.2 Abordagem Bayesiana

Nesta seção, o método BHS é usado para estimar a relação entre a variável resposta e as covariáveis. Antes da apresentação os resultados, deve ser feita uma consideração com respeito à estimação. Existem vários valores empatados, tanto na covariável N quanto para R . Um exemplo é o fato de haver vários sítios totalmente planos, ou seja, com declive igual a 0. Estes empates podem causar problemas na seleção dos dados para locação dos nós, pois não podem haver dois nós com o mesmo valor. Para eliminar os empates acrescentou-se um pequeno ruído ($\xi \sim U(-0.1, 0.1)$) aos valores empatados nas covariáveis. Como nem os declives e nem as populações podem ser negativos, para $R = 0$ e $N = 0$ foi adicionado $|\xi|$. Tendo considerado esta correção, foram amostradas 1000 curvas e então obtidas a curva estimada e os limites bayesianos. A Figura 7.1 mostra estas estimativas separadamente. Informações sobre as distribuições marginais *a posteriori* para λ_g , λ_h , σ^2 e α podem ser vistas na Tabela 7.3. As modas *a posteriori* para K_g e K_h são respectivamente 13 e 15.

Tabela 7.3: Estimativa e limites bayesianos para as marginais *a posteriori* do parâmetros λ_g , λ_h , σ^2 e α para os dados de Hiroshima.

Parâmetro	LI 2,5%	Média	LS 97,5%
α	0,6112	0,6176	0,6243
σ^2	0,0822	0,0847	0,0871
λ_g	0,2107	0,5259	0,9725
λ_h	0,2248	0,5068	0,9093

A análise poderia terminar aqui, pois a partir destes resultados já teríamos uma estimativa

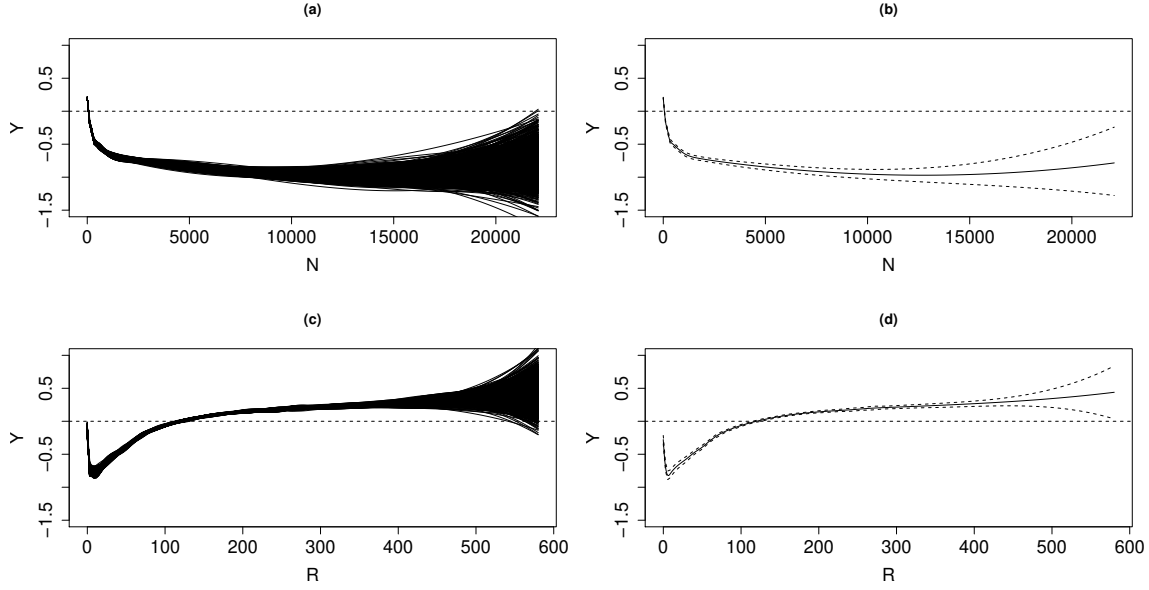


Figura 7.1: Em (a) as 1000 curvas *a posteriori* estimadas para a função g , e em (b) a curva média com os limites bayesianos. Em (c) e (d) o mesmo que em (a) e (b) respectivamente em relação à função h .

para a relação entre a variável resposta Y e as covariáveis N e R . Mas como o objetivo dos autores era escolher um dos vários modelos candidatos, usaremos essa amostra de curvas *a posteriori* para apontar qual dos modelos candidatos é o mais adequado.

7.2.1 Teste Bayesiano

Será aplicado ao conjunto de dados de Hiroshima o procedimento baseado na evidência bayesiana a favor de H_0 apresentado no Capítulo 3. A Tabela 7.4 mostra a evidência a favor de cada modelo candidato $g_k + h_l$, para $k = 1, 2$ e $l = 1, 2, 3, 4$. Os modelos com uma maior evidência a seu favor são $g_2 + h_3$ e $g_2 + h_4$. Esses modelos se destacaram em relação aos demais, e estão praticamente empatados em relação à evidência. Considerando apenas a evidência a favor de H_0 , o modelo $g_2 + h_4$ deve ser escolhido. Mas se considerarmos os modelos $g_2 + h_3$ e $g_2 + h_4$ como empatados, deveríamos escolher o modelo $g_2 + h_3$ pois $\cos(\phi_{g_2+h_3, \hat{f}}) > \cos(\phi_{g_2+h_4, \hat{f}})$ e $DQI(g_2 + h_3, \hat{f}) < DQI(g_2 + h_4, \hat{f})$.

Vejamos como são os resultados usando o modelo DFGAM e em seguida o teste DQI .

Tabela 7.4: Resultados dos testes de hipóteses bayesianos do tipo $H_0 : f = g_k + h_l$ contra $H_1 : f \neq g_k + h_l$, para $k = 1, 2$ e $l = 1, 2, 3, 4$.

Modelo	$EV(g + h \mathbf{y}, \mathbf{x})$
$g_2 + h_1$	0,1422
$g_2 + h_2$	0,1474
$g_2 + h_3$	0,2344
$g_2 + h_4$	0,2364
$g_3 + h_1$	0,1169
$g_3 + h_2$	0,1061
$g_3 + h_3$	0,1991
$g_3 + h_4$	0,1907

7.3 Abordagem Sequencial

Assim como na abordagem bayesiana, foram eliminados os empates nas covariáveis para que não houvessem problemas no posicionamento dos nós. O procedimento SHS foi utilizado para obtenção da superfície estimada. O número de funções base para cada covariável foi $K_g = 17$ e $K_h = 24$, e os parâmetros de suavização foram $\lambda_g = 3,8685 \times 10^{-7}$ e $\lambda_h = 6,3780 \times 10^4$. Observe que foram necessárias muitas funções base para estimar a função h . Mas, como forma de compensação, a estimativa obtida para λ_h foi bastante elevada, o que ajudou a controlar a suavização. Por outro lado, o valor de λ_g mostra que praticamente não houve penalização na estimação da função g . A Figura 7.2 mostra as curvas estimadas para g e h .

Analisando as Figuras 7.1 e 7.2, pode-se observar que as estimativas obtidas pelos métodos BHS e SHS são similares.

Novamente, poderíamos apenas obter uma estimativa para relação entre a variável resposta Y e as covariáveis N e R . Mas como os autores desejam escolher um entre os vários modelos candidatos, usaremos o teste DQI para apontar qual deles é o mais adequado. Na próxima seção, as estimativas obtidas pelo método SHS serão utilizadas no teste DQI .

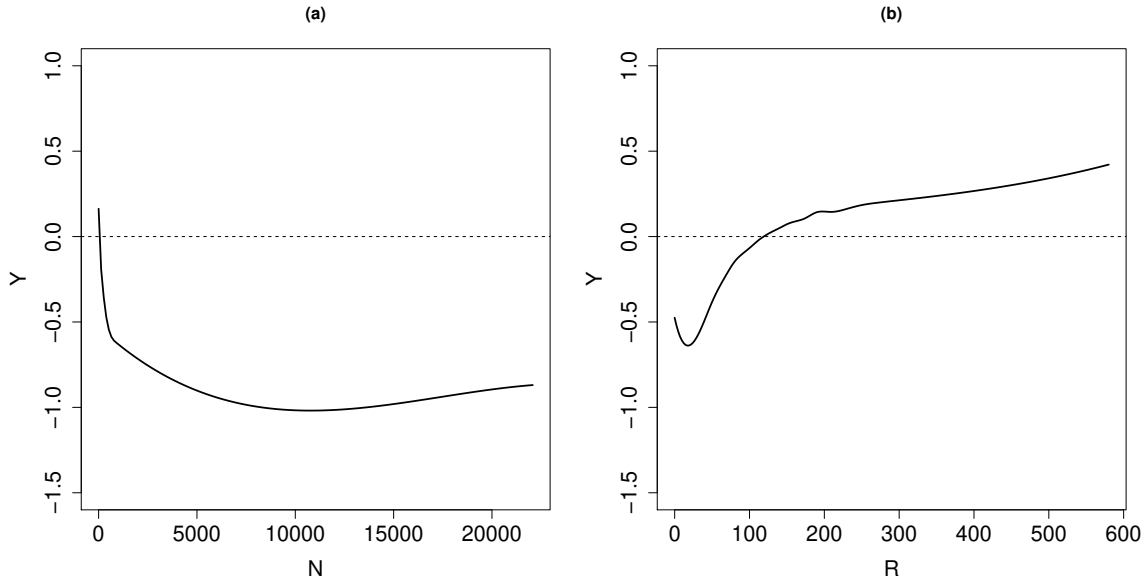


Figura 7.2: Em (a) a estimativa para função $g(N)$ e em (b) a estimativa para $h(R)$.

7.3.1 Teste DQI

Agora é considerado o procedimento descrito no Capítulo 5, o teste DQI . Como o tamanho da amostra é grande, 8538 observações, o valor- p do teste DQI é ajustado usando a proposta de Good (1982). O autor propõe o seguinte valor- p ajustado chamado valor- q :

$$\text{valor-}q = \text{valor-}p \sqrt{\frac{n}{100}}.$$

Good (1982) baseou sua proposta em Jeffreys (1939), que mostrou que o fator de Bayes contra a hipótese nula para um dado valor- p é inversamente proporcional à raiz quadrada do tamanho da amostra. Mais formalmente, esse valor- p ajustado é definido por

$$\text{valor-}q = \min \left(0, 5; \text{valor-}p \sqrt{\frac{n}{100}} \right),$$

para que o valor- q não ultrapasse 1. A interpretação de Good (1982) para esse valor seria “a evidência a favor de H_0 como se esta fosse obtida a partir de uma amostra de tamanho 100”.

As quantidades valor- p e valor- q obtidas nos testes para cada modelo candidato são dadas na Tabela 7.5.

Tabela 7.5: Resultados dos testes DQI de hipóteses do tipo $H_0 : f = g_k + h_l$ contra $H_1 : f \neq g_k + h_l$, para $k = 1, 2$ e $l = 1, 2, 3, 4$.

Modelo	valor- p	valor- q
$g_2 + h_1$	<0,001	<0,001
$g_2 + h_2$	0,0038	0,0356
$g_2 + h_3$	0,0236	0,2180
$g_2 + h_4$	0,0119	0,1099
$g_3 + h_1$	<0,001	<0,001
$g_3 + h_2$	<0,001	<0,001
$g_3 + h_3$	<0,001	<0,001
$g_3 + h_4$	<0,001	<0,001

Os únicos modelos considerados adequados, observando o valor- q , foram $g_2 + h_3$ e $g_2 + h_4$. Portanto, ambos os modelos podem ser usados. Observe que os dois modelos apresentam o mesmo número de parâmetros. Com isso, para escolher apenas um modelo recomenda-se utilizar medidas como valor- q , o cosseno e a DQI . Nas três medidas o desempenho do modelo $g_2 + h_3$ é superior ao $g_2 + h_4$. Portanto, na situação de escolher apenas um modelo, o candidato escolhido seria o modelo $g_2 + h_3$.

7.4 Análise dos Resultados

Em seu trabalho, Tanaka e Nishii (2009) analisaram 20 modelos com erros independentes para os dados de Hiroshima. Para selecionar o modelo mais adequado para este conjunto de dados os autores utilizaram a medida AIC . Segundo este critério, o modelo apontado como sendo o mais adequado é $g_2 + h_3$.

Neste trabalho foram analisados oito destes modelos. Foram utilizadas os procedimentos DF-GAM e BHS para estimar o modelo $Y = f(R, N) = g(R) + h(N)$. Baseados nas estimativas foram aplicados os testes DQI e FBST para igualdade de curvas. Para o teste bayesiano, considerando apenas a evidência a favor de H_0 , o modelo escolhido é $g_2 + h_4$. Mas como a evidência a favor dos modelos $g_2 + h_3$ e $g_2 + h_4$ é muito próxima, optou-se por observar as medidas do cosseno e DQI para desempate. Neste caso, o modelo escolhido é $g_2 + h_3$. Para o teste DQI , ocorreu algo

parecido, em que $g_2 + h_3$ e $g_2 + h_4$ foram considerados os únicos modelos adequados, mas o modelo escolhido foi $g_2 + h_3$.

Em resumo, tanto o critério AIC quanto os testes propostos, apontaram o modelo $g_2 + h_3$ como sendo o mais adequado.

Capítulo 8

Considerações Finais

8.1 Conclusões

Foram propostos dois novos procedimentos de estimações para modelos aditivos. Ambos são generalizações de procedimentos de regressão *h-splines*, um usando uma abordagem adaptativo sequencial e outro usando uma abordagem bayesiana. Usando dados simulados pôde-se verificar que os dois obtiveram estimativas de superfície próximas às verdadeiras. Além disso, a cadeia gerada pelo método BHS foi iniciada para diferentes valores de K . Ora com K_1 e K_2 grandes, hora com eles pequenos, com K_1 grande e K_2 pequeno e o contrário, e em todas as situações a cadeia convergiu para a mesma configuração. O procedimento bayesiano obtém estimativas mais suaves, mesmo em situações em que utiliza muitas funções base. Isso ocorre porque a estimativa final do método BHS é uma média de várias curvas *a posteriori*. O procedimento SHS é mais rápido do que o BHS. O método SHS demorou poucos segundos para estimar a superfície para os dados com ou sem restrição (demorando um pouco mais para dados com restrição). Nas superfícies consideradas na abordagem bayesiana, o método BHS demorou 12 minutos para obter 1000 amostras da distribuição *a posteriori*, sendo usado um descarte de 4000 amostras. Portanto, em uma situação em que seja necessário estimar muitos modelos, o procedimento SHS é mais indicado.

Como o método SHS é mais rápido, recomendamos o uso de suas estimativas como valores iniciais na cadeia do RJ-MCMC para o procedimento BHS. Em geral, os valores iniciais não inter-

ferem na configuração final da cadeia. Mas existem certas situações, como curvas extremamente estruturadas e/ou covariáveis muito correlacionadas, em que o uso das estimativas do método SHS como valores iniciais para BHS melhoram consideravelmente os resultados obtidos. Uma possível explicação para este fenômeno é que as estimativas do método SHS estejam próximas do estado estacionário da cadeia. Este fenômeno será estudado em trabalhos futuros.

Nas comparações do procedimento BHS contra DP e do SHS contra GAMLSS e MGCV, os resultados obtidos por métodos baseados em *h-splines* sempre foram equivalentes ou superiores. No contexto bayesiano, o método referente ao pacote **DPpackage** tem como padrão o uso de muitas bases na estimação. Isso pode gerar estimativas de curvas com muita rugosidade. Por outro lado, no contexto frequentista, os procedimentos referentes aos pacotes **gamlss** e **mgcv** utilizam uma quantidade reduzida de funções base como padrão. Desse modo, não conseguem capturar toda a estrutura de algumas funções. Os métodos SHS e BHS são baseados em *h-splines*. Isso significa que a quantidade de funções base a serem utilizadas na estimação não é predeterminada, mas sim estimada juntamente com os demais parâmetros do modelo. Desse modo, evita-se tanto o excesso quanto a escassez de funções base e, conseqüentemente, o risco de haver sobreajustamento ou subajustamento dos dados torna-se reduzido.

Os testes de hipóteses propostos também obtiveram bons desempenhos. No estudo de simulação do Capítulo 3, a medida de evidência proposta crescia (decretava) conforme a função testada se aproximava (afastava) da verdadeira. Um comportamento esperado para uma medida de evidência. Além disso, o poder do teste *DQI* proposto também se mostrou adequado no estudo de simulação realizado no Capítulo 5. Na aplicação dos testes a um conjunto de dados reais, o objetivo era escolher um entre oito modelos candidatos. Os dois testes apontaram um mesmo par de modelos com os mais adequados, o que mostra uma coerência nos testes.

Por último, lembramos que os procedimentos de estimação e testes propostos aqui são para modelo aditivos. Portanto, em geral, referem-se à aproximações do verdadeiro modelo no caso de estimação, e de indicativos de decisão no caso dos testes.

8.2 Propostas para Trabalhos Futuros

Uma questão não abordada neste trabalho é a criação de um ponto de corte para regra de decisão no teste de hipóteses bayesiano. Pode-se indicar o uso de novas funções perda ou mesmo indicar a construção de uma regra de decisão mais robusta, que não dependa da opinião do pesquisador sobre o erro mais (ou menos) danoso na sua decisão.

Uma extensão deste trabalho, que parece ser imediata, é a criação de modelos aditivos generalizados *h-splines* utilizando a abordagem bayesiana. O método SHS para o modelo DFGAM já faz essa generalização de modo sequencial.

Outra extensão seria ampliar o procedimento BHS para uma abordagem de seleção de variáveis. A idéia seria adicionar novos passos de nascimento e morte, porém que não tratassem de adição ou remoção de bases, mas sim de covariáveis.

Na abordagem sequencial, a extensão seria a criação de um teste de hipóteses DFGAM, que auxiliaria na escolha tanto da forma funcional f quanto da distribuição. Um exemplo seria um pesquisador em dúvida se usa a função g ou h , e se usa a distribuição gama, log-normal ou Weibull. O teste DFGAM auxiliaria na escolha de uma das seis possíveis combinações.

Referências

- Craven, Peter e Grace Wahba (1978). “Smoothing noisy data with spline functions”. Em: *Numerische Mathematik* 31.4, pp. 377–403.
- Denison, DGT, BK Mallick e AFM Smith (1998). “Automatic Bayesian curve fitting”. Em: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2, pp. 333–350.
- Dias, Ronaldo (1999). “Sequential adaptive nonparametric regression via H-splines”. Em: *Communications in Statistics-Simulation and Computation* 28.2, pp. 501–515.
- Dias, Ronaldo e Dani Gamerman (2002). “A Bayesian approach to hybrid splines non-parametric regression”. Em: *Journal of Statistical Computation and Simulation* 72.4, pp. 285–297.
- DiMatteo, Ilaria, Christopher R Genovese e Robert E Kass (2001). “Bayesian curve-fitting with free-knot splines”. Em: *Biometrika* 88.4, pp. 1055–1071.
- Duchesne, P. e P. Lafaye de Micheaux (2010). “Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods”. Em: *Computational Statistics and Data Analysis* 54, pp. 858–862.
- Good, IJ (1982). “C140. Standardized tail-area probabilities”. Em: *Journal of Statistical Computation and Simulation* 16.1, pp. 65–66.
- Green, Peter J (1995). “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”. Em: *Biometrika* 82.4, pp. 711–732.
- Green, Peter J e Bernard W Silverman (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall, London.

- Gu, Chong e Grace Wahba (1991). “Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method”. Em: *SIAM Journal on Scientific and Statistical Computing* 12.2, pp. 383–398.
- Hastie, Trevor e Robert Tibshirani (1990). *Generalized additive models*. Chapman & Hall, London.
- Jara, Alejandro et al. (2011). “DPpackage: Bayesian non-and semi-parametric modelling in R”. Em: *Journal of Statistical Software* 40.5, p. 1.
- Jeffreys, H (1939). *61, Theory of Probability*. Oxford University Press, Oxford.
- Kimeldorf, George S e Grace Wahba (1970). “A correspondence between Bayesian estimation on stochastic processes and smoothing by splines”. Em: *The Annals of Mathematical Statistics* 41.2, pp. 495–502.
- Madruga, M Regina, Luis G Esteves e Sergio Wechsler (2001). “On the bayesianity of Pereira-Stern tests”. Em: *Test* 10.2, pp. 291–299.
- Pereira, Carlos Alberto de Bragança e Julio Michael Stern (1999). “Evidence and credibility: full Bayesian significance test for precise hypotheses”. Em: *Entropy* 1.4, pp. 99–110.
- Provost, Serge B e Edmund M Rudiuk (1996). “The exact distribution of indefinite quadratic forms in noncentral normal vectors”. Em: *Annals of the Institute of Statistical Mathematics* 48.2, pp. 381–394.
- Souza, Camila P E e Ronaldo Dias (2008). “Testes de hipóteses para dados funcionais baseados em distâncias: um estudo usando splines”. Em:
- Stasinopoulos, D Mikis e Robert A Rigby (2007). “Generalized additive models for location scale and shape (GAMLSS) in R”. Em: *Journal of Statistical Software* 23.7, pp. 1–46.
- Tanaka, Shojiro e Ryuei Nishii (2009). “Nonlinear Regression Models to Identify Functional Forms of Deforestation in East Asia”. Em: *IEEE Transactions on Geoscience and Remote Sensing* 47.8, pp. 2617–2626.
- Wahba, Grace (1981). “Data-based optimal smoothing of orthogonal series density estimates”. Em: *The Annals of Statistics*, pp. 146–156.
- (1983). “Bayesian “confidence interval” for the cross-validated smoothing spline”. Em: *Journal of the Royal Statistical Society. Series B (Methodological)* 1, pp. 133–150.

- Wood, Simon N (2001). “mgcv: GAMs and generalized ridge regression for R”. Em: *R news* 1.2, pp. 20–25.
- (2004). “Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models”. Em: *Journal of the American Statistical Association* 99, pp. 673–686.

Apêndice A

Forma Matricial para $\int [f''(t)]^2 dt$

Seja $\mathbf{b} = (b_1, \dots, b_K)^t$ um vetor com K funções base conhecidas. Considere uma função f que pode ser escrita (ou aproximada/estimada) como uma combinação linear desta bases, ou seja,

$$f(t) = \boldsymbol{\theta}^t \mathbf{b}(t),$$

em que $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^t$ são os coeficientes da combinação linear. Desde que $f''(t)$ seja um escalar e a transposta de um escalar seja igual ao próprio, a obtenção da forma matricial do termo de penalização à não suavidade segue abaixo

$$\begin{aligned} \int [f''(u)]^2 du &= \int [\boldsymbol{\theta}^t \mathbf{b}''(u)]^2 du \\ &= \int \boldsymbol{\theta}^t \mathbf{b}''(u) \mathbf{b}''(u)^t \boldsymbol{\theta} du \\ &= \boldsymbol{\theta}^t \left[\int \mathbf{b}''(u) \mathbf{b}''(u)^t du \right] \boldsymbol{\theta} \\ &= \boldsymbol{\theta}^t \boldsymbol{\Omega} \boldsymbol{\theta} \end{aligned}$$

em que $\boldsymbol{\Omega}$ é uma matriz quadrada de dimensão K com entradas dadas por

$$\Omega_{ij} = \int b_i''(u) b_j''(u) du.$$

Para caso em que as funções base são calculadas em valores observados x_1, \dots, x_n , temos

$$\Omega_{ij} = \sum_{l=1}^n b_i''(x_l) b_j''(x_l).$$

Apêndice B

Identificabilidade do Modelo Aditivo

Seja o modelo aditivo dado por

$$\mathbf{y} = \boldsymbol{\alpha} + \sum_{j=1}^p \mathbf{f}_j + \boldsymbol{\epsilon} = \boldsymbol{\alpha} + \sum_{j=1}^p \mathbf{X}_j \boldsymbol{\theta}_j + \boldsymbol{\epsilon} = \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\epsilon}.$$

Este modelo, em geral, é não identificável. Uma alternativa para torna-lo identificável é impor restrições do tipo

$$\mathbf{C} \boldsymbol{\theta} = \mathbf{0}.$$

Uma restrição adequada para este fim é que a soma (ou média) dos elementos de \mathbf{f}_j deva ser zero, para $j = 1, \dots, p$. Essa restrição pode ser escrita como

$$\mathbf{1}_n^t \mathbf{X}_j \boldsymbol{\theta}_j = 0.$$

Como são p covariáveis, esta é a quantidade de restrições deste tipo a se fazer. Deste modo temos

$$\mathbf{C} = \begin{bmatrix} 0 & \mathbf{1}_n^t \mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{1}_n^t \mathbf{X}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_n^t \mathbf{X}_p \end{bmatrix}$$

Uma abordagem para impor estas restrições é através de uma reparametrização que utiliza a decomposição QR. Seja

$$\mathbf{C}^t = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

em que \mathbf{Q} é uma matriz ortogonal $K \times K$ e \mathbf{R} é uma matriz triangular superior $p \times p$, com $K = 1 + K_1 + \dots + K_p$. A matriz \mathbf{Q} pode ser particionada como $\mathbf{Q} \equiv [\mathbf{D} : \mathbf{Z}]$, em que \mathbf{Z} é uma matriz $K \times (K - p)$.

Deste modo, $\boldsymbol{\theta} = \mathbf{Z}\boldsymbol{\theta}_z$ vai atender as restrições para qualquer vetor $\boldsymbol{\theta}_z$ de dimensão $K - p$. E isso é simples de verificar:

$$\mathbf{C}\boldsymbol{\theta} = \begin{bmatrix} \mathbf{R}^t & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{D}^t \\ \mathbf{Z}^t \end{bmatrix} \mathbf{Z}\boldsymbol{\theta}_z = \begin{bmatrix} \mathbf{R}^t & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{K-p} \end{bmatrix} \boldsymbol{\theta}_z = \mathbf{0}.$$

Daí, para minimizar $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \boldsymbol{\theta}^t \boldsymbol{\Omega} \boldsymbol{\theta}$ com relação a $\boldsymbol{\theta}$ sujeito a $\mathbf{C}\boldsymbol{\theta} = \mathbf{0}$, o seguinte algoritmo pode ser usado.

1. Obtenha a decomposição QR para \mathbf{C}^t e defina \mathbf{Z} como sendo as $K - p$ últimas colunas da matriz ortogonal \mathbf{Q} .
2. Seja $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{Z}$ e $\widetilde{\boldsymbol{\Omega}} = \mathbf{Z}^t \boldsymbol{\Omega} \mathbf{Z}$. Minimize $\|\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}_z\|^2 + \boldsymbol{\theta}_z^t \widetilde{\boldsymbol{\Omega}} \boldsymbol{\theta}_z$ com relação a $\boldsymbol{\theta}_z$, obtendo $\widehat{\boldsymbol{\theta}}_z$.
3. $\widehat{\boldsymbol{\theta}} = \mathbf{Z}\widehat{\boldsymbol{\theta}}_z$.

Anexo I

Licença

Copyright (c) 2014 de Saulo Almeida Morellato.

Exceto quando indicado o contrário, esta obra está licenciada sob a licença Creative Commons Atribuição-CompartilhaIgual 3.0 Não Adaptada. Para ver uma cópia desta licença, visite <http://creativecommons.org/licenses/by-sa/3.0/>.



A marca e o logotipo da UNICAMP são propriedade da Universidade Estadual de Campinas. Maiores informações sobre encontram-se disponíveis em <http://www.unicamp.br/unicamp/a-unicamp/logotipo/normas%20oficiais-para-uso-do-logotipo>.

I.1 Sobre a licença dessa obra

A licença Creative Commons Atribuição-CompartilhaIgual 3.0 Não Adaptada utilizada nessa obra diz que:

1. Você tem a liberdade de:

- Compartilhar – copiar, distribuir e transmitir a obra;
- Remixar – criar obras derivadas;

- fazer uso comercial da obra.

2. Sob as seguintes condições:

- Atribuição – Você deve creditar a obra da forma especificada pelo autor ou licenciante (mas não de maneira que sugira que estes concedem qualquer aval a você ou ao seu uso da obra).
- Compartilhamento pela mesma licença – Se você alterar, transformar ou criar em cima desta obra, você poderá distribuir a obra resultante apenas sob a mesma licença, ou sob uma licença similar à presente.