**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Estruturas unidimensionais e bidimensionais utilizando P-*splines* nos
modelos mistos aditivos generalizados com aplicação na produção de
cana-de-açúcar**

Natalie Verónika Rondinel Mendoza

Tese apresentada para obtenção do título de Doutora
em Ciências. Área de concentração: Estatística e Ex-
perimentação Agronômica

**Piracicaba
2017**

Natalie Verónika Rondinel Mendoza
Licenciada en Matemática - Mención Estadística

**Estruturas unidimensionais e bidimensionais utilizando *P-splines* nos
modelos mistos aditivos generalizados com aplicação na produção de
cana-de-açúcar**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Tese apresentada para obtenção do título de Doutora
em Ciências. Área de concentração: Estatística e Ex-
perimentação Agronômica

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Rondinel Mendoza, Natalie Verónica

Estruturas unidimensionais e bidimensionais utilizando *P-splines* nos modelos mistos aditivos generalizados com aplicação na produção de cana-de-açúcar / Natalie Verónica Rondinel Mendoza. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2017 .

84 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. *P-splines* 2. *B-splines* 3. Produto tensor 4. Modelo misto aditivo generalizado I. Título

DEDICATÓRIA

Com amor e carinho agradeço a **Deus**, meu Pai do céu
por ter colocado muitas provas neste caminho da vida
(e que ainda irá colocar), porque eu sei que com
dedicação, fé e paciência fará que eu nunca perca as forças .

Com todo meu amor incondicional a meus pais,
Sócrates Rondinel e **Fortunata Mendoza**,
a minha irmãzinha, **Katy Rondinel Mendoza**,
que sempre me apoiam com esse amor incondicional e
alegram minha vida, mesmo estando a quilômetros de
distância.

A meu irmão **Juan Carlos Rondinel Mendoza**
(*in memoriam*), quem foi meu anjo da guarda
durante toda esta jornada, dedico este trabalho,
saudades eternas.....

AGRADECIMENTOS

Aos meus pais: Sócrates Rondinel e Fortunata Mendoza, os quais eu amo profundamente, por estarem ao meu lado, mesmo a quilômetros de distância durante todo este tempo de estudos e permanência aqui no Brasil, meu segundo lar, agradeço a Deus por ter me dado pais tão amorosos como vocês.

A minha irmã Katy Rondinel, por todo apoio, ajuda e pensamento positivo. Sempre me fazendo olhar para frente, nunca me deixando abalar com adversidades e NUNCA desistir.

A minha orientadora Dra. Sônia Maria De Stefano Piedade, por ter sido para mim como uma mãe e amiga aqui no meu segundo lar o Brasil, fico muito agradecida pelas conversas, sugestões tanto para meus estudos como para minha vida pessoal e por nunca ter desistido de mim mesmo, por sempre estar presente para me auxiliar e direcionar no meu caminho.

Aos meus amigos: Cesia Flores, Marcus Gurgel, Andrés Trujillo e Thiago Oliveira, pelos momentos de convivência, de alegria e discussões sobre valores que sempre fizeram abrir meus olhos para outros pontos de vista, e como ser melhor pessoa e principalmente por estarem ao meu lado nos momentos de aperto sempre oferecendo ajuda.

Aos meus amigos de pós-graduação: Ana Julia Riguetto, Luiz Ricardo Nakamura, Rafael Moral, Ricardo Klein, Maria Cristina Martins e Cristiane Rodriguez, pelo companheirismo.

Às secretárias do programa de pós-graduação: Luciane Brajão e Solange Paes Sabadin por toda a ajuda nos momentos cruciais e na correria da reta final dos meus estudos.

Aos funcionários do programa de pós-graduação: Jorge Wiendl e Eduardo Bonilha por me ajudar quando eu precisava de algum apoio técnico e informático e pelas longas conversas.

Agradecimento muito especial à Dra. Clarice Garcia Borges Demétrio, por ter-me dado a oportunidade de estudar no departamento desde a época do Mestrado, por sempre estar preocupada comigo, pelas conversas e pelos conselhos.

Aos professores do programa de pós-graduação em Estatística e Experimentação Agronômica: Decio Barbin, Carlos Tadeu dos Santos Dias, Silvio Sandoval Zocchi, Idemauro Antonio R. de Lara, Edwin Moisés Marcos Ortega, Roseli Aparecida, César Gonçalves De Lima, Renata Alcarde, Cristian Villegas e Taciana Villela Savian pelos conselhos e ajuda nos momentos de dúvidas.

Agradecimento especial ao Henrique Boriolo Dias, que me ajudou com algumas informações de dados meteorológicos para o desenvolvimento de meu trabalho.

Agradecimento especial à Professora Dra. Maria Durbán, que foi a pessoa que me inspirou no desenvolvimento deste trabalho e graças a um curso outorgado por ela em nosso departamento.

Ao Programa Capes pelo apoio financeiro com uma bolsa de estudos para toda etapa de trabalho e estadia no Brasil.

Por fim a todos que sempre acreditaram em mim.

OBRIGADA!!!!!!

EPÍGRAFE

O trabalho duro
e a dedicação,
não são garantia
de sucesso, mas sem isso,
não temos a menor chance.

Pat Riley

Quando eu era menina, gostava de pedalar a bicicleta com bastante força para lá na frente os pés do pedal e deixar a bicicleta correr sem esforço algum, apenas sentindo o vento no rosto. O que eu não sabia é que a vida é dada essas coisas também. Há o tempo de pedalar e o tempo de relaxar. Talvez você ainda não sinta o vento no rosto, e a fase que está vivendo dói tanto quanto cortar o dedo numa folha de papel afiada. Mas experimente confiar ... Confiar em Deus, confiar no momento que você vive confiar nos rumos que sua vida tomou. Aquiete seu coração e confie no tempo de esperas, acalme seu espírito e confie no novo dia que está nascendo, nas novas mudas que estão florindo, no cabelo que está crescendo. Confie na renovação que ocorre todo dia sem que a gente se dê conta, nas possibilidades que a vida nos dá diariamente, e nos finais que precisamos digerir e aceitar...

Fabíola Simões

SUMÁRIO

Resumo	8
Abstract	9
Lista de Figuras	10
Lista de Tabelas	12
Lista de Abreviaturas e Siglas	13
1 Introdução	15
2 Desenvolvimento	17
2.1 Regressão paramétrica versus regressão não-paramétrica	17
2.2 Função de regressão	18
2.3 Métodos de suavização não-paramétrica	18
2.4 Espaço de funções	19
2.4.1 <i>Scatterplot smoothing</i>	20
2.4.2 <i>Penalidade à falta de suavidade das curvas</i>	20
2.4.2.1 O princípio do ajuste da curva	20
2.4.2.2 Estimação da função de regressão	21
2.5 <i>Splines</i>	22
2.6 Métodos de suavização utilizando <i>splines</i>	23
2.6.1 Suavização por <i>splines</i>	23
2.6.1.1 <i>Spline</i> cúbico	24
2.6.2 <i>Splines</i> de Regressão	24
2.6.3 Suavização por <i>splines</i> penalizados	26
2.6.3.1 <i>Splines</i> penalizados de O'Sullivan	26
2.6.4 Suavização por P- <i>splines</i>	27
2.6.5 Bases	29
2.6.5.1 Bases polinomiais truncadas	31
2.6.5.2 Bases B- <i>splines</i>	33
2.6.6 Propriedades dos P- <i>splines</i>	34
2.6.6.1 Matriz de suavização	34
2.6.6.2 Graus de liberdade efetivos	35
2.6.6.3 Graus de liberdade dos resíduos	35
2.6.6.4 Seleção do parâmetro de suavização	37
2.7 P- <i>Splines</i> como modelos mistos	38
2.7.1 Revisão geral de modelos mistos	39
2.7.2 Estimação de efeitos fixos	40
2.7.3 Predição de efeitos aleatórios	40
2.7.4 Estimação de componentes de variância	41
2.7.5 Representação dos P- <i>splines</i> como modelos mistos	42
2.7.6 Intervalos de confiança para a função suave	46
2.8 Suavizado bidimensional	48
2.9 P- <i>splines</i> bidimensional	48
2.9.1 Produto tensor de bases	49
2.9.2 Produto tensor das penalizações	49
2.10 P- <i>splines</i> bidimensionais como modelos mistos	50
2.11 Modelo Misto Aditivo Generalizado	53
2.11.1 Modelo linear generalizado	53

2.11.2	Modelos mistos lineares generalizados	54
2.11.3	Modelo misto aditivo generalizado	55
2.11.4	Suavizado dos componentes com representação <i>P-spline</i>	56
3	Material e Métodos	59
3.1	Materiais	59
3.2	Métodos	59
3.2.1	Análise exploratória dos dados	60
3.2.2	Abordagem 1: Modelos mistos	62
3.2.3	Abordagem 2: Representação <i>P-splines</i> como modelos mistos	63
3.2.4	Abordagem 3: Modelo misto aditivo generalizado com <i>P-splines</i>	64
3.2.4.1	Modelo aditivo e bidimensional suave para temperatura máxima e precipitação	65
4	Resultados	67
4.1	Abordagem 1: Modelos mistos	67
4.2	Abordagem 2: Representação <i>P-splines</i> como modelos mistos	67
4.2.1	Abordagem 3: Modelo misto aditivo generalizado com <i>P-splines</i>	68
4.2.1.1	Suavizado unidimensional	68
4.2.1.2	Suavizado do modelo aditivo e bidimensional para temperatura máxima e precipitação	71
5	Discussão	77
6	Conclusão	79
	Referências	81

RESUMO

Estruturas unidimensionais e bidimensionais utilizando *P-splines* nos modelos mistos aditivos generalizados com aplicação na produção de cana-de-açúcar

Os *P-splines* de EILERS e MARX (1996) são métodos de suavização que é uma combinação de bases *B-splines* e uma penalização discreta sobre os coeficientes das bases utilizados para suavizar dados normais e não normais em uma ou mais dimensões, no caso de várias dimensões utiliza-se como suavização o produto tensor dos *P-splines*. Também os *P-splines* são utilizados como representação de modelos mistos CURRIE *et al.* (2006) pela presença de características tais como: efeitos fixos, efeitos aleatórios, correlação espacial ou temporal e utilizados em modelos mais generalizados tais como os modelos mistos lineares generalizados e modelos mistos aditivos generalizados. Neste trabalho apresentou-se toda a abordagem, metodologia e descrição dos *P-splines* como modelos mistos e como componentes das estruturas suavizadoras de variáveis unidimensionais e bidimensionais dos modelos mistos aditivos generalizados, mostrando essa abordagem e propondo seu uso em uma aplicação no comportamento dos níveis médios da produção de cana-de-açúcar sob a influência das alterações das variáveis climáticas como temperatura e precipitação, que foram medidos ao longo de 10 anos em cada mesoregião do Estado de São Paulo. O motivo de usar essa abordagem como método de suavização é que muitas vezes não é conhecido a tendência dessas covariáveis climáticas mas sabe-se que elas influenciam diretamente sobre a variável resposta. Além de permitir essa abordagem inclusão de efeitos fixos e aleatórios nos modelos a serem propostos, permitirá a inclusão do processo autoregressivo AR(1) como estrutura de correlação nos resíduos.

Palavras-chave: *P-splines*, *B-splines*, Produto tensor, Modelo misto aditivo generalizado

ABSTRACT

Unidimensional and bidimensional structures using *P-splines* in generalized additive mixed models with application in the production of sugarcane

P-splines of EILERS e MARX (1996) are methods of smoothing that is a combination of *B-splines* bases and penalty the coefficients of the bases used to smooth normal and non-normal data in one or more dimensions; in the case of several dimensions it is used as smoothing the tensor product of the *P-splines*. Also the *P-splines* are used as representation of mixed models CURRIE *et al.* (2006) by the presence of characteristics such as: fixed effects, random effects, spatial or temporal correlation and used in more generalized models such as generalized linear mixed models and generalized additive mixed models. In this work the whole approach, methodology and description of the *P-splines* as mixed models and as components of the smoothing structures of one-dimensional and two-dimensional variables of generalized additive mixed models were presented, showing this approach and proposing its application in the behavior of the average levels of sugarcane production, which is influenced by changes in climatic variables such as temperature and precipitation, which were measured over 10 years in each mesoregion of the state of São Paulo. The reason for using this approach as a smoothing method is that the tendency of these climate covariables is not known for the most part, but is known that they influence directly the response variable, besides allowing this approach to include fixed and random effects in the models to be proposed, will allow the inclusion of the autoregressive process AR(1) as a correlation structure in the residuals.

Keywords: *P-splines*, *B-splines*, Tensor product, Generalized additive mixed model

LISTA DE FIGURAS

2.1	Gráfico de dispersão de dados simulados	21
2.2	Curvas ajustadas utilizando suavização por <i>splines</i> , com 5, 15, 25 e 35 nós, para diferentes valores dos parâmetros de suavização para a função $f(x) = \cos(5\pi x)$, para um conjunto de 200 dados simulados.	24
2.3	Curvas ajustadas utilizando <i>Splines</i> de regressão com 5, 15, 25 e 35 nós, para a função $f(x) = \cos(5\pi x)$, para um conjunto de 200 dados simulados.	26
2.4	Curvas ajustadas utilizando P- <i>splines</i> , com 5, 15, 25 e 35 nós, com diferentes valores dos parâmetros de suavização para a função $f(x) = \cos(5\pi x)$, com 200 dados simulados.	29
2.5	Bases de polinômios truncados com diferentes graus: (a) Grau 0, (b) Grau 1, (c) Grau 2 e (d) Grau 3.	32
2.6	Bases de B- <i>splines</i> : (a) B- <i>splines</i> de grau 1, (b) sequência de B- <i>splines</i> de grau 1, (c) B- <i>splines</i> de grau 3 e (d) sequência de B- <i>splines</i> de grau 3.	34
2.7	Curva ajustada sem penalizar os coeficientes (a) utilizando os <i>splines</i> de regressão, e penalizando os coeficientes (b) utilizando os P- <i>splines</i> . Para a penalização, considerou-se 30 nós com um conjunto 200 dados simulados para a função $f(x) = \cos(5\pi x)$	35
2.8	Curva estimada usando P- <i>splines</i> com diferentes valores de λ , com 200 dados simulados para a função $f(x) = \cos(5\pi x)$	37
2.9	Curvas estimadas usando diferentes métodos de suavização: <i>Splines</i> de regressão, suavização por <i>splines</i> e P- <i>splines</i> , utilizando neste último 35 nós para as bases de B- <i>splines</i> com $\lambda = 4$, para a função $f(x) = \cos(5\pi x)$ para um conjunto de 200 dados simulados.	39
2.10	Curvas estimadas usando diferentes valores de nós para as bases de B- <i>splines</i> , usando representação P- <i>splines</i> como modelos mistos para a função $f(x) = \cos(5\pi x)$ para um conjunto de 200 dados simulados.	46
2.11	Intervalo de confiança para as bases de B- <i>splines</i> , utilizando P- <i>splines</i> como modelos mistos para a função $f(x) = \cos(5\pi x)$ com 200 dados simulados.	47
3.1	Mapa edafoclimático da produção de cana-de-açúcar do Estado de São Paulo.	60
3.2	Matriz de dispersão das variáveis produção, precipitação, temperatura máxima, temperatura mínima e ano.	61
3.3	Gráfico semi-normal de probabilidades com envelope simulado para a distribuição dos resíduos do modelo.	62
3.4	Gráfico do comportamento da produção de cana-de-açúcar num período de 10 anos.	63
4.1	Intervalo de confiança para a média da produção de cana-de-açúcar ao longo do tempo.	69
4.2	Formas das funções suaves estimadas para a covariável Ano na produção de cana-de-açúcar utilizando modelo generalizado aditivo misto em que (a) Curva estimada sem considerar estrutura de correlação nos resíduos e (b) Função de autocorrelação sem considerar o efeito dos resíduos. (c) Curva estimada considerando estrutura de correlação nos resíduos e (d) Função de autocorrelação considerando estrutura autoregressiva AR(1).	70
4.3	Forma dos suavizadores para temperatura mínima, máxima e precipitação utilizando modelo generalizado aditivo misto.	71
4.4	Curvas ajustadas com um modelo aditivo para a) temperatura máxima, b) precipitação e c) superfície aditiva considerando as duas variáveis climáticas, temperatura máxima e precipitação.	72

4.5	Curvas ajustadas em um modelo bidimensional para temperatura máxima e precipitação: a) curvas de nível de superfície e b) superfície bidimensional.	73
4.6	Valores preditos para produção de cana-de-açúcar ao longo dos anos.	75

LISTA DE TABELAS

3.1	Estimativas e erros padrão dos parâmetros da regressão.	62
3.2	Modelos mistos (MM) considerando efeito aleatório no intercepto e no tempo (Ano).	63
3.3	Modelos com função suave na variável Ano, efeito aleatório no intercepto, utilizando P- <i>splines</i> com representação como modelos mistos - MPS	64
4.1	Critérios de informação AIC e BIC para os modelos mistos propostos.	67
4.2	Estimativas dos parâmetros dos efeitos fixos e componentes de variância para o modelo MM1, erro padrão, p-valor.	67
4.3	Critérios de informação AIC e BIC para os modelos P- <i>splines</i> propostos.	68
4.4	Estimativas dos parâmetros dos efeitos fixos e componentes de variância para o modelo MPS1, erros padrão, p-valor.	68
4.5	Comparação de modelos considerando estrutura de correlação autoregressiva nos resíduos para a covariável Ano.	69
4.6	Aproximação das funções suaves estimadas individuais das covariáveis no GAMM.	70
4.7	Funções suaves estimadas do modelo aditivo para as covariáveis temperatura máxima e precipitação no GAMM.	72
4.8	Função suave estimada do modelo bidimensional das covariáveis temperatura máxima e precipitação no GAMM.	73
4.9	AIC e BIC dos modelos aditivo e bidimensional de temperatura máxima e precipitação. . .	73
4.10	Funções suaves estimadas das covariáveis por mesorregião.	74

LISTA DE ABREVIATURAS E SIGLAS

MM	Mixed Models
GLM	Generalized Linear Models
GLMM	Generalized Linear Mixed Models
GAMM	Generalized Additive Mixed Models
AIC	Critério de informação de Akaike
BIC	Critério de informação Bayesiano
CV	Validação Cruzada
GCV	Validação Cruzada Generalizada
BP	Melhor preditor
BLP	Melhor preditor linear
BLUP	Melhor preditor linear não viesado
DVS	Decomposição em valores singulares
PAM	Produção Agrícola Municipal
IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto Nacional de Meteorologia
CIIAGRO	Centro Integrado de Informações Agrometeorológicas
UNICA	União da Indústria de Cana-de-Açúcar.
ESALQ	Escola Superior de Agricultura “Luiz de Queiroz”

1 INTRODUÇÃO

Uma característica comum de dados reais é que frequentemente possuem complexidade na relação linear ou não linear entre as variáveis respostas e suas covariáveis, já que na maioria dos estudos, o comportamento dos preditores ou covariáveis não segue um determinado padrão probabilístico; portanto, ao longo dos anos vários autores propuseram diferentes alternativas para modelar dados complexos, assim: 1) Os modelos de regressão paramétrica, supõem uma forma definida e paramétrica para a função da regressão os quais garantem as propriedades das estimações das curvas. No entanto, esses modelos caracterizam-se por ser métodos pouco flexíveis e de difícil adaptação às diferentes situações reais e 2) os modelos de regressão não paramétrica, os quais permitem uma maior flexibilidade nas curvas e não precisam de hipóteses sobre a estrutura e comportamento dos dados. Utilizar modelos não-paramétricos implica considerar algum método de suavização, entre os quais podem ser mencionados, *Splines* penalizados (O’SULLIVAN (1986), RUPPERT *et al.* (2003)) e *P-splines* (EILERS e MARX (1996), DURBAN (2002), CURRIE *et al.* (2006) e EILERS *et al.* (2015)). Esse tipo de abordagem desenvolve procedimentos estatísticos nos quais não precisa de hipóteses sobre o comportamento, distribuição ou tendência que relaciona as covariáveis com as variáveis respostas.

Portanto, o objetivo do presente trabalho, é no uso de métodos de suavização e em particular, no uso de *P-splines*. Estes métodos são bem estabelecidos para suavizar dados normais e não normais em uma ou mais dimensões (CURRIE *et al.*, 2006) e eles têm sido utilizados em diferentes contextos. Também os *P-splines* foi utilizado como representação de modelos mistos (CURRIE e DURBAN (2002); WAND (2003); CURRIE *et al.* (2006); LEE e DURBAN (2011) e (WOOD *et al.*, 2015)) pela presença de características tais como: efeitos fixos, efeitos aleatórios, correlação espacial ou temporal, assim como o uso em modelos mais generalizados tais como os modelos mistos lineares generalizados e modelos mistos aditivos generalizados (DURBÁN, 2014).

No presente trabalho, propõe-se o uso, descrição, aplicação e avaliação dos *P-splines* nas estruturas de componentes aditivos do modelo misto aditivo generalizado para estudar o comportamento dos níveis médios da produção de cana-de-açúcar de cada mesorregião do estado de São Paulo, sob a influência das alterações nos níveis de temperatura e precipitação em um determinado período de tempo. Existem vários motivos de usar essa abordagem, 1) muitas vezes o comportamento ou distribuição dessas variáveis climáticas que influenciam sobre a produção de cana-de-açúcar ou qualquer outro cultivar são desconhecidos, 2) a relação existente entre a produção e as variáveis climáticas nem sempre é linear e 3) permite a inclusão de efeitos fixos e aleatórios nos modelos que serão propostos bem como a possível inclusão de correlação que existe entre as medições tomadas nas mesmas unidades observacionais ao longo do tempo sendo analisados unidimensional e bidimensionalmente utilizando os *P-splines*.

Com a motivação apresentada, este trabalho visa inserir a abordagem dos métodos de suavização utilizando *P-splines*, já que este tipo de dados possui estrutura complexa.

Para o desenvolvimento do presente trabalho, será estruturado nos seguintes temas de estudo: uma breve introdução e descrição dos métodos de regressão paramétrica e não paramétrica, uma revisão geral dos *splines* e os métodos de suavização com *splines*, revisão de *P-splines* como modelos mistos no caso unidimensional e bidimensional, uso dos *P-splines* como componentes suavizadores nos modelos mistos aditivos generalizados e finalmente, aplicação de toda a metodologia dos *P-splines* no caso de estudo de produção de cana-de-açúcar.

2 DESENVOLVIMENTO

2.1 Regressão paramétrica versus regressão não-paramétrica

Seguindo a abordagem de DELICADO (2008), inicia-se esse capítulo com uma breve comparação entre a regressão paramétrica e não-paramétrica, para um melhor entendimento e descrição do tema *Splines*.

Definição 2.1.1 A função de distribuição da variável aleatória X , representada por $F(x)$ ou simplesmente por F , é definida por

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

A variável aleatória X segue um “Modelo paramétrico”, denotado por F_θ , se a função de distribuição F , pertence a uma família de distribuições indexado por um parâmetro θ de dimensão finita definida por,

$$X \sim F \in \mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\} \quad (2.1)$$

em que Θ é o espaço paramétrico contido nos reais de dimensão k .

Por outro lado, a variável aleatória X segue um “Modelo não-paramétrico”, se a função de distribuição F cumpre com determinadas pressuposições de regularidade:

- F é uma função de distribuição absolutamente contínua,
- F apresenta simetria em torno a sua mediana,
- F tem função de densidade $f(x)$ com duas derivadas contínuas.

Com essas restrições de regularidade de F indicam de que essa função pertence a um subconjunto da família de distribuições com dimensão infinita; isto significa, não é possível indexar um parâmetro de dimensão finita.

Definição 2.1.2 A variável aleatória X é (absolutamente) contínua se existe uma função $f(x) \geq 0$ tal que

$$F(x) = \int_{-\infty}^x f(u)du, \quad \forall x \in \mathbb{R}.$$

Diz-se que $f(x)$ é uma função de densidade de X , ou simplesmente densidade de X .

Seja o conjunto de n observações $\{(x_i, y_i), i = 1, 2, \dots, n\}$ de uma variável bidimensional (X, Y) satisfazendo o seguinte modelo:

$$y_i = f(x_i) + \epsilon_i, \quad (2.2)$$

em que as variáveis aleatórias ϵ_i , são os erros do modelo. Supondo que existe uma relação funcional entre as variáveis explicativa X e a resposta Y , então a função $f(x)$ têm por objetivos:

- Analisar a relação existente entre as variáveis X e Y ;
- Prever o comportamento de Y para valores desconhecidos de X ;
- Estimar as derivadas de $f(x)$ e propriedades relacionadas.

Para conseguir esses objetivos podemos realizar por duas aproximações:

1. Supondo um modelo paramétrico para a função $f(x)$ ou;
2. Não impor nenhuma restrição, e supor que a função $f(x)$, apresente algumas propriedades geralmente relacionadas com a derivabilidade.

Segundo ABDELKADER (2009), o primeiro caso corresponde a uma Regressão Paramétrica, e supõe que a função $f(\cdot)$ pertence a alguma classe paramétrica de funções, $f \in \{f_\theta | \theta \in \Theta\}$, onde o problema poderia ser resolvido encontrando essa função sob alguns critérios de otimização (normalmente o critério de mínimos quadrados). Para o segundo caso, refere-se à Regressão não-Paramétrica, cuja única restrição que é imposta na função, $f(\cdot)$, é que seja “suave”, deve-se entender por suavização em termos de derivabilidade, é assim que recebem o nome de métodos de suavização.

Os métodos de suavização são métodos de inferência estatística não-paramétricos, utilizados quando não se precisa de hipóteses paramétricas sobre a distribuição da variável, sendo necessário o uso de técnicas estatísticas que tenham por objetivo estimar a curva relacionada com essa distribuição de probabilidade da variável observada sem necessidade de especificar essas hipóteses; tais técnicas são chamadas “Métodos de estimação não paramétrica de curvas”. Entre as funções $f(x)$ que estimam-se as curvas são:

- A função de densidade $f(x)$, sua derivada e integral (*função de distribuição*), expressada por:
Se $X \sim f(x)$, então $F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du$, logo $f(x) = F'(x)$.
- A função de regressão $f(x)$, e suas derivadas,
Se $(X, Y) \sim F(x, y)$, então $f(x) = E(Y|X = x)$.
- A função de risco $\lambda(x)$, suas derivadas e a integral (*função de risco acumulada*, $\Lambda(x)$):
Se $X \sim f(x)$, então $\lambda(x) = \frac{f(x)}{1-F(x)}$, com $\Lambda(x) = \int_{-\infty}^x \lambda(u)du$, logo $\lambda(x) = \Lambda'(x)$.

2.2 Função de regressão

Considerando o modelo de regressão da Equação 2.2, em que as variáveis aleatórias ϵ_i são não correlacionadas com $E(\epsilon_i) = 0$ e $Var(\epsilon_i) = \sigma^2$ para todo i , dado que a variância dos erros ϵ_i é constante, então diz-se que o modelo é “Homocedástico”. Supondo também que essa variância é função das variáveis explicativas x_i , $Var(\epsilon_i) = \sigma^2(x_i)$, então diz-se que o modelo é “Heterocedástico”; de acordo com DELICADO (2008), a função $f(\cdot)$, é uma função desconhecida das observações *a priori*, chamada “função da regressão”, portanto, não especifica-se a forma dessa função $f(x)$, mas tem-se a suposição de que ela é regular; isto significa que tem segunda derivada contínua. A função de regressão também é conhecida como “curva da regressão” ou “função esperança”, definida por, $f(x) = E(Y|X = x)$; logo

$$Y = f(x) + \epsilon, \quad \text{com } E(\epsilon|X) = 0, \quad Var(\epsilon|X) = \sigma^2(X), \quad \epsilon = Y - f(x).$$

Portanto, o interesse está em estudar a esperança de Y condicionada a X que toma um valor conhecido x_i . Uma vez estabelecido o modelo, o seguinte passo é ajustar $f(x)$ baseando-se nas n observações disponíveis; isso significa, que deve-se encontrar um estimador $\hat{f}(x)$ da função de regressão e um estimador $\hat{\sigma}^2$ da variância do erro. Tais procedimentos de estimação de $f(x)$ são conhecidos como “Métodos de suavização não-paramétrica”.

2.3 Métodos de suavização não-paramétrica

Os primeiros estimadores de regressão não-paramétrica foram propostos por (WATSON, 1964) denominados “estimadores tipo núcleo de Nadaraya-Watson”, os quais foram melhorados com os métodos de regressão polinomial local, tornando-se uns dos métodos mais utilizados e fundamenta-se na escolha adequada dos parâmetros tanto da largura da banda mais conhecido como “parâmetro de suavização” e os graus de ajuste polinomiais, que são os que definem o bom comportamento dos estimadores resultantes da regressão.

EUBANK (1988), diz: “Supondo que a função de regressão desconhecida, $f(\cdot)$, é suave, então podemos esperar que as observações tomadas em pontos próximos a um dado x , pode informar do vetor da função em x .”

Para estimar com métodos não-paramétricos a função de regressão, existem diferentes abordagens, mas todas elas têm algo em comum, estão baseadas nos dados próprios para especificar a forma que tem o modelo; isto é, a curva em um ponto dado, só depende das observações vizinhas. Entre as principais técnicas de estimação, tem-se

- **Ajuste local de modelos paramétricos:** A estimação de $f(x)$, é realizada utilizando regressão polinomial com pesos (*Lowess*), consiste em fazer vários ajustes paramétricos levando em consideração unicamente os dados mais próximos do ponto onde deve ser feita a estimação da função (DELICADO, 2008).
- **Métodos baseados em séries ortogonais de funções:** Essa estimação é feita utilizando bases ortonormais do espaço vetorial de funções e estimam-se os coeficientes da base da função de regressão. Os mais utilizados neste método são os ajustes por Séries de Fourier e os *wavelets* (WASSERMANN, 2006).
- **Suavizado com splines:** Esse método consiste em buscar um estimador $\hat{f}(x)$ que minimiza a soma de quadrados dos resíduos ($\epsilon_i = y_i - \hat{f}(x_i)$) adicionando um termo que penaliza a falta de suavidade das funções $\hat{f}(x)$ que sejam candidatas; cuja penalidade expressada em termos da segunda derivada do quadrado integrável. A solução desse método de suavização é um *spline* cúbico com nós nos pontos observados x_i (GREEN e SILVERMAN, 1994).
- **Técnicas de aprendizado supervisionado:** Segundo Delicado (2008), entre os mais utilizados estão as redes neurais, os k vizinhos mais próximos e as árvores de regressão que são utilizados para estimar $f(x)$, portanto para serem utilizadas essas técnicas, a variável resposta e o preditor devem ser contínuos para estimar não-parametricamente a função de regressão (HASTIE *et al.*, 2009).

Entre as vantagens e desvantagens a serem consideradas em relação a regressão paramétrica e a não-paramétrica, salienta-se:

1. Os modelos de regressão paramétrica, são modelos mais simples e de processos computacionais mais rápidos. Além disso, supõe-se que eles apresentam uma forma definida e paramétrica para a função da regressão os quais garantem as propriedades das estimações das curvas. No entanto, esses modelos caracterizam-se por ser métodos pouco flexíveis e de difícil adaptação às diferentes situações reais;
2. Os modelos de regressão não-paramétrica, permitem uma maior flexibilidade na modelagem dos dados e também não precisa de realizar muitas hipóteses sobre a estrutura e comportamento dos dados, traduz-se em soluções mais complexas e com um alto custo computacional.

2.4 Espaço de funções

O método de regressão não-paramétrica apresenta uma vantagem, o qual permite uma maior flexibilidade na forma da curva de regressão, $f(x)$. Assume-se que essa curva de regressão pertence a uma coleção infinito dimensional de funções, assim por exemplo, $f(x)$, pode pertencer a uma família de funções que são diferenciáveis e a segunda derivada do quadrado integrável (DIAS, 2001). Logo, para construir um modelo de regressão não-paramétrica precisa escolher o espaço de funções adequado para $f(x)$.

Definição 2.4.1 Seja f uma função numérica, definida em um subconjunto aberto $\Omega \in \mathbb{R}^n$, representado por $L^p(\Omega)$, em que $1 \leq p < \infty$, chamado de “espaço de Banach” das funções numéricas f definidas em Ω , mensuráveis cuja potência p , $|f(x)|^p$, é integrável em Ω , definido no espaço vetorial com norma

$$\|f\|_{L^p(\Omega)} = \left(\int_{\Omega} |f(x)|^p dx \right)^{1/p}.$$

Caso particular, utilizando a Definição 2.4.1, em que $p = 2$ e definido no intervalo $\Omega = [a, b]$, é chamado “espaço de Hilbert”, denotado por $L^2([a, b])$ e definido por

$$\|f\|_{L^2([a, b])} = \left(\int_a^b |f(x)|^2 dx \right)^{1/2},$$

com o produto interno de duas funções $f(x)$ e $g(x)$

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx, \quad \forall f, g \in L^2[a, b].$$

Portanto, deve-se procurar uma função de regressão f em um subconjunto de $L^2[a, b]$, tal que, essas funções desse subconjunto sejam “suaves” (ou seja, tem m derivadas contínuas). A escolha de tal subconjunto é um novo espaço chamado “espaço de Sobolev”, representado por $W_2^m[a, b]$, sendo o espaço vetorial de todas as funções $f^{(m)}$ que pertence a $L^2[a, b]$, em que $f^{(m)}$ é a m -ésima derivada de f , definido por

$$W_2^m[a, b] = \{f : f^{(m)} \text{ é absolutamente contínua, e } f^{(m)} \in L^2[a, b]\}. \quad (2.3)$$

2.4.1 Scatterplot smoothing

Os *scatterplot smoothing*, tornou-se uma ferramenta comum no contexto da análise de dados, quando o interesse é destacar a “tendência subjacente ou fundamental dos dados” num gráfico de dispersão (RUPPERT *et al.*, 2003). Os pontos do gráfico de dispersão são simplesmente tratados como uma coleção de pontos sobre um plano, sem qualquer relação com um modelo probabilístico subjacente. Para um melhor entendimento, considere-se a Equação 2.2 para o caso univariado; logo, deve-se pensar na posição vertical de cada ponto como uma realização de uma variável aleatória Y condicionada a uma variável X com valor correspondente à posição horizontal de cada ponto, dada pela função de regressão, $f(x) = E[Y|X = x]$; assim a função não especificada $f(x)$, precisa ser estimada de (x_i, y_i) com $i = 1, \dots, n$ (RUPPERT *et al.*, 2003).

A Figura 2.1, mostra o gráfico de dispersão para dados simulados de um conjunto de 200 observações, utilizando como função, $f(x) = \cos(5\pi x)$, e cuja variável independente x varia no intervalo $[0, 1]$.

2.4.2 Penalidade à falta de suavidade das curvas

2.4.2.1 O princípio do ajuste da curva

Existem muitos dados em que claramente fazendo um ajuste linear não é apropriado utilizar um modelo de regressão linear simples, portanto um modelo mais apropriado a ser utilizado seria

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim Normal(0, \sigma^2) \quad (2.4)$$

em que $f(x)$ é uma curva. Neste caso, uma abordagem clássica é utilizar uma regressão polinomial de ordem baixa para $f(x)$ e cujos coeficientes são estimados por mínimos quadrados. O uso da regressão polinomial é uma técnica bastante popular, só isso apresenta vários inconvenientes, um deles é dado

Figura 2.1: Gráfico de dispersão de dados simulados

pelas observações individuais que podem exercer uma influência e de forma inesperada sobre as partes da curva. Para resolver esse problema de uma forma bastante simples usa-se a abordagem chamada *Roughness penalty* ou penalidade à falta de suavidade das curvas.

Considerando primeiro a seguinte pergunta: o que aconteceria se o modelo (2.4) fosse estimado por mínimos quadrados sem levar em consideração algumas restrições sobre a curva $f(x)$. Nesse caso, a soma de quadrados residual $\sum_{i=1}^n (y_i - f(x_i))^2$ será reduzida a zero, portanto estaríamos interpolando os dados (falta de suavidade). Para resolver esse problema, deve-se considerar condição de suavidade em $f(x)$, de acordo com GREEN e SILVERMAN (1994), isto significa “considerar variação muito local na curva como “ruído” (*noise*) aleatório a fim de estudar a tendência mais variável dos dados”. Portanto, a ideia básica da abordagem *Roughness penalty*, é quantificar a flutuação rápida da curva, por meio da adição de uma penalidade à falta de suavidade da curva.

2.4.2.2 Estimação da função de regressão

Considerando o modelo definido em (2.4), a estimação de $f(x)$ é feita utilizando a estimação por mínimos quadrados ordinários como solução ao problema de minimização da soma de quadrados de resíduos, dado por

$$\min_{f: \mathbb{R} \rightarrow \mathbb{R}} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 \right\}, \quad (2.5)$$

reformulando a expressão anterior com mínimos quadrados penalizados,

$$\min_{f: \mathcal{M}} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 + \Phi(f) \right\}, \quad (2.6)$$

em que $\mathcal{M}$ é uma classe de funções suaves que pertence ao espaço de Sobolev definido em (2.3), e $\Phi(f)$ é uma função ($\Phi: \mathcal{M} \rightarrow \mathbb{R}$) encarregada de penalizar a falta de suavidade de f . Como as observações x_i

estão no intervalo $[a, b] \subseteq \mathbb{R}$, uma escolha ideal da classe de funções $\mathcal{M}$ é considerar como o espaço das funções da segunda derivada do quadrado integrável em $[a, b]$,

$$\mathcal{M} = W_2^2[a, b] = \left\{ f : [a, b] \rightarrow \mathbb{R} : \int_a^b (f(x))^2 dx < \infty, \text{ existe } f''(x) \text{ e } \int_a^b (f''(x))^2 dx < \infty \right\} \quad (2.7)$$

portanto, a função de penalização:

$$\Phi(f) = \lambda \int_a^b (f''(x))^2 dx, \quad \lambda > 0 \quad (2.8)$$

em que λ é um parâmetro positivo e regulador da penalização chamado “parâmetro de suavização”. Reescrevendo (2.6), a minimização da soma de quadrados penalizados está definido por

$$S(f) = \min_{f \in W_2^2[a, b]} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx \right\}. \quad (2.9)$$

como deseja-se procurar uma estimativa de f para ajustar bem aos dados e ao mesmo tempo que tenha um determinado grau de suavidade, essa última expressão é a combinação de dois critérios: uma medida de bondade de ajuste para os dados que é determinada por $\sum_{i=1}^n (y_i - f(x_i))^2$; e por outra parte, uma medida de suavidade para f expressada por $\int_a^b (f''(x))^2 dx$, (DIAS, 2001).

Se o parâmetro de suavização λ é muito grande, a contribuição sobre a componente principal em $S(f)$ será no termo da penalização, aqui a estimativa $\hat{f}(x)$ mostrará uma pequena curvatura; isto é, no caso que λ tende para o infinito, o termo $\int_a^b (f''(x))^2 dx$ tenderá para zero e a curva $\hat{f}(x)$, se aproximará a uma reta. No entanto, se λ é bastante pequeno, então a contribuição principal para $S(f)$ será na soma de quadrados dos resíduos, logo a curva $\hat{f}(x)$ irá interpolar os dados.

Portanto, para um valor dado λ e minimizando $S(f)$ se determina um melhor valor entre suavizado e a bondade de ajuste.

2.5 Splines

Considerando o modelo de (2.4), como uma forma de representar a função de regressão $f(x)$, em termos dos *splines*. Um *spline*, é uma função flexível capaz de se moldar aos dados; portanto, é uma curva definida por segmentos de funções polinomiais, que são utilizadas para aproximar curvas com formas mais complicadas. Tal função de regressão $f(x)$, recebe o nome de “função suave”.

Algumas referências em destaque relacionados ao estudo dos *splines*, são: WAHBA (1990), GREEN e SILVERMAN (1994), quem apresentam trabalhos matemáticos sobre os *splines* e GU e WAHBA (1991) que apresentam um rigoroso estudo para a seleção de modelos e inferência com Modelos Aditivos Generalizados (MAGs) construídos a partir dos *splines*, mas esses modelos são computacionalmente muito complexos.

Definição 2.5.1 Um spline de grau p com nós em $\xi_1, \dots, \xi_k$ é qualquer função da forma

$$f(x) = \sum_{i=0}^{p-1} \theta_i x^i + \sum_{i=1}^k \delta_i (x - \xi_i)_+^{p-1},$$

em que $(t)_+ = \max(0, t)$ e os coeficientes $\theta_0, \dots, \theta_{p-1}, \delta_1, \dots, \delta_k$, são números reais. Portanto, verificamos que:

- $f(x)$ é um polinômio por partes de grau p em qualquer subintervalo $[\xi_i, \xi_{i+1})$;
- $f(x)$ tem $(p - 2)$ derivadas contínuas;

- $f(x)$ tem $(p - 1)$ derivadas contínuas em cada ξ_i , portanto é uma função escada com saltos em $\xi_1, \dots, \xi_k$.

Segundo a definição 2.5.1, pode-se concluir que um *spline* é um espaço vetorial cuja dimensão é $p + k$, desde que as funções $1, x, \dots, x^{p-1}, (x - \xi_1)_+^{p-1}, \dots, (x - \xi_k)_+^{p-1}$, sejam linearmente independentes.

Com todas as propriedades mencionadas anteriormente, pode-se dizer: uma *spline* é uma curva formada por segmentos polinomiais unidos nos pontos $\xi_1, \dots, \xi_k$ chamados “nós”, com o objetivo de garantir continuidade em cada nó.

Basicamente, os *splines* dependem de três elementos, ou seja do grau do polinômio, do número de nós e da localização dos nós, os quais serão explicados segundo o tipo de métodos de suavização utilizando *splines*.

2.6 Métodos de suavização utilizando *splines*

A estimação da função de regressão ou curva $f(x)$ do modelo (2.4) utilizando *splines* pode ser realizada por diferentes métodos de suavização utilizando como base *splines*. No presente trabalho abordaremos os principais, dentre eles encontram-se:

- a) Suavização por *splines* (*Smoothing splines*): o modelo é ajustado pela soma de quadrados penalizados,
- b) *Splines* de regressão (*Regression splines*): o modelo é ajustado por mínimos quadrados.

2.6.1 Suavização por *splines*

Suavização por *splines* chegou a ser um método estabelecido para ajustar curvas mais flexíveis determinado pelos dados. Na utilização do conceito de suavização por *splines*, deve-se assumir a definição de espaço de Sobolev como expresso em (2.3), com as observações x_i no intervalo $[a, b] \subseteq \mathbb{R}$, e que $f(x)$ seja a escolha ideal da classe de funções considerando como o espaço das funções da segunda derivada do quadrado integrável em $[a, b]$,

$$f(x) \in W_2^2[a, b] = \left\{ f : [a, b] \rightarrow \mathbb{R} : \int_a^b (f(x))^2 dx < \infty, \text{ existe } f''(x) \text{ e } \int_a^b (f''(x))^2 dx < \infty \right\} \quad (2.10)$$

Como deseja-se obter uma estimativa de $f(x)$ que se ajuste bem aos dados e ao mesmo tempo que tenha um determinado grau de suavidade, é utilizando mínimos quadrados penalizados

$$S(f) = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx, \quad (2.11)$$

em que $f(x)$ é gerada a partir de uma base de *splines* e $\lambda > 0$, sendo o parâmetro de suavização que determina o grau de suavidade da estimativa, controlando o quanto estamos indo na direção da interpolação dos dados ou na direção de suavização excessiva. Segundo DIAS (1997): “o parâmetro de suavização é o que governa o balanço entre o critério de bondade de ajuste e o critério de suavização”, logo

- Se λ é grande, é devido ao critério de suavização;
- Se λ é pequeno, é devido ao critério de qualidade de ajuste.

Uma característica de suavização por *splines*, evita o problema da seleção de nós, já que eles utilizam como entrada as próprias covariáveis (x_i) como nós e ao mesmo tempo, controlam o super ajuste, reduzindo os coeficientes da função estimada. As bases utilizadas nas funções suaves são os *splines* cúbicos.

2.6.1.1 Spline cúbico

Segundo GREEN e SILVERMAN (1994): “ Seja $x_1, \dots, x_n$ sobre o intervalo $[a, b]$, satisfazendo que $a < x_1 < x_2 < \dots < x_n < b$. A função $f(x)$ definida sobre o intervalo $[a, b]$ é uma *spline* cúbico, se duas condições são satisfeitas. Primeiro, em cada intervalo $(a, x_1), (x_1, x_2), (x_2, x_3), \dots, (x_n, b)$, $f(x)$ é um polinômio cúbico; segundo os segmentos polinomiais ajustam-se nos pontos x_i de tal forma que a primeira e segunda derivadas são contínuas em cada x_i . A seguir um exemplo de um *spline* cúbico

$$f(x) = d_i(x - x_i)^3 + c_i(x - x_i)^2 + b_i(x - x_i) + a_i, \quad \text{para } x_i \leq x \leq x_{i+1} \quad (2.12)$$

com coeficientes a_i, b_i, c_i, d_i , $i = 0, \dots, n$, e define-se $t_0 = a$ e $t_{n+1} = b$.

Um *spline* cúbico no intervalo $[a, b]$ será dito um “*spline* cúbico natural”, se a segunda e terceira derivadas são zeros nos extremos a e b .

Na Figura 2.2 mostra-se as curvas ajustadas utilizando a abordagem suavização por *splines*, para diferentes valores nos parâmetros de suavização e com quantidades diferentes nos nós. Para esse gráfico e os demais que serão apresentados, foi necessário simular um conjunto de 200 observações, utilizando como função, $f(x) = \cos(5\pi x)$, e cuja variável independente x varia no intervalo $[0, 1]$. Pode-se observar que as curvas que melhor ajustam-se aos dados, são aquelas que foram consideradas com 15, 25 e 35 nós em que o parâmetro de suavização mostra uma estabilidade aproximada de 0.00026 respectivamente.

Figura 2.2: Curvas ajustadas utilizando suavização por *splines*, com 5, 15, 25 e 35 nós, para diferentes valores dos parâmetros de suavização para a função $f(x) = \cos(5\pi x)$, para um conjunto de 200 dados simulados.

2.6.2 Splines de Regressão

Segundo MA e RACINE (2013), a diferença fundamental entre suavização por *splines* e *splines* de regressão, é que na primeira abordagem penalizam explicitamente à falta de suavidade e usam os próprios dados como nós potenciais, enquanto os *splines* de regressão colocam nós em pontos equidistantes e equitativos e nenhuma penalidade à falta de suavidade é utilizada como acontece na suavização por *splines*.

Os *splines* de regressão estão formados pela combinação linear de um conjunto de funções de base *splines* da forma $\sum_{j=1}^{b+p+1} a_j B_j(x_i; p)$, em que $a_1, \dots, a_{b+p+1}$, são os coeficientes e $B_1, \dots, B_{b+p+1}$ são as funções de base de potência truncada de p -ésima ordem sobre os b nós: $\xi_1, \dots, \xi_b$.

Considerando a expressão (2.4), sendo que $f(x_i) = \sum_{j=1}^{b+p+1} a_j B_j(x_i; p)$; os coeficientes $a_1, \dots, a_{b+p+1}$ são estimados por mínimos quadrados a partir de,

$$S(f(x)) = \sum_{i=1}^n \left(y_i - \sum_{j=1}^{b+p+1} a_j B_j(x_i; p) \right)^2, \quad (2.13)$$

obtendo os estimadores dos coeficientes da regressão,

$$\hat{a}_j = (B_j' B_j)^{-1} B_j' y_i, \quad (2.14)$$

portanto, a curva ajustada é dada por

$$\hat{f}(x) = \sum_{j=1}^{b+p+1} \hat{a}_j B_j(x). \quad (2.15)$$

que é conhecida como “*splines* de regressão”. A estimativa de $f(x)$ é calculada por mínimos quadrados, neste método nenhuma penalidade à falta de suavidade é utilizada como acontece na suavização por *splines*.

A metodologia mencionada para o caso de uma única covariável, será generalizada em termos matriciais para um melhor desenvolvimento dos temas apresentados a seguir.

Considerando o modelo (2.4) na forma matricial, expressa por

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (2.16)$$

em que $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$, a curva do modelo $\mathbf{f}(\mathbf{x}) = \mathbf{B}\mathbf{a}$, e a base da regressão $\mathbf{B} \in \mathbb{R}^{n \times (b+p+1)}$ chamada “matriz base” (contém qualquer tipo de funções de bases) construída a partir de x , e definida por, $\mathbf{B} = \mathbf{B}(x) = (B_1(x_i), \dots, B_{b+p+1}(x_i))$, com elementos

$$B_{ij} = B_j(x_i; p), \quad i = 1, \dots, n, \quad j = 1, \dots, b + p + 1. \quad (2.17)$$

Significando que a j -ésima coluna de $\mathbf{B}$ avalia os B_j sobre os pontos $x_1, \dots, x_n$; e $\mathbf{a}$ é o vetor de coeficientes da regressão. Logo pode-se reescrever (2.13) como

$$S(f) = \|\mathbf{y} - \mathbf{B}\mathbf{a}\|^2 = (\mathbf{y} - \mathbf{B}\mathbf{a})'(\mathbf{y} - \mathbf{B}\mathbf{a}), \quad (2.18)$$

derivando a expressão (2.18) e igualando a zero, obtém-se a solução explícita dos coeficientes da regressão,

$$\hat{\mathbf{a}} = (\mathbf{B}'\mathbf{B})^{-1} \mathbf{B}'\mathbf{y} \quad (2.19)$$

com valores ajustados

$$\hat{\mathbf{y}} = \mathbf{B}\hat{\mathbf{a}} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} \mathbf{B}'\mathbf{y}, \quad (2.20)$$

Portanto, a estimativa do *spline* de regressão, é dado por,

$$\hat{f}(x) = \mathbf{B}'\hat{\mathbf{a}} \quad (2.21)$$

Os *splines* de regressão são suavizadores lineares determinados pelo número e posição de nós selecionados; isto significa, a quantidade de suavizado é determinado pelo pesquisador, e muitas vezes a solução 2.21 apresenta um grande problema ou seja, o super ajuste (*overfit*) nos dados, o que piora quando se aumenta o número de bases. Na Figura 2.3, pode-se observar que a medida em que se aumenta o número de nós, estaríamos chegando a uma interpolação dos dados; portanto, a curva ajustada é menos suave e é mais ondulada. Neste caso, muitos autores propuseram o uso de uma penalidade para controlar a suavidade do ajuste, dentre eles O’SULLIVAN (1986); EUBANK (1988); GREEN e SILVERMAN (1994); WAHBA (1990) e EILERS e MARX (1996).

Figura 2.3: Curvas ajustadas utilizando *Splines* de regressão com 5, 15, 25 e 35 nós, para a função $f(x) = \cos(5\pi x)$, para um conjunto de 200 dados simulados.

2.6.3 Suavização por splines penalizados

O termo *splines* penalizados surgiu para descrever um *spline* geral ajustado sujeito a uma penalização (esse tipo de penalização difere segundo o tipo de suavização a ser utilizado com os *splines*). Os *splines* penalizados foram introduzidos por O’SULLIVAN (1986) os quais são uma classe de *splines* penalizados baseados em funções de bases B-*splines* com uma aproximação de penalidade discreta que é uma aproximação da segunda derivada do quadrado integrável da curva ajustada.

Portanto, os *splines* penalizados, combinam o melhor de duas abordagens de suavização utilizando *splines*: da suavização por *splines* tem a ideia da penalização e dos *splines* de regressão, expressa a função como combinação linear de B-*splines*, logo utiliza-se menos parâmetros que os apresentados na suavização por *splines*, mas a seleção dos nós não é tão determinante como nos *splines* de regressão.

Os *splines* penalizados (O’SULLIVAN, 1986; RUPPERT *et al.*, 2003) e os P-*splines* (EILERS e MARX, 1996, 2010) e (EILERS *et al.*, 2015), utilizam bases completas sobre um conjunto de nós reduzidos com uma aproximação discreta à penalidade da suavização por *splines*.

2.6.3.1 *Splines* penalizados de O’Sullivan

Como dito na seção anterior, os *splines* penalizados foram introduzidos por O’SULLIVAN (1986), e a seguir tem-se uma breve apresentação desse tipo de *splines*:
Considerando uma simples regressão não-paramétrica,

$$y_i = f(x_i) + \epsilon_i, \quad 1 \leq i \leq n, \quad \epsilon \sim N(0, \sigma_\epsilon^2) \quad (2.22)$$

em que $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. A função suave $f(x)$ será definida como, $f(x_i) = a_j B_j(x_i)$, em que $B_1, \dots, B_{b+4}$, são as funções de base B-*spline*, $a_1, a_2, \dots, a_{b+4}$, os coeficientes dessas bases, e como $f(x)$ está no intervalo $[x_{min}, x_{max}]$ que contém os valores de x'_i s. Para um inteiro $b \leq n$, define-se uma sequência de nós:

$\xi_1, \dots, \xi_{b+8}$, tal que

$$x_{min} = \xi_1 = \xi_2 = \xi_3 = \xi_4 < \xi_5 < \dots < \xi_{b+4} < \xi_{b+5} = \xi_{b+6} = \xi_{b+7} = \xi_{b+8} = x_{max}.$$

A estimativa de $f(x)$ é calculada utilizando o critério de otimização, tendo em vista que O'SULLIVAN (1986) introduziu uma penalização definida como a segunda derivada do quadrado integrável nas funções de base. Utilizando 2.11 obtém-se o critério de otimização,

$$S(f) = \sum_{i=1}^n \left(y_i - \sum_{j=1}^b a_j B_j(x_i) \right)^2 + \lambda \int_{x_{min}}^{x_{max}} \left(\sum_{j=1}^b a_j B_j''(x_i) \right)^2 dx, \quad (2.23)$$

sendo que a segunda parte de (2.23) corresponde à penalização e pode-se escrever em termos da matriz de bases, $\mathbf{B}$ de ordem $n \times (b+4)$ com entradas $B_j(x_i)$. Essa penalização é denotada por, $\mathbf{\Omega}$ cuja matriz é de ordem $(b+4) \times (b+4)$ definida por

$$\mathbf{\Omega} = \int_{x_{min}}^{x_{max}} B_j''(x) B_{j'}''(x) dx$$

A penalização está interpretada simplesmente como “curvatura”. Portanto, uma estimativa de $f(x)$ na posição $x \in \mathbb{R}$ pode ser obtida como

$$\hat{f}(x) = \mathbf{B}_x \hat{\mathbf{a}} \quad \text{em que} \quad \hat{\mathbf{a}} = (\mathbf{B}'\mathbf{B} + \lambda\mathbf{\Omega})^{-1} \mathbf{B}'\mathbf{y}$$

com $\mathbf{B}_x = [B_1(x), \dots, B_{b+4}(x)]$ e $\lambda > 0$ o parâmetro de suavização. Posteriormente, EILERS e MARX (1996) generalizaram o problema para a p -ésima derivada e a ideia foi substituir a penalização de O'Sullivan por outra baseada em diferenças finitas dos coeficientes adjacentes dos B-*splines*.

2.6.4 Suavização por P-*splines*

Uma classe de *splines* penalizados mais próximos aos *splines* penalizados de O'Sullivan são os P-*splines* de EILERS e MARX (1996). O nome de P-*splines* é devido a uma simples combinação de duas ideias para a curva ajustada: regressão sobre as funções de bases B-*splines* e uma penalidade de diferenças sobre os coeficientes da regressão dos B-*splines*, utilizando nós igualmente espaçados. Portanto, a penalidade é a parte principal do ajuste da curva; significando que a suavização da curva é ajustada pela alteração dos pesos dos coeficientes da regressão.

RUPPERT e CARROLL (2000) propuseram a utilização dos P-*splines* utilizando funções de base de potência truncada com nós desigualmente espaçados (utilizando os quantis de x) e uma penalidade *ridge*.

Seguindo a aproximação de EILERS e MARX (1996), utilizaram uma penalidade baseada nas “diferenças de ordem d ” entre os coeficientes adjacentes das bases de B-*splines*, esse tipo de penalidade é mais flexível por ser independente do grau do polinômio utilizado para construir as bases B-*splines*, essa abordagem é uma boa aproximação discreta do quadrado da d -ésima diferença dos coeficientes das bases B-*splines*. Reescrevendo a Equação (2.23) em termos das diferenças de ordem d , obtém-se o critério de otimização de mínimos quadrados penalizados

$$S(f) = \sum_{i=1}^n (y_i - \sum_{j=1}^b a_j B_j(x_i))^2 + \lambda \sum_{j=d+1}^b (\Delta^d a_j)^2 dx \quad (2.24)$$

em que Δ^d é operador diferença de ordem d para a_j . Para o vetor dos coeficientes da regressão $\mathbf{a}$, o operador diferença é definido recursivamente por

$$\begin{aligned}\Delta^1 a_j &= a_j - a_{j-1} \\ \Delta^2 a_j &= \Delta^1(\Delta^1 a_j) = a_j - 2a_{j-1} + a_{j-2} \\ &\vdots \\ \Delta^d a_j &= \Delta^1(\Delta^{d-1} a_j).\end{aligned}$$

Além disso, Δ^d , é uma matriz $[(b+1+p-d) \times (b+1+p)]$ do operador diferença de ordem, $d = p-1$; essa ordem da penalidade d , “controla” a mudança entre os coeficientes adjacentes das bases B-splines. A primeira diferença de ordem $d = 1$, penaliza os saltos entre os coeficientes sucessivos, e a segunda diferença de ordem, penaliza os desvios da tendência linear ($2a_{j-1} + a_{j-2}$), tendo como resultado o ajuste do P-splines (EILERS e MARX, 1996). Por exemplo, a matriz diferença de ordem $d = 2$ ou de segunda ordem Δ^2 , com $b = 4$ nós e $p = 3$ grau polinomial da base B-spline, obtém-se a seguinte matriz

$$\Delta^2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix},$$

com base na matriz diferença Δ^d , e reescrevendo 2.24 na forma matricial,

$$S(f) = (\mathbf{y} - \mathbf{B}\mathbf{a})'(\mathbf{y} - \mathbf{B}\mathbf{a}) + \lambda \mathbf{a}' \mathbf{P}_d \mathbf{a}, \quad (2.25)$$

em que $\mathbf{P}_d = \lambda(\Delta^d)'(\Delta^d)$, é o termo da penalidade sendo uma matriz quadrada, semidefinida positiva e simétrica. A função da penalidade é forçar aos coeficientes para que possam variar suavemente e consequentemente obter uma curva suave. Como qualquer outra técnica de suavização, o parâmetro λ é conhecido como o “parâmetro de suavização” e serve para controlar a suavidade do ajuste da curva.

Para um valor dado $\lambda > 0$, os valores ajustados dos coeficientes das bases B-splines obtidos pela solução da soma de quadrados penalizados de (2.25), agora são definidos por

$$\hat{\mathbf{a}} = (\mathbf{B}'\mathbf{B} + \lambda \mathbf{P}_d)^{-1} \mathbf{B}'\mathbf{y}, \quad (2.26)$$

ou que pode ser expressada também como

$$(\mathbf{B}'\mathbf{B} + \lambda \mathbf{P}_d) \hat{\mathbf{a}} = \mathbf{B}'\mathbf{y}. \quad (2.27)$$

com curva ajustada,

$$\hat{f}(\mathbf{x}) = \mathbf{B}'\hat{\mathbf{a}}. \quad (2.28)$$

A forma da curva ajustada, $\hat{f}(\mathbf{x})$, é influenciada pelo valor do parâmetro de suavização λ , que controla a mudança entre o modelo ajustado e o modelo suavizado, portanto:

- Quando $\lambda \rightarrow \infty$, a curva ajustada $\hat{f}(\mathbf{x})$ tende a um polinômio de grau $(d-1)$, e acontece se o grau do B-splines é igual ou maior à ordem da penalidade;
- Quando $\lambda = 0$, o resultado se reduz a simples equações normais de uma regressão linear de $\mathbf{y}$ sobre $\mathbf{B}$, cuja solução $\hat{\mathbf{a}}$ se reduz a, $\hat{\mathbf{a}} = (\mathbf{B}'\mathbf{B})^{-1} \mathbf{B}'\mathbf{y}$, que é exatamente igual a expressão (2.19) de splines de regressão.

Portanto, observa-se que para ajustar uma curva do modelo depende da seleção do parâmetro de suavização λ e o tipo de penalidade a ser utilizada dependerá do tipo de base (EILERS e MARX, 1996). No caso de bases com polinômios truncados, utiliza-se uma penalidade *ridge* independente do grau do polinômio truncado, assim

$$S(f) = (\mathbf{y} - \mathbf{B}\mathbf{a})'(\mathbf{y} - \mathbf{B}\mathbf{a}) + \lambda \mathbf{a}'\mathbf{a}.$$

Na Figura 2.4, pode-se observar que a curva que melhor se ajustam aos dados, são as que utilizam suavização com P-splines com 25 e 35 nós. No entanto, utilizando para esse caso menor quantidade de nós, pode-se observar que o ajuste tende a ser linear.

Figura 2.4: Curvas ajustadas utilizando P-splines, com 5, 15, 25 e 35 nós, com diferentes valores dos parâmetros de suavização para a função $f(x) = \cos(5\pi x)$, com 200 dados simulados.

2.6.5 Bases

Nesta seção e conforme RUPPERT *et al.* (2003), são tratados os tipos de funções de bases utilizados na penalidade com P-splines, mostrando como construir essas funções de bases partindo de um simples modelo de regressão linear. Considerando o seguinte modelo:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (2.29)$$

com $i = 1, \dots, n$. A correspondente “base” do modelo é dada pelos vetores

$$\mathbf{1} = [1, \dots, 1]' \quad e \quad \mathbf{x} = [x_1, \dots, x_n]', \quad (2.30)$$

que são “funções de bases”. Nota-se que o lado direito de (2.29) é uma combinação linear das funções de (2.30). Portanto, essas “funções de base” correspondem às colunas da matriz $\mathbf{X}$, expressa por

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

sendo $\mathbf{X}$ uma base para o espaço vetorial de todos os polinômios lineares em x para ajustar a uma regressão. O vetor de valores estimados, $\hat{\mathbf{y}}$, pode ser obtido da matriz $\mathbf{X}$ e do vetor $\mathbf{y}$, através da seguinte expressão:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}. \quad (2.31)$$

em que $\mathbf{H}$ é a matriz de projeção.

Uma simples extensão do modelo linear é o modelo quadrático, ou seja,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad (2.32)$$

com $i = 1, \dots, n$, em que β_0, β_1 e β_2 , são os coeficientes do modelo de regressão. Nota-se que (2.32) tem uma função de base adicional x^2 , que corresponde à adição do termo $\beta_2 x_i^2$ no modelo; logo a matriz $\mathbf{X}$ para o modelo quadrático é dada por:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

e vetor ajustado $\hat{\mathbf{y}}$, é obtido utilizando a expressão (2.31) com esta nova matriz $\mathbf{X}$. O modelo quadrático é um exemplo de como um modelo linear simples, pode ser estendido a uma estrutura não linear.

Considere-se outro tipo de extensão de estrutura não linear, o modelo *Broken stick*, assim chamado por ser composto por duas linhas de diferentes inclinações que se encontram em um único ponto $x = \xi$, a introdução de uma função de base que é zero a esquerda de ξ , se obtem uma função com inclinação de coeficiente angular positiva de ξ .

Assim, a forma matemática de expresar essa função de base é dada por $(x - \xi)_+$, para algum

$$(x - \xi)_+ = \begin{cases} (x - \xi), & \text{se } (x - \xi) > 0; \\ 0, & \text{em outro caso.} \end{cases}$$

Considerando o modelo *Broken stick* com o ponto de ruptura em $x = 0.6$, dado por

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.6)_+ + \epsilon_i, \quad i = 1, \dots, n \quad (2.33)$$

logo, a matriz $\mathbf{X}$ para o modelo (2.33) é

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1 - 0.6)_+ \\ 1 & x_2 & (x_2 - 0.6)_+ \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - 0.6)_+ \end{bmatrix}$$

o vetor ajustado $\hat{\mathbf{y}}$ é obtido utilizando (2.31) com esta nova matriz $\mathbf{X}$.

Pode-se obter uma estrutura mais complexa adicionando mais funções da forma $(x - \xi)_+$ para a base, ou aumentando colunas de valores $(x_i - \xi)_+$ à matriz $\mathbf{X}$ e o valor de ξ da função $(x - \xi)_+$, usualmente refere-se a um *nó*. Um exemplo da estrutura da matriz $\mathbf{X}$ com k nós em 0.5, 0.55, ..., 0.95 é dado por

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1 - 0.5)_+ & (x_1 - 0.55)_+ & \dots & (x_1 - 0.95)_+ \\ 1 & x_2 & (x_2 - 0.5)_+ & (x_2 - 0.55)_+ & \dots & (x_2 - 0.95)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - 0.5)_+ & (x_n - 0.55)_+ & \dots & (x_n - 0.95)_+ \end{bmatrix}.$$

Uma função com a estrutura $(x - 0.6)_+$, é chamada função de base *splines* linear e o conjunto de tais funções é chamado de “base *splines* linear”. Nota-se que, a combinação linear de funções de base *splines* linear: $1, x, (x - \xi_1)_+, \dots, (x - \xi_b)_+$, é uma função linear por partes com nós em $\xi_1, \dots, \xi_b$, definida por

$$f(x) = \beta_0 + \beta_1 x + \sum_{j=1}^b \beta_{1j} (x - \xi_j)_+,$$

em que $(x - \xi_j)_+ = \max[0, (x - \xi_j)]$, (diz-se, truncado já que a partir de certo ponto vale 0). Neste texto, discute-se a natureza linear das funções e que são expressas como uma combinação linear de funções lineares por partes da forma, $(x - \xi)_+$. Segundo RUPPERT *et al.* (2003), uma maneira simples de escapar da linearidade por partes é adicionar x^2 à base e substituir cada $(x - \xi)_+$ por seu quadrado $(x - \xi)_+^2$, tal que esta função tem continuidade na primeira derivada. A combinação linear dessas funções é dada por

$$1, x, x^2, (x - \xi_1)_+^2, \dots, (x - \xi_b)_+^2, \quad (2.34)$$

Lê-se “funções de base *splines* quadráticos” com nós: $\xi_1, \dots, \xi_b$. Uma generalização desses modelos, são os modelos *splines* de grau geral; isto é, utilizando funções de base de potência truncada, logo a base é dada por

$$1, x, \dots, x^p, (x - \xi_1)_+^p, \dots, (x - \xi_b)_+^p, \quad (2.35)$$

tal modelo é conhecido como “base de potência truncada de grau p ”. A função $(x - \xi_b)_+^p$, tem $(p - 1)$ derivadas contínuas, e maiores valores de p fazem com que as funções sejam mais suaves. Portanto, o modelo *splines* de grau p é dado por,

$$f(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{j=1}^b \beta_{pj} (x - \xi_j)_+^p. \quad (2.36)$$

que permite a continuidade na derivada de ordem $(p - 1)$ em cada nó ξ_j .

2.6.5.1 Bases polinomiais truncadas

Considerando a expressão (2.36), sabe-se que $f(x)$ é a função suave não especificada da covariável x ; e considerando n observações, com $i = 1, \dots, n$. Pode-se reescrever

$$f(x_i) = \sum_{l=0}^p \beta_l x_i^l + \sum_{j=1}^b \beta_{pj} z_{ij}, \quad (2.37)$$

em que $z_{ij} = (x_i - \xi_j)_+^p = (x_i - \xi_j)^p I(x_i > \xi_j)$, sendo $I(x_i > \xi_j)$ a função indicadora dada por;

$$I(x_i > \xi_j) = \begin{cases} 1, & \text{se } (x_i - \xi_j) \geq 0; \\ 0, & \text{se } (x_i - \xi_j) < 0. \end{cases}$$

O vetor $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_b)$, contém os nós que são escolhidos como quantis da variável x para determinar a posição desses nós e geralmente não são igualmente espaçados. Expressando (2.37) na forma matricial,

$$f(\mathbf{x}) = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 = \mathbf{X} \boldsymbol{\beta}, \quad (2.38)$$

com $\mathbf{X} = [\mathbf{X}_1 : \mathbf{X}_2]$, em que $\mathbf{X}_1 = [1, x_i, \dots, x_i^p]_{1 \leq i \leq n}$, é a matriz base dos polinômios da covariável x ; $\mathbf{X}_2 = [(x_i - \xi_1)_+^p, (x_i - \xi_2)_+^p, \dots, (x_i - \xi_b)_+^p]_{1 \leq i \leq n}$, é a estrutura da matriz que contém as funções de bases de polinômios truncados; com vetor de coeficientes $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, em que, $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, \dots, \beta_p)'$, vetor de coeficientes de regressão de $\mathbf{X}_1$ e $\boldsymbol{\beta}_2 = (\beta_{p1}, \beta_{p2}, \dots, \beta_{pb})'$ o vetor de coeficientes das bases de potências

truncadas de $\mathbf{X}_2$. A matriz base do p -ésimo grau de polinômios truncados, é dado por

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p & (x_1 - \xi_1)_+^p & \cdots & (x_1 - \xi_b)_+^p \\ 1 & x_2 & \cdots & x_2^p & (x_2 - \xi_1)_+^p & \cdots & (x_2 - \xi_b)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p & (x_n - \xi_1)_+^p & \cdots & (x_n - \xi_b)_+^p \end{bmatrix}$$

e portanto, utilizando a expressão (2.25) e o critério de otimização para minimizar a função, tem-se

$$S(f) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta}, \quad (2.39)$$

em que $\mathbf{P} = \text{diag}(\mathbf{0}_{p+1}, \mathbf{I}_b)$, e observa-se que é uma penalidade *ridge* sobre $\boldsymbol{\beta}$. A base de polinômios truncados é simples mas não é sempre numericamente estável; o que acontece, quando o número de nós é grande e o parâmetro de suavização λ é próximo de zero. Na Figura 2.5 pode-se verificar alguns gráficos de funções de base de potência truncada com diferentes graus de polinômios, bem como o comportamento de tais polinômios. Assim a Figura 2.5.(a) representa uma base com polinômio de grau 0, portanto o resultado é de linhas constantes, na Figura 2.5.(b) apresenta linhas paralelas com inclinação positiva por ser utilizado polinômio de grau 1; no entanto a Figura 2.5.(c) corresponde a base com polinômios de grau 2 o que resulta numa família de parábolas e na Figura 2.5.(d) está representada uma base com polinômios de grau 3.

Figura 2.5: Bases de polinômios truncados com diferentes graus: (a) Grau 0, (b) Grau 1, (c) Grau 2 e (d) Grau 3.

2.6.5.2 Bases B-splines

Entre as principais referências básicas de B-splines pode-se citar DE BOOR (1979) que define matematicamente os B-splines. EILERS e MARX (1996), utiliza os B-splines como funções de base no caso de P-splines. Segundo BIEZUNER (2014), para ilustrar a ideia básica, define-se: “B-splines são curvas bem formadas e compostas por segmentos polinomiais interligados”.

Seja p um inteiro não negativo e uma sequência de nós (pontos de controle) não decrescente dado pelo vetor $\xi = (\xi_1, \xi_2, \dots, \xi_b)$, de comprimento $(p + 2)$. O j -ésimo B-spline de grau p é uma curva formada por segmentos polinomiais interligados nos nós, definida por

$$B_j^p(\xi) = S_1(\xi)B_j^{p-1}(\xi) + S_2(\xi)B_{j+1}^{p-1}(\xi), \quad (2.40)$$

em que

$$S_1(\xi) = \frac{\xi - \xi_j}{\xi_{j+p} - \xi_j} \quad \text{e} \quad S_2(\xi) = \frac{\xi_{j+1+p} - \xi}{\xi_{j+1+p} - \xi_{j+1}}, \quad (2.41)$$

para todo número real ξ com

$$B_j^0(\xi) = \begin{cases} 1, & \text{se } \xi_j \leq \xi < \xi_{j+1} \\ 0, & \text{caso contrário} \end{cases}.$$

Logo, a relação de recorrência é expressa por

$$B_j^p(\xi) = \begin{cases} 0, & \text{se } \xi_j = \xi < \xi_{j+1+p}; \\ S_1(\xi), & \text{se } \xi_j < \xi_{j+p} \text{ e } \xi_{j+1} = \xi_{j+1+p}; \\ S_2(\xi), & \text{se } \xi_j = \xi_{j+p} \text{ e } \xi_{j+1} < \xi_{j+1+p}; \\ S_1(\xi) + S_2(\xi), & \text{caso contrário.} \end{cases}$$

B-splines são muito populares entre os estatísticos pela flexibilidade e fácil desenvolvimento numérico. Um exemplo muito simples e conforme sugere DURBÁN (2014) é mostrado na Figura 2.6 onde na Figura 2.6(a): um B-splines de grau 1 formado por duas partes ou segmentos de polinômios lineares; uma linha de x_1 a x_2 , a outra de x_2 a x_3 . Os nós são x_1, x_2 e x_3 . A esquerda de x_1 e à direita de x_3 , este B-splines é zero. Na Figura 2.6(b), uma sequência de B-splines de grau 1 é mostrada, cada um baseado sobre 3 nós. Análogamente, na Figura 2.6(c), é mostrado um B-splines de grau 3, formado por 5 nós adjacentes: x_1, x_2, x_3, x_4 e x_5 , e 4 partes polinomiais cúbicas interligadas pelos nós x_2, x_3 e x_4 . Na Figura 2.6(d) observa-se a sequência das funções da base que tem a mesma forma, mas estão deslocados horizontalmente e o deslocamento é uma função da distância entre os nós.

Em geral, com base nesses exemplos da Figura 2.6 e segundo EILERS e MARX (1996), as propriedades gerais de um B-splines de grau p são:

- constituído de $(p + 1)$ partes de um polinômio de grau p ,
- as partes polinomiais são unidas em p nós internos,
- nos pontos de conexão, as derivadas até a ordem $(p - 1)$ são contínuas,
- B-splines é positivo no domínio expandido por $(p + 2)$ nós e zero em caso contrário,
- para cada valor x , $(p + 1)$ B-splines são não nulos.

O intervalo de x é dividido por nós, em que as funções de bases são calculadas, tal que, $x_{min} = \xi_1 < \xi_2 < \dots < \xi_b = x_{max}$, cada intervalo é constituído por $(p + 1)$ B-splines de grau p . O número total de nós para a construção dos B-splines será $(b + 2p + 1)$ nós, e o número de B-splines para uma regressão base, ou número de colunas de $\mathbf{B}$, é $k = b + p$.

Figura 2.6: Bases de B-splines: (a) B-splines de grau 1, (b) sequência de B-splines de grau 1, (c) B-splines de grau 3 e (d) sequência de B-splines de grau 3.

É possível escolher a localização dos nós na maioria das situações, mas a sugestão de RUPPERT e CARROLL (2000), é utilizar uma quantidade maior de nós igualmente espaçados entre 20 e 40 nós.

A Figura 2.7 mostra os ajustes de uma curva utilizando como bases B-splines sem penalização e com penalização por meio de *splines* de regressão e P-splines respectivamente. A Figura 2.7(a) mostra as funções que formam as bases ou as colunas da matriz $\mathbf{B}$ multiplicadas pelos coeficientes (círculo). Observa-se que o comportamento dos coeficientes dá origem a uma curva pouco suave; no entanto, na Figura 2.7(b), quando se impõem esses coeficientes nessas funções, mostra uma curva mais suave; portanto, esse gráfico mostra o objetivo da penalização: forçar os coeficientes a seguir em um padrão suave.

2.6.6 Propriedades dos P-splines

A seguir são apresentadas algumas propriedades dos P-splines com funções de bases B-splines.

2.6.6.1 Matriz de suavização

Seja o modelo de regressão linear em termos das funções de bases B-splines

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\epsilon}, \quad (2.42)$$

O ajuste dos valores de um P-spline utilizando o critério de mínimos quadrados penalizados de (2.25) é dado por

$$\hat{\mathbf{y}} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{P}_d)^{-1}\mathbf{B}'\mathbf{y} = \mathbf{H}_\lambda\mathbf{y} \quad (2.43)$$

Figura 2.7: Curva ajustada sem penalizar os coeficientes (a) utilizando os *splines* de regressão, e penalizando os coeficientes (b) utilizando os *P-splines*. Para a penalização, considerou-se 30 nós com um conjunto 200 dados simulados para a função $f(x) = \cos(5\pi x)$.

em que

$$\mathbf{H}_\lambda = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{P}_d)^{-1}\mathbf{B}' \quad (2.44)$$

sendo que a matriz $\mathbf{H}_\lambda$ é chamada de “matriz *hat*” ou “matriz de suavização” (RUPPERT *et al.*, 2003). Lembrando que essa não é uma matriz de projeção, por não ser idempotente; isto é, $\mathbf{H}_\lambda^2 \neq \mathbf{H}_\lambda$, e os elementos da matriz $\mathbf{H}_\lambda$, são os valores ajustados $\hat{\mathbf{y}}$ para um valor dado λ , o que faz com que o suavizador seja linear (DURBÁN, 2014).

2.6.6.2 Graus de liberdade efetivos

De acordo com DURBÁN (2014), se no modelo com *P-splines* o parâmetro de suavização $\lambda = 0$, então os graus de liberdade de ajuste do modelo correspondem à dimensão da matriz base $\mathbf{B}$ menos o número de restrições do modelo; no entanto, se λ é grande, então o modelo é pouco flexível, portanto tem poucos graus de liberdade. Essa medida de graus de liberdade de ajuste é correspondente ao valor do parâmetro de suavização λ e para determinar essa medida, utiliza-se o traço da matriz $\mathbf{H}_\lambda$. HASTIE e TIBSHIRANI (1990), propuseram a dimensão efetiva ou “graus de liberdade efetivos”, cuja dimensão é uma boa aproximação da dimensão do vetor de parâmetros definido pela matriz de suavização. Para um modelo com *P-splines* com b nós e diferença de ordem d , essa dimensão é medida por

$$de(\lambda) = \text{traço}(\mathbf{H}_\lambda) = \text{traço}[\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{P}_d)^{-1}\mathbf{B}'], \quad (2.45)$$

em que o traço da matriz dependerá do parâmetro de suavização e obtém-se que o número de valores positivos de λ corresponde aos graus de liberdade efetivos, ou seja

$$d + 1 < de(\lambda) < d + 1 + b$$

isto significa, se $\lambda \rightarrow 0$ então $\text{traço}(\mathbf{H}_\lambda) = d + 1 + b$, no outro extremo se $\lambda \rightarrow \infty$ então $\text{traço}(\mathbf{H}_\lambda) = d + 1$.

2.6.6.3 Graus de liberdade dos resíduos

Como na estimação paramétrica, os graus de liberdade dos resíduos são usados para estimar σ_ϵ^2 . Segundo RUPPERT *et al.* (2003), define-se o valor esperado da soma de quadrados dos resíduos como, $E(SQR) = (n - p)\sigma_\epsilon^2$, em que o fator $(n - p)$, são os graus de liberdade para os resíduos e é usado para corrigir o viés quando usamos a soma de quadrados dos resíduos (SQR) para estimar σ_ϵ^2 , em que n é o

número de observações e p , o número de parâmetros do modelo. Na regressão não-paramétrica, os graus de liberdade para os resíduos dependem do viés introduzido pela matriz de suavização e é definido por,

$$df_{res} = n - 2\text{traço}(\mathbf{H}_\lambda) + \text{traço}(\mathbf{H}_\lambda \mathbf{H}'_\lambda), \quad (2.46)$$

Para obter os graus de liberdade dos resíduos, é utilizado o processo sugerido por RUPPERT *et al.* (2003). Supondo o modelo geral

$$\mathbf{y} = f(\mathbf{x}) + \epsilon, \quad \text{com} \quad E(\mathbf{y}) = f(\mathbf{x}) \quad \text{e} \quad \text{Cov}(\epsilon) = \sigma_\epsilon^2 \mathbf{I},$$

sendo,

$$\hat{f}(\mathbf{x}) = \mathbf{H}_\lambda \mathbf{y},$$

a soma de quadrados residual é definida por, $SQR = \|\hat{f}(\mathbf{x}) - \mathbf{y}\|^2$. Usando o seguinte resultado:

$$E(\mathbf{V}' \mathbf{A} \mathbf{V}) = E(\mathbf{V}') \mathbf{A} E(\mathbf{V}) + \text{traço}\{\mathbf{A} \text{Cov}(\mathbf{V})\},$$

obtém-se,

$$\begin{aligned} E(SQR) &= E \|\hat{f}(\mathbf{x}) - \mathbf{y}\|^2 \\ &= E \|\mathbf{H}_\lambda - \mathbf{I}\| \mathbf{y} \|^2 \\ &= E(\mathbf{y})' \|\mathbf{H}_\lambda - \mathbf{I}\| E(\mathbf{y}) + \text{traço}[(\mathbf{H}_\lambda - \mathbf{I})'(\mathbf{H}_\lambda - \mathbf{I}) \text{Cov}(\epsilon)]. \end{aligned}$$

Desta maneira, a última equação pode ser escrita como:

$$E(SQR) = \|\mathbf{H}_\lambda - \mathbf{I}\| f(\mathbf{x}) \|^2 + \underbrace{\sigma_\epsilon^2 [n - 2\text{traço}(\mathbf{H}_\lambda) + \text{traço}(\mathbf{H}'_\lambda \mathbf{H}_\lambda)]}_{df_{res}}. \quad (2.47)$$

Assumindo que o termo $\|\mathbf{H}_\lambda - \mathbf{I}\| f(\mathbf{x}) \|^2$ é desprezível, obtém-se uma estimativa não viesada para σ_ϵ^2 dado por

$$\frac{SQR}{df_{res}}.$$

Sabe-se que o ajuste de uma curva é de interesse, então é comum para medir o erro global por meio de vários valores, assim tem-se por exemplo, o erro quadrático médio (EQM), definido por

$$EQM[\hat{f}(\mathbf{x})] = E \left\{ \left[\hat{f}(\mathbf{x}) - f(\mathbf{x}) \right]^2 \right\},$$

que também pode ser expressa como

$$EQM[\hat{f}(\mathbf{x})] = \text{Var}[\hat{f}(\mathbf{x})] + \left\{ E[\hat{f}(\mathbf{x})] - f(\mathbf{x}) \right\}^2, \quad (2.48)$$

que representa um balanço da contribuição do erro entre o quadrado de viés e a variância da curva ajustada (RUPPERT *et al.*, 2003). Com base nessa medida, existe um *trade-off* definido pela média do erro quadrático médio ($MEQM$)

$$MEQM(\lambda) = \frac{1}{n} \sum_{i=1}^n EQM[\hat{f}(\mathbf{x})] = \frac{1}{n} \left\{ \|\mathbf{H}_\lambda - \mathbf{I}\| f(\mathbf{x}) \|^2 + \sigma_\epsilon^2 \text{traço}(\mathbf{H}'_\lambda \mathbf{H}_\lambda) \right\} \quad (2.49)$$

O primeiro termo da expressão (2.49) representa a contribuição do quadrado do viés, e o segundo termo corresponde à soma de todas as variâncias. Em geral, com a expressão (2.49) obtém-se valores maiores do parâmetro de suavização λ , pois incrementa-se o viés mas diminui-se a variância. No entanto, para valores menores de λ , o resultado mostra o comportamento contrário (HASTIE *et al.*, 2009).

Para se encontrar um valor ótimo de suavização, utilizam-se funções mais complexas para um melhor ajuste; conforme apresentado a seguir.

2.6.6.4 Seleção do parâmetro de suavização

O papel do parâmetro de suavização nos *P-splines*, é o mesmo que se tem em qualquer outro método de suavização sendo “ controlar a suavidade da curva”, mas neste caso e conforme EILERS e MARX (1996), o que ele faz é “penalizar os coeficientes que estão muito separados entre si” e quanto maior λ , os coeficientes se aproximarão de zero; isto significa, que se $\lambda \rightarrow \infty$ aproxima-se de um ajuste polinomial. Pelo contrário, quando $\lambda \rightarrow 0$, utiliza-se mínimos quadrados ordinários. Na Figura 2.8, encontram-se as curvas ajustadas usando diferentes valores do parâmetro de suavização.

Figura 2.8: Curva estimada usando *P-splines* com diferentes valores de λ , com 200 dados simulados para a função $f(x) = \cos(5\pi x)$

Existem muitos métodos para escolher o valor ótimo de λ , os quais podem ser classificados em dois grandes grupos: métodos baseados em Validação Cruzada (*Cross Validation - CV*) e métodos baseados no critério de informação.

- **Métodos de Validação Cruzada:** A idéia da Validação Cruzada Ordinária, é tirar uma observação e ajustar o modelo com os demais dados e calcular o quadrado da diferença entre o valor perdido e a sua predição, isto é, $\sum_{i=1}^n (y_i - \hat{y}_{-i})^2$, (STONE, 1974). Ao invés de repetir o processo n vezes, uma alternativa eficiente é dada por HASTIE e TIBSHIRANI (1990), ou seja

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{y}_i}{1 - H_{\lambda}^{ii}} \right\}^2, \quad (2.50)$$

em que H_{λ}^{ii} , são os elementos diagonais da matriz de suavização, $\mathbf{H}_{\lambda}$. Uma modificação deste critério, proposto por CRAVEN e WAHBA (1978) e WAHBA (1990) é a Validação Cruzada Generalizada (*Generalized Cross Validation - GCV*), dada por,

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{y}_i}{1 - \frac{\sum_{i=1}^n H_{\lambda}^{ii}}{n}} \right\}^2 = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{y}_i}{1 - \frac{\text{traço}(\mathbf{H}_{\lambda})}{n}} \right\}^2, \quad (2.51)$$

em que $\sum_{i=1}^n H_{\lambda}^{ii}$ é substituído por $\text{traço}(\mathbf{H}_{\lambda})$. A diferença entre as duas quantidades é geralmente pequena; o melhor λ será o valor que minimiza $CV(\lambda)$ ou $GCV(\lambda)$.

- **Métodos de Critério de Informação:** A ideia básica é corrigir o logaritmo da máxima verossimilhança de um modelo ajustado pela dimensão efetiva, análogo ao efetivo número de parâmetros. EILERS e MARX (1996) propuseram minimizar o Critério de Informação:

$$IC = Dev(\mathbf{y}; \lambda) + d \times de(\lambda). \quad (2.52)$$

A *deviance* (Dev) é uma medida de qualidade de ajuste e é dada por,

$$Dev(\mathbf{y}, \hat{\mathbf{y}}) = 2 \{L(\mathbf{y}) - L(\hat{\mathbf{y}})\}, \quad (2.53)$$

em que, $L(\cdot)$, denota a função logaritmo da máxima verossimilhança. Para dados normais, a *deviance* é simplesmente a soma de quadrados residual, $SQR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Para dados não normais, a *deviance* está baseada numa generalização da soma de quadrados, e poderia tomar diferentes expressões. O termo d , penaliza a dimensão efetiva do modelo ($de(\lambda)$) em (2.45). Quando $d = 2$ tem-se o critério de informação de Akaike - *AIC* AKAIKE (1973) e quando $d = \ln(n)$, tem-se o critério de informação Bayesiano - *BIC*, (SCHWARZ, 1978) respectivamente. O *BIC* penaliza o modelo mais “complexo” que o *AIC*, especialmente quando n é grande. Esta penalidade é feita por adicionar o fator multiplicado pela dimensão do modelo e levando em conta o valor de n . Os dois critérios são:

- Critério de informação de Akaike - *AIC*

$$AIC(\lambda) = Dev(\mathbf{y}; \lambda) + 2de(\lambda),$$

- Critério de informação Bayesiano - *BIC*

$$BIC(\lambda) = Dev(\mathbf{y}; \lambda) + \ln(n)de(\lambda).$$

Na Figura 2.9, mostra-se as curvas estimadas usando diferentes métodos de suavização utilizando *splines* e pode-se observar que ajustando o modelo aos dados com os *splines* de regressão, resulta em uma curva pouco suave que está quase interpolando os dados, por não utilizar uma penalização. A curva que melhor ajusta-se aos dados e é mais suave, é a que utiliza os P-*splines* com 35 nós para as bases B-*splines* e com $\lambda = 4$. O tipo de penalidade dependerá do tipo de base a ser utilizada (EILERS e MARX, 1996).

2.7 P-Splines como modelos mistos

Os P-*splines* estudados nos últimos anos surgem da possibilidade de escrever um modelo não paramétrico ou semiparamétrico, em que se utilizam os P-*splines* como um modelo misto (DURBÁN, 2014). A vantagem de utilizar essa abordagem permite, por um lado utilizar a metodologia desenvolvida para os modelos mistos, e por outro lado, utilizar o *software* para modelos mistos que está disponível na maioria dos programas estatísticos. Essa ligação entre regressão não paramétrica e os modelos mistos, iniciou-se na década de 90 com os trabalhos de: SPEED (1991), WANG (1998b), BRUMBACK e RICE (1998) e VERBYLA *et al.* (1999), entre outros; estes desenvolveram a teoria com maiores detalhes, ou seja, a ligação entre suavização por *splines* e os modelos mistos utilizando como bases os *splines* cúbicos. WANG (1998a) considerou o modelo misto com erros correlacionados e WAND (2003), utilizou os polinômios

Figura 2.9: Curvas estimadas usando diferentes métodos de suavização: *Splines* de regressão, suavização por *splines* e *P-splines*, utilizando neste último 35 nós para as bases de *B-splines* com $\lambda = 4$, para a função $f(x) = \cos(5\pi x)$ para um conjunto de 200 dados simulados.

truncados como bases de regressão no modelo misto. Neste trabalho, aborda-se a representação dos *P-splines* como modelos mistos conforme CURRIE e DURBAN (2002), CURRIE *et al.* (2006) e LEE e DURBAN (2011). Essa escolha foi realizada, pelo fato de que as bases *B-splines* são bases mais estáveis nos processos computacionais.

2.7.1 Revisão geral de modelos mistos

Os modelos mistos (*Mixed Models* - MM) ou modelos de efeitos mistos são uma extensão dos modelos de regressão nos quais incorpora-se efeitos aleatórios. De acordo com PINHEIRO e BATES (2000) “os modelos de efeitos mistos fornecem uma ferramenta flexível e poderosa para a análise de dados agrupados”. Esses dados agrupados surgem em muitas áreas tais como agricultura, biologia, economia, etc. Matricialmente o modelo misto descrito em HARVILLE (1977) é expresso por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad (2.54)$$

em que $\mathbf{y}$, é o vetor $(n \times 1)$ de variáveis resposta; $\boldsymbol{\beta}$, o vetor de parâmetros desconhecidos de efeito fixo de ordem $q + 1 \times 1$, também conhecido como de “efeito marginal”; $\mathbf{X}$, é a matriz de incidência de efeitos fixos conhecido de ordem $n \times q + 1$; $\boldsymbol{\alpha}$, é o vetor aleatório $k \times 1$ cujas componentes são chamados de efeitos aleatórios desconhecidos; $\mathbf{Z}$ é a matriz de incidência dos efeitos aleatórios conhecidos de dimensão $n \times k$ e $\boldsymbol{\epsilon}$ é o vetor de erros aleatórios $(n \times 1)$; $\mathbf{X}$ e $\mathbf{Z}$ são matrizes de posto completo. Em que n , é o número de observações, q é o número de parâmetros e k é o número de efeitos aleatórios. Com as seguintes pressuposições:

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}) \quad \text{e} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}),$$

$\boldsymbol{\alpha}$ e $\boldsymbol{\epsilon}$ são independentes; $\mathbf{G}$ e $\mathbf{R}$, são matrizes positivas definidas por hipótese, e portanto, não singular, logo tem-se

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right),$$

em que, $\mathbf{G} = \sigma_{\alpha}^2 \mathbf{I}_k$, é a matriz de variâncias-covariâncias dos efeitos aleatórios presentes no vetor $\boldsymbol{\alpha}$; $\mathbf{R} = \sigma_{\epsilon}^2 \mathbf{I}_n$, é a matriz de variâncias-covariâncias dos efeitos fixos (ou residuais). Tirar conclusões sobre o comportamento dos modelos de efeitos mistos, refere-se: à estimação dos parâmetros de efeitos fixos, à predição de efeitos aleatórios e à estimação dos componentes da variância-covariância.

2.7.2 Estimação de efeitos fixos

Do modelo (2.54) e sob a pressuposição de normalidade dos erros que são independentes e identicamente distribuídos (*iid*), a distribuição marginal de $\mathbf{y}$ é normal com média, $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, e variância,

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}, \quad (2.55)$$

em que os componentes de variância, σ_{α}^2 e σ_{ϵ}^2 , estão incluídos através da matriz de variâncias-covariâncias de (2.55). Portanto,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$$

com função de densidade marginal de $\mathbf{y}$, definida por

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mathbf{V}^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \right\}, \quad (2.56)$$

e função de verossimilhança de (2.56) é dada por

$$L(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_{\alpha}^2, \sigma_{\epsilon}^2) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\mathbf{V}) - \frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]. \quad (2.57)$$

Derivando o logaritmo da máxima verossimilhança de (2.57) em relação a $\boldsymbol{\beta}$, obtém-se as estimativas dos efeitos fixos

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}. \quad (2.58)$$

Às vezes essa última expressão refere-se mínimos quadrados generalizados.

2.7.3 Predição de efeitos aleatórios

Adotando-se um modelo misto pode-se fazer a predição de efeitos aleatórios, na presença de efeitos fixos, pelos BLUPs. Como descrito por SEARLE *et al.* (2009), se $\mathbf{y}$ é o vetor de dados observados e $\boldsymbol{\alpha}$ é o vetor aleatório, então o melhor preditor (*Best Prediction - BP*) corresponde por minimizar o critério

$$E \{ \| \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \|^2 \},$$

cujas solução é dada por

$$\tilde{\boldsymbol{\alpha}} = BP(\boldsymbol{\alpha}) = E(\boldsymbol{\alpha} | \mathbf{y}), \quad (2.59)$$

e o melhor preditor linear (*Best Linear Prediction - BLP*), está definido por

$$\tilde{\boldsymbol{\alpha}} = BLP(\boldsymbol{\alpha}) = \mathbf{G}\mathbf{Z}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (2.60)$$

Em modelos mistos, tais predições são dadas pelo melhor preditor linear não viesado (*Best Linear Unbiased Prediction - BLUP*), supondo a distribuição condicional de $\mathbf{y}$ dado $\boldsymbol{\alpha}$,

$$\mathbf{y} | \boldsymbol{\alpha} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \mathbf{R})$$

com função de densidade, $f(\mathbf{y}|\boldsymbol{\alpha})$, dada por

$$f(\mathbf{y}|\boldsymbol{\alpha}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})] \right\}, \quad (2.61)$$

e como, $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G})$, a função de densidade, $f(\boldsymbol{\alpha})$, é definida por

$$f(\boldsymbol{\alpha}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{G}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [\boldsymbol{\alpha}' \mathbf{G}^{-1} \boldsymbol{\alpha}] \right\}. \quad (2.62)$$

Da definição de função de densidade conjunta, $f(\mathbf{y}, \boldsymbol{\alpha}) = f(\mathbf{y}|\boldsymbol{\alpha}) \cdot f(\boldsymbol{\alpha})$, e das expressões (2.61) e (2.62), obtém-se

$$f(\mathbf{y}, \boldsymbol{\alpha}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{R}|^{\frac{1}{2}} |\mathbf{G}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}) + \boldsymbol{\alpha}' \mathbf{G}^{-1} \boldsymbol{\alpha}] \right\}, \quad (2.63)$$

com função de verossimilhança,

$$l(\mathbf{y}, \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ n \log(2\pi) + \log(\mathbf{R}) + \log(\mathbf{G}) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}) + \boldsymbol{\alpha}' \mathbf{G}^{-1} \boldsymbol{\alpha} \right\}, \quad (2.64)$$

o que mostra que a estimação $BLUP(\boldsymbol{\beta} \text{ e } \boldsymbol{\alpha})$, envolve mínimos quadrados generalizados com um termo de penalização, $\boldsymbol{\alpha}' \mathbf{G}^{-1} \boldsymbol{\alpha}$. Uma forma de unificar os resultados é por meio do BLUP (*Best Linear Unbiased Prediction*). Conforme mostra ROBINSON (1991), as soluções para $\tilde{\boldsymbol{\beta}}$ e $\tilde{\boldsymbol{\alpha}}$, são

$$\begin{aligned} BLUP(\boldsymbol{\beta}) &= \tilde{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \\ \text{e } BLUP(\boldsymbol{\alpha}) &= \tilde{\boldsymbol{\alpha}} = \mathbf{GZ}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \end{aligned}$$

Nota-se que o $BLUP$ para $\boldsymbol{\beta}$ é idêntico à solução de mínimos quadrados generalizados expressos em (2.58), o $BLUP$ para $\boldsymbol{\alpha}$ é o BLP com $\boldsymbol{\beta}$ substituído por $BLUP(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}}$.

Portanto o $BLUP(\boldsymbol{\beta}, \boldsymbol{\alpha})$, pode ser expresso por

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = [\mathbf{C}' \mathbf{R}^{-1} \mathbf{C} + \mathbf{P}]^{-1} \mathbf{C}' \mathbf{R}^{-1} \mathbf{y},$$

em que $\mathbf{C} = [\mathbf{X} \quad \mathbf{Z}]$ e $\mathbf{P} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$.

Outra forma de encontrar o $BLUP$ de $\boldsymbol{\alpha}$, é utilizando a distribuição conjunta de $\mathbf{y}$ e $\boldsymbol{\alpha}$, expresso por

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{\alpha} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V} & \mathbf{Z}\mathbf{G} \\ \mathbf{GZ}' & \mathbf{G} \end{bmatrix} \right).$$

Com o resultado das esperanças condicionais da parte multivariada $\mu_{2.1} = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1)$, obtém-se a predição dos efeitos aleatórios a partir de $\boldsymbol{\alpha}|\mathbf{y}$,

$$\hat{\boldsymbol{\alpha}} = \mathbf{GZ}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (2.65)$$

Os valores ajustados para $\mathbf{y}$, são dados por

$$BLUP(\mathbf{y}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\alpha}} = \mathbf{C}[\mathbf{C}' \mathbf{R}^{-1} \mathbf{C} + \mathbf{P}]^{-1} \mathbf{C}' \mathbf{R}^{-1} \mathbf{y}. \quad (2.66)$$

2.7.4 Estimação de componentes de variância

Da expressão (2.57), podemos encontrar os estimadores por máxima verossimilhança de σ_α^2 e σ_ϵ^2 , maximizando o perfil log-verossimilhança para V , definido por

$$l_p(\mathbf{V}) = -\frac{1}{2} [\log(\mathbf{V}) + \mathbf{y}' \mathbf{V}^{-1} \{ \mathbf{I} - \mathbf{X}(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \} \mathbf{y}] - \frac{n}{2} \log(2\pi) \quad (2.67)$$

Vale lembrar que as estimativas por máxima verossimilhança são viesadas, pois não consideram os graus de liberdade usados para a estimação dos efeitos fixos. Um método de estimação alternativo que leva em conta essa perda de graus de liberdade é a “Estimação de máxima verossimilhança restrita” (*restricted maximum likelihood estimation* - REML), citado por PATTERSON e THOMPSON (1971) e SCHALL (1991). Portanto, a função log-verossimilhança restrita é dado por

$$l_{REML}(\mathbf{V}) = l_p(\mathbf{V}) - \frac{1}{2} \log(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}) \quad (2.68)$$

que derivada estima os componentes de variância

$$\hat{\mathbf{V}}^{-1} = \frac{1}{\sigma_\epsilon^2} [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \sigma_\epsilon^2\mathbf{G}^{-1})^{-1}\mathbf{Z}'] . \quad (2.69)$$

2.7.5 Representação dos P-splines como modelos mistos

De acordo com DURBÁN (2014) o interesse pela representação dos P-splines como modelos mistos, surgiu das dificuldades que os modelos aditivos apresentam com funções suaves, pois existem problemas de identificabilidade nos modelos. Outra vantagem de utilizar P-splines e não tem necessidade de métodos iterativos para a estimação das curvas, mas é necessário modificar as bases, de forma que uma curva possa ser decomposta como uma soma de um componente polinomial, da mesma ordem que a penalidade, e outro não polinomial. BRUMBACK e RICE (1998) apresentam os P-splines com funções de base de polinômios truncados e utiliza uma aproximação de modelos mistos para estimar o parâmetro de suavização. VERBYLA *et al.* (1999), faz a decomposição dessa curva utilizando suavização por splines tendo como base os splines cúbicos, os quais são expressos como a soma de um efeito fixo linear, desde que a penalidade seja de ordem 2 para uma suavização por splines, e um efeito aleatório. Na abordagem de LEE e DURBAN (2011), a função suave $f(\mathbf{x}) = \mathbf{B}\mathbf{a}$, é expressa como a soma de efeitos fixos e aleatórios.

Seja o modelo de regressão,

$$\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \quad (2.70)$$

em que $\mathbf{B}$, é a base da regressão e $\mathbf{a}$, os coeficientes de regressão da base $\mathbf{B}$. A proposta de LEE e DURBAN (2011) é reparametrizar os componentes em uma nova base, $\mathbf{B} \rightarrow [\mathbf{X} : \mathbf{Z}]$. Esta representação decompõe os valores ajustados como a soma da parte polinomial ($\mathbf{X}\boldsymbol{\beta}$) não penalizada e o termo de suavização não linear penalizado ($\mathbf{Z}\boldsymbol{\alpha}$) e, com essa nova reparametrização, encontrar o critério de otimização expresso por mínimos quadrados penalizados,

$$\mathbf{S} = (\mathbf{y} - \mathbf{B}\mathbf{a})'(\mathbf{y} - \mathbf{B}\mathbf{a}) + \lambda\mathbf{a}'\mathbf{P}\mathbf{a}. \quad (2.71)$$

Reformulando o modelo (2.70) como modelo misto,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}) \quad \text{e} \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}) \quad (2.72)$$

com $\mathbf{X}$ e $\mathbf{Z}$ matrizes de incidência dos efeitos fixos e aleatórios respectivamente; $\boldsymbol{\beta}$, vetor de parâmetros desconhecidos dos efeitos fixos; $\boldsymbol{\alpha}$, vetor de parâmetros dos efeitos aleatórios e $\mathbf{G} = \sigma_\alpha^2\boldsymbol{\Lambda}$, a matriz de variâncias-covariâncias dos efeitos aleatórios, $\boldsymbol{\alpha}$, para alguma matriz positiva definida, $\boldsymbol{\Lambda}$, e $\mathbf{R} = \sigma_\epsilon^2\mathbf{I}$, a matriz de variâncias-covariâncias dos erros.

A aproximação de CURRIE e DURBAN (2002) e CURRIE *et al.* (2006), usa os B-splines e a penalidade de diferenças $\mathbf{P}$, para reparametrizar o modelo original em um modelo misto, através dos seguintes teoremas, lemas e corolários:

Lema 2.7.1 A matriz de transformação $\mathbf{T}$ para reparametrizar o modelo base e coeficientes de um modelo P-splines em uma representação do modelo misto, é dado por

$$\mathbf{T} = [\mathbf{T}_0 : \mathbf{T}_+] \quad (2.73)$$

em que, $\mathbf{T}$ é uma matriz ortogonal, com sub matrizes $\mathbf{T}_0$ e $\mathbf{T}_+$, que contêm respectivamente os autovetores da parte nula e não nula da decomposição em valores singulares da matriz de penalidade $\mathbf{P} = \mathbf{D}'\mathbf{D}$, com $\mathbf{D} = \mathbf{\Delta}^d$.

Demonstração de ortogonalidade de $\mathbf{T}$

Seja $\mathbf{P} = \mathbf{D}'\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}'$, a decomposição em valores singulares (*DVS*) da matriz de penalidade. A matriz de autovetores $\mathbf{U}$, que pode ser dividida em duas partes:

$$\mathbf{U} = [\mathbf{U}_0 : \mathbf{U}_+] \quad (2.74)$$

em que $\mathbf{U}_0$ contém a parte nula de autovetores de dimensão $(k+1+p) \times d$, e $\mathbf{U}_+$ contém a parte não nula de autovetores da decomposição de dimensão $(k+1+p) \times (k+1+p-d)$. A matriz diagonal de $\mathbf{\Sigma}$ contém os autovalores da *DVS* de $\mathbf{P} = \mathbf{D}'\mathbf{D}$ com d autovetores nulos. Então, pode-se decompor a penalidade como

$$\mathbf{D}'\mathbf{D} = \begin{bmatrix} \mathbf{U}_0 & \mathbf{U}_+ \end{bmatrix} \begin{bmatrix} \mathbf{0}_d & \\ & \tilde{\mathbf{\Sigma}}_{c-d} \end{bmatrix} \begin{bmatrix} \mathbf{U}'_0 \\ \mathbf{U}'_+ \end{bmatrix},$$

em que $\mathbf{0}_d$, é uma matriz quadrada de zeros de ordem d e $\tilde{\mathbf{\Sigma}}_{c-d}$ é a matriz de autovalores de dimensão, $(c-d)$, em que $c = k+1+p$, é o número de colunas da base *B-splines* e d , é a ordem da penalidade.

Considerando $\mathbf{T} = [\mathbf{U}_0 : \mathbf{U}_+]$, a matriz $\mathbf{T}$ é ortogonal, isto é, $\mathbf{T}\mathbf{T}' = \mathbf{I}$, assim

$$\mathbf{T}\mathbf{T}' = \mathbf{U}_0\mathbf{U}'_0 + \mathbf{U}_+\mathbf{U}'_+ = \mathbf{I}_c.$$

Lema 2.7.2 Dada a matriz de transformação $\mathbf{T}$ em (2.73), tal que $\mathbf{B}\mathbf{T} = [\mathbf{X} : \mathbf{Z}]$. As matrizes de efeitos fixos e aleatórios são dados por:

$$\mathbf{X} = \mathbf{B}\mathbf{U}_0 \quad \text{e} \quad (2.75)$$

$$\mathbf{Z} = \mathbf{B}\mathbf{U}_+ \quad (2.76)$$

com coeficientes

$$\begin{aligned} \beta &= \mathbf{U}'_0 \mathbf{a} \\ \alpha &= \mathbf{U}'_+ \mathbf{a} \end{aligned} \quad .$$

Demonstração

Define-se uma função suave, $\mathbf{B}\mathbf{a} = \mathbf{X}\beta + \mathbf{Z}\alpha$, em que $[\mathbf{X} : \mathbf{Z}]$ tem posto completo. Fazendo uso da transformação injetora de $\mathbf{T} = [\mathbf{U}_0 : \mathbf{U}_+]$ da matriz ortogonal $\mathbf{T}$, tal que $\mathbf{B} \rightarrow [\mathbf{X} : \mathbf{Z}]$, conhecida como Fatoração QR e utilizando as rotações de Householder, temos que

$$\begin{aligned} \mathbf{B}\mathbf{T} &= [\mathbf{X} : \mathbf{Z}], \\ \mathbf{B}\mathbf{T}\mathbf{T}'\mathbf{a} &= [\mathbf{X} : \mathbf{Z}]\mathbf{T}'\mathbf{a}, \end{aligned}$$

como $\mathbf{T}' = [\mathbf{U}'_0 : \mathbf{U}'_+]$ e fazendo comparação de termos,

$$\mathbf{T}'\mathbf{a} = \begin{bmatrix} \mathbf{U}'_0 \mathbf{a} \\ \mathbf{U}'_+ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \beta \\ \alpha \end{bmatrix},$$

portanto, temos a nova reparametrização

$$\mathbf{B}\mathbf{T}\mathbf{T}'\mathbf{a} = [\mathbf{X} : \mathbf{Z}] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \mathbf{X}\beta + \mathbf{Z}\alpha.$$

Teorema 2.7.1 Dada a matriz de transformação ortogonal em (2.73), e a matriz de penalidade $\mathbf{P} = \mathbf{D}'\mathbf{D}$. O bloco diagonal da penalidade do modelo misto é dado por

$$\Phi = \mathbf{T}'\mathbf{P}\mathbf{T} = \text{blockdiag}(\mathbf{0}_d, \mathbf{F}) \quad \text{com} \quad \mathbf{F} = \lambda\tilde{\Sigma}, \quad (2.77)$$

em que $\mathbf{0}_d$, é a matriz quadrada de zeros de ordem igual ao número de efeitos fixos d , e $\lambda\tilde{\Sigma}$, é a matriz diagonal com elementos na diagonal iguais aos autovalores não nulos da DVS de $\mathbf{P}$.

Demonstração

Como $\mathbf{T} = [\mathbf{U}_0 : \mathbf{U}_+]$ é ortogonal e $\mathbf{P} = \mathbf{D}'\mathbf{D}$ do termo da penalidade $\lambda\mathbf{a}'\mathbf{P}\mathbf{a}$ da expressão (2.71), a reparametrização da nova penalidade é dada por

$$\lambda\mathbf{a}'\mathbf{P}\mathbf{a} = \lambda\mathbf{a}'\mathbf{T}'\mathbf{P}\mathbf{T}\mathbf{a}.$$

Dessa expressão, considerando somente o termo, $\lambda\mathbf{T}'\mathbf{P}\mathbf{T}$ como

$$\begin{aligned} \lambda\mathbf{T}'\mathbf{D}'\mathbf{D}\mathbf{T} &= \lambda \begin{bmatrix} \mathbf{U}'_0 \\ \mathbf{U}'_+ \end{bmatrix} \mathbf{D}'\mathbf{D}[\mathbf{U}_0 : \mathbf{U}_+] \\ &= \lambda \begin{bmatrix} \mathbf{U}'_0\mathbf{D}'\mathbf{D}\mathbf{U}_0 & \\ & \mathbf{U}'_+\mathbf{D}'\mathbf{D}\mathbf{U}_+ \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0}_d & \\ & \mathbf{F} = \lambda\tilde{\Sigma} \end{bmatrix}, \end{aligned} \quad (2.78)$$

obtém-se o resultado de (2.77), que é a nova penalidade do modelo misto.

Corolário 2.7.1 Dado $\mathbf{T}'\mathbf{P}\mathbf{T} = \text{blockdiag}(\mathbf{0}_d, \mathbf{F})$ com $\mathbf{F} = \lambda\tilde{\Sigma}$, os efeitos fixos β são não penalizados, e somente os efeitos aleatórios α são penalizados pela matriz diagonal $\mathbf{F}$, que contém os autovalores positivos da DVS de $\mathbf{D}'\mathbf{D}$. Então, dada a matriz de transformação $\mathbf{T} = [\mathbf{T}_0 : \mathbf{T}_+]$, a penalidade do modelo misto pode ser obtida como:

$$\mathbf{F} = \mathbf{T}'_+\mathbf{P}\mathbf{T}_+, \quad (2.79)$$

em que $\mathbf{T}_+$ é a matriz de autovetores correspondente dos autovalores não nulos.

Demonstração

Para o caso unidimensional, fazendo $\mathbf{T}_+ = \mathbf{U}_+$, e sabe-se que $\mathbf{F} = \lambda\tilde{\Sigma}$, então tem-se que

$$\mathbf{F} = \mathbf{U}'_+\mathbf{D}'\mathbf{D}\mathbf{U}_+ = \lambda\mathbf{T}'_+\mathbf{P}\mathbf{T}_+.$$

Corolário 2.7.2 A matriz da penalidade $\mathbf{P}$, pode ser construída a partir da nova penalidade $\mathbf{F}$, dada por

$$\mathbf{P} = \mathbf{T}'_+\mathbf{F}\mathbf{T}_+. \quad (2.80)$$

Demonstração

Do Teorema (2.7.1), sabe-se que $\Phi = \mathbf{T}'_+\mathbf{P}\mathbf{T}_+$, então

$$\mathbf{T}\Phi\mathbf{T}' = \mathbf{T}\mathbf{T}'\mathbf{P}\mathbf{T}\mathbf{T}' = \mathbf{P},$$

portanto,

$$\mathbf{P} = [\mathbf{T}_0 : \mathbf{T}_+] \begin{bmatrix} \mathbf{0}_d & \\ & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{T}'_0 \\ \mathbf{T}'_+ \end{bmatrix} = \mathbf{T}_+\mathbf{F}\mathbf{T}'_+.$$

Sendo que os parâmetros fixos β' s são não penalizados, a matriz de efeitos fixos $\mathbf{X} = \mathbf{B}\mathbf{U}_0$, pode ser substituída por qualquer sub matriz, tal que:

- A matriz composta $[\mathbf{X} : \mathbf{Z}]$ tem posto completo, isso implica que individualmente essas matrizes $\mathbf{X}$ e $\mathbf{Z}$ também são de posto completo na coluna.
- $\mathbf{X}$ e $\mathbf{Z}$ são ortogonais; isto é, $\mathbf{X}'\mathbf{Z} = \mathbf{0}$.

Supondo uma penalidade de segunda ordem, $d = 2$, a matriz diagonal de autovalores, tem dois zeros e $(c - 2)$ autovetores positivos, logo, a matriz de efeitos fixos pode ser escrita como:

$$\mathbf{X} = [\mathbf{1} : \mathbf{x}],$$

em que, $\mathbf{1}$ é o vetor de uns; e $\mathbf{x}$ é o vetor de covariáveis. Então, quando $\lambda \rightarrow \infty$, o modelo nulo é a parte polinomial $\mathbf{X}\boldsymbol{\beta}$, de grau $(p - 1)$, e a parte aleatória $\mathbf{Z}\boldsymbol{\alpha}$, é considerada como o desvio em relação ao modelo nulo.

Após a reparametrização do modelo misto em termos da nova base $\mathbf{B}$ para a expressão (2.71), reparametrizando a parte da penalidade em termos da nova penalidade $\mathbf{F}$,

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}) + \mathbf{a}'\mathbf{P}\mathbf{a},$$

no caso unidimensional, utilizando o Corolário (2.7.2), fazendo alguns cálculos no vetor de coeficientes $\mathbf{a}$ e considerando a parte dos efeitos aleatórios, a nova matriz diagonal da penalidade do modelo misto é $\mathbf{F} = \lambda\tilde{\boldsymbol{\Sigma}}$. Portanto, dada as novas bases e a nova penalidade, a soma dos quadrados penalizados de (2.71) resulta em

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}) + \boldsymbol{\alpha}'\mathbf{F}\boldsymbol{\alpha}. \quad (2.81)$$

Obtendo as derivadas da expressão (2.81) em relação aos parâmetros, facilmente obtém-se as estimativas do modelo para $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ respectivamente, ou seja

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\boldsymbol{\alpha}} &= \hat{\sigma}_{\alpha}^2\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad . \end{aligned}$$

Assim, utilizando a expressão (2.69), a estimação de $\mathbf{V}$ é dada por

$$\hat{\mathbf{V}}^{-1} = \frac{1}{\sigma_{\epsilon}^2} \left[\mathbf{I} - \mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma_{\epsilon}^2}{\sigma_{\alpha}^2} \mathbf{I}_{c-2} \right)^{-1} \mathbf{Z}' \right]. \quad (2.82)$$

Da mesma maneira, é possível obter a matriz de suavização e os graus de liberdade para *P-splines* como modelo misto, expressos em (2.44) e (2.45) respectivamente,

$$\begin{aligned} \mathbf{H}_{\lambda} &= \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{P}_d)^{-1}\mathbf{B}' \\ \mathbf{H}_{\lambda} &= [\mathbf{X} : \mathbf{Z}] \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} [\mathbf{X} : \mathbf{Z}]' \quad , \end{aligned} \quad (2.83)$$

e com a reparametrização, a matriz de componentes de variância-covariância é definida como $\mathbf{G} = \sigma_{\alpha}^2\mathbf{F}^{-1}$ e os graus de liberdade são dadas por

$$de(\lambda) = \text{traço}(\mathbf{H}_{\lambda}) = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix}. \quad (2.84)$$

Na Figura 2.10 pode-se observar o comportamento das diferentes curvas usando a representação de *P-splines* como modelos mistos, com números diferentes de nós.

Figura 2.10: Curvas estimadas usando diferentes valores de nós para as bases de B-splines, usando representação P-splines como modelos mistos para a função $f(x) = \cos(5\pi x)$ para um conjunto de 200 dados simulados.

2.7.6 Intervalos de confiança para a função suave

Observa-se que a curva ajustada é dada por, $\hat{f}(\mathbf{x}) = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha} = \mathbf{H}_\lambda \mathbf{y}$, e o interesse dessa curva estimada, é levar em consideração a variabilidade considerando homocedasticidade, $Cov(\mathbf{y}) = \sigma_\epsilon^2 \mathbf{I}$, então a variância de $\hat{f}(\mathbf{x})$ é

$$Var(\hat{f}(\mathbf{x})) = Var(\mathbf{H}_\lambda \mathbf{y}), \quad (2.85)$$

com bandas de variabilidade definidas como

$$\hat{f}(\mathbf{x}) \pm 2 \times sd(\hat{f}(\mathbf{x})), \quad (2.86)$$

em que sd é o desvio padrão. Segundo RUPPERT *et al.* (2003), o ajuste dever ser feito em relação ao desvio padrão de $\hat{f}(\mathbf{x})$, porque considera,

- A variabilidade em $\hat{\sigma}_\epsilon$ como uma estimativa de σ_ϵ ,
- o viés devido à curvatura, e
- a variabilidade em $\mathbf{H}_\lambda$, devido à estimação do parâmetro de suavização.

Pode-se notar também que $\hat{f}(\mathbf{x})$ é aleatória devido à aleatoriedade em α , então de acordo com RUPPERT *et al.* (2003) “ α é considerada como uma simples ferramenta para modelar a curvatura, enquanto ϵ explica a variabilidade ao redor da curva”. Portanto, a variabilidade de $\hat{f}(\mathbf{x})$ deve ser em relação à distribuição condicional $\mathbf{y}|\alpha$, com variância, $Var(\mathbf{y}|\alpha) = \sigma_\epsilon^2 \mathbf{I}$; e a variância da condicional da curva ajustada, está expressa por

$$Var(\hat{f}(\mathbf{x})|\alpha) = Var(\mathbf{H}_\lambda \mathbf{y}|\alpha) = \hat{\sigma}_\epsilon^2 \mathbf{B} Cov \left(\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} \middle| \alpha \right) \mathbf{B}', \quad (2.87)$$

mas, a matriz de covariância condicional é dada por

$$\text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} \middle| \alpha \right) = \hat{\sigma}_\epsilon^2 \left(\mathbf{B}'\mathbf{B} + \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\alpha^2} \mathbf{P}_d \right)^{-1} \mathbf{B}'\mathbf{B} \left(\mathbf{B}'\mathbf{B} + \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\alpha^2} \mathbf{P}_d \right)^{-1},$$

consequentemente da expressão (2.87), obtém-se o desvio padrão da condicional

$$sd(\hat{f}(\mathbf{x})|\alpha) = \hat{\sigma}_\epsilon \sqrt{\mathbf{B} \left(\mathbf{B}'\mathbf{B} + \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\alpha^2} \mathbf{P}_d \right)^{-1} \mathbf{B}'\mathbf{B} \left(\mathbf{B}'\mathbf{B} + \frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\alpha^2} \mathbf{P}_d \right)^{-1} \mathbf{B}'} \quad (2.88)$$

obtendo como resultado intervalo de confiança aproximado $100(1 - \alpha)\%$,

$$\hat{f}(\mathbf{x}) \pm z_{(1-\frac{\alpha}{2})} \times sd(\hat{f}(\mathbf{x})|\alpha). \quad (2.89)$$

A estimativa do parâmetro de suavização λ é definida como a razão das estimativas das variâncias $\frac{\hat{\sigma}_\epsilon^2}{\hat{\sigma}_\alpha^2}$, definidas na expressão (2.88), essas estimativas são obtidas utilizando REML.

Na Figura 2.11 mostra-se o intervalo de confiança de 95% para uma curva ajustada utilizando a representação *P-splines* como modelo misto. Pode-se observar também que esses intervalos de confiança são maiores pois levam em consideração os dois componentes do erro, ou seja, a variância e o viés ao quadrado.

Figura 2.11: Intervalo de confiança para as bases de *B-splines*, utilizando *P-splines* como modelos mistos para a função $f(x) = \cos(5\pi x)$ com 200 dados simulados.

Há casos em que a matriz de variâncias-covariâncias dos efeitos fixos ou residual, $\mathbf{R} = \sigma_\epsilon^2 \mathbf{I}_n$, não apresenta uma estrutura de correlação nos resíduos, no entanto quando ajusta-se uma curva suave com *P-splines* como representação de modelos mistos, imediatamente deve-se de introduzir uma estrutura de correlação e assim estimar juntamente com a curva suave conforme DURBAN (2002).

Seja o modelo expresso em (2.72), sob a pressuposição que os resíduos são distribuídos de forma independente com média 0 e variância σ_ϵ^2 , conforme ZUUR *et al.* (2009), tem-se

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

$$Cov(\epsilon_i, \epsilon_{i'}) = \begin{cases} \sigma_\epsilon^2, & \text{se } i = i' \\ 0, & \text{caso contrário} \end{cases}, \quad (2.90)$$

A segunda linha da expressão (2.90) é devido a pressuposição de independência, ao invés de usar 0, modela-se a autocorrelação entre os resíduos em diferentes pontos temporais introduzindo uma função de correlação, $h(\cdot)$:

$$Cor(\epsilon_i, \epsilon_{i'}) = \begin{cases} 1, & \text{se } i = i' \\ h(\epsilon_i, \epsilon_{i'}), & \text{caso contrário} \end{cases}, \quad (2.91)$$

A função $h(\cdot)$ é conhecida como função de correlação e toma valores entre -1 e 1 . Assumindo estacionariedade, ou seja, a correlação entre os resíduos ϵ_i e $\epsilon_{i'}$ somente depende da diferença entre os pontos temporais i e i' conforme PINHEIRO e BATES (2000). Uma possível estrutura de correlação para os resíduos é o processo autoregressivo de ordem 1, $AR(1)$. Segundo ZUUR *et al.* (2009), essa estrutura modela os resíduos no tempo $i - 1$ mais um pequeno erro e_i :

$$\epsilon_i = \rho\epsilon_{i-1} + e_i,$$

o parâmetro ρ é desconhecido e necessita ser estimado das observações. Logo, a estrutura de correlação é dado por

$$Cor(\epsilon_i, \epsilon_{i'}) = \begin{cases} 1, & \text{se } i = i' \\ \rho^{|i'-i|}, & \text{caso contrário} \end{cases}. \quad (2.92)$$

2.8 Suavizado bidimensional

Para a construção dos suavizadores ou curvas em duas dimensões, deve-se obter uma base para uma regressão bidimensional com suporte local da mesma forma em que os *B-splines* proporcionam uma base estável para uma regressão unidimensional (CURRIE *et al.*, 2004). Para entender o processo de construção de uma regressão de duas dimensões, supondo como o exemplo, um simples modelo polinomial de duas dimensões com duas variáveis regressoras x_1 e x_2 , com preditor linear expresso por

$$1 + x_1 + x_2 + x_1^2 + x_1x_2 + x_1x_1^2 \quad (2.93)$$

e as observações da variável resposta estão dispostas em uma matriz $\mathbf{Y}$, de tamanho $n_1 \times n_2$. As linhas de $\mathbf{Y}$ são indexadas por x_1 e as colunas por x_2 . Para os efeitos de regressão, supõe-se que os dados são dispostos em colunas, isto é, $\mathbf{y} = \text{vec}(\mathbf{Y})$.

A fatorização do preditor linear de (2.93), é expresso como $(1 + x_1 + x_1^2)(1 + x_2)$ e pode-se observar que tais fatores podem ser representados como dois modelos, um em função de x_1 para a coluna de $\mathbf{Y}$, e outro em função de x_2 para a linha de $\mathbf{Y}$. Logo, pode-se denotar $\mathbf{X}_1$, como a matriz da regressão do modelo com preditor linear, $(1 + x_1 + x_1^2)$, e $\mathbf{X}_2$ a matriz de regressão que corresponde ao preditor linear, $(1 + x_2)$. Utilizando a definição de produto tensor (RAO e KLEFFE, 1988), essa fatorização é dado por

$$\mathbf{X} = \mathbf{X}_2 \otimes \mathbf{X}_1 \quad (2.94)$$

em que $\otimes$ denota o produto Kronecker de duas matrizes. Portanto, na expressão (2.94) a matriz $\mathbf{X}$, é a matriz do modelo polinomial com preditor linear dado por (2.93), expressa como o produto de dois modelos marginais dados em termos de $\mathbf{X}_1$ e $\mathbf{X}_2$.

2.9 P-splines bidimensional

A metodologia dos *P-splines* para o caso unidimensional pode ser estendido para o caso bidimensional, e requer a extensão das funções de bases *B-splines* com estrutura bidimensional, o que significa, utilizar o produto tensor para as bases *B-splines*.

2.9.1 Produto tensor de bases

No item (2.6.4), mostrou-se em detalhe a suavização do *P-splines* com uma simples variável explicativa utilizando uma base *B-spline*. Neste item trata-se da suavização usando *P-splines* no caso bidimensional utilizando duas bases marginais *B-splines* para cada covariável. (DURBAN, 2002), propôs o uso do produto tensor de bases *B-splines* como bases para o suavizado de duas ou mais dimensões.

Seguindo a abordagem de DURBAN (2002), considere-se um modelo de regressão geral não-paramétrico k -dimensional,

$$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k) + \epsilon, \quad \epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}) \quad (2.95)$$

em que $f(\cdot)$ é a função suave de k variáveis explicativas, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$, de tamanho $n_1, n_2, \dots, n_k$, respectivamente, cuja função suave é expressa em termos das bases *B-splines* por

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k) = \mathbf{B}\mathbf{a}, \quad (2.96)$$

em que $\mathbf{B}$ é definida como o produto tensor das k bases *B-splines* correspondentes as k variáveis explicativas

$$\mathbf{B} = \mathbf{B}_k \otimes \dots \otimes \mathbf{B}_2 \otimes \mathbf{B}_1 \quad (2.97)$$

e o sinal $\otimes$ representa o produto Kronecker de duas matrizes e $\mathbf{B}_l = \mathbf{B}(\mathbf{x}_l)$, para $l = 1, 2, \dots, k$.

No caso bidimensional, o uso do produto tensor das bases *B-splines* para duas variáveis explicativas ($k = 2$), com função suave expressa por, $f(\mathbf{x}_1, \mathbf{x}_2)$, e variáveis explicativas definidas, $\mathbf{x}_1 = (x_{11}, x_{12}, \dots, x_{1n_1})'$ e $\mathbf{x}_2 = (x_{21}, x_{22}, \dots, x_{2n_2})'$, o vetor de observações $\mathbf{y}$, é organizado em uma matriz $\mathbf{Y}$ de tamanho n_1 linhas e n_2 colunas, assim $\mathbf{y} = \text{vec}(\mathbf{Y})$, é dado por

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1n_2} \\ y_{21} & y_{22} & \dots & y_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n_1 1} & y_{n_1 2} & \dots & y_{n_1 n_2} \end{bmatrix},$$

com y_{ij} , em que $i = 1, 2, \dots, n_1$ e $j = 1, 2, \dots, n_2$. Como o objetivo é buscar uma “superfície suave” tal que, $f(\mathbf{x}_1, \mathbf{x}_2)$ dê uma boa aproximação aos dados, então a base da regressão do suavizado é dada pela matriz de base em duas dimensões, definida como o produto tensor

$$\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1, \quad (2.98)$$

em que a base *B-splines* $\mathbf{B}_1$, de tamanho $n_1 \times c_1$ suaviza ao longo de $\mathbf{x}_1$, e a base *B-splines* $\mathbf{B}_2$ de tamanho $n_2 \times c_2$, suaviza ao longo de $\mathbf{x}_2$, portanto a dimensão da matriz $\mathbf{B}$ expressa em (2.98) é de tamanho $n_1 n_2 \times c_1 c_2$.

2.9.2 Produto tensor das penalizações

Pode-se observar que na expressão (2.96), a matriz $\mathbf{B}$, associa-se ao vetor de coeficientes da regressão $\mathbf{a}$; portanto, o preditor linear $\mathbf{B}\mathbf{a}$, é definido por,

$$\mathbf{B}\mathbf{a} = \mathbf{B}_1 \mathbf{A} \mathbf{B}_2, \quad (2.99)$$

e os elementos de $\mathbf{a}$ estão organizados em linhas e colunas na matriz $\mathbf{A}$ de tamanho $c_1 \times c_2$, tal que $\mathbf{a} = \text{vec}(\mathbf{A})$, é dado por

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1c_2} \\ a_{21} & a_{22} & \dots & a_{2c_2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{c_1 1} & a_{c_1 2} & \dots & a_{c_1 c_2} \end{bmatrix}.$$

Seguindo o proposto por EILERS e MARX (2003), o objetivo é penalizar as linhas e colunas da matriz $\mathbf{A}$; o que significa que o preditor linear da j -ésima coluna de $\mathbf{Y}$, pode ser escrito como combinação linear de c_2 suavizados em $\mathbf{x}_1$, ou seja, aplicar uma penalização à falta de suavidade em cada uma das colunas c_2 de $\mathbf{A}$,

$$\sum_{j=1}^{c_2} a'_{.j} \mathbf{D}'_1 \mathbf{D}_1 a_{.j} = \mathbf{a}' (\mathbf{I}_{c_2} \otimes \mathbf{D}'_1 \mathbf{D}_1) \mathbf{a} = \mathbf{a}' \mathbf{P}_1 \mathbf{a}, \quad (2.100)$$

De forma similar, considerando o preditor linear que corresponde a i -ésima linha de $\mathbf{Y}$, a correspondente penalidade sobre as linhas c_1 de $\mathbf{A}$ suavizados em $\mathbf{x}_2$, dada por

$$\sum_{i=1}^{c_1} a'_{i.} \mathbf{D}'_2 \mathbf{D}_2 a_{i.} = \mathbf{a}' (\mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_1}) \mathbf{a} = \mathbf{a}' \mathbf{P}_2 \mathbf{a}, \quad (2.101)$$

em que $\mathbf{D}_1$ e $\mathbf{D}_2$ são as matrizes de diferença de ordem d_1 e d_2 respectivamente. Finalmente a matriz de penalização em duas dimensões é dado por

$$\mathbf{P} = \lambda_1 (\mathbf{I}_{c_2} \otimes \mathbf{D}'_1 \mathbf{D}_1) + \lambda_2 (\mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_1}) = \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2, \quad (2.102)$$

em que λ_1 e λ_2 são os parâmetros de suavização para cada dimensão do modelo, e de forma geral os valores desses parâmetros são diferentes, o que faz com que o modelo seja “não isotrópico”; ou seja, que exista a possibilidade do suavizado ser maior em uma dimensão do que em outra (DURBAN, 2002). $\mathbf{P}_1$ e $\mathbf{P}_2$ são matrizes semidefinidas positivas de coeficientes.

2.10 P-splines bidimensionais como modelos mistos

Seguindo a representação de P-splines como modelo misto no caso unidimensional, como expresso no modelo (2.72); estende-se para o caso bidimensional como modelo misto (LEE e DURBAN, 2011). Considerando a função suave

$$f(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}) \quad , \quad (2.103)$$

e representando esse modelo como P-splines, precisa-se da reparametrização das bases e coeficientes do modelo em termos de P-splines, tal como

$$\mathbf{B} \rightarrow [\mathbf{X} : \mathbf{Z}] \quad \text{e} \quad \mathbf{a} \rightarrow (\boldsymbol{\beta}, \boldsymbol{\alpha}) \quad , \quad (2.104)$$

com matriz de penalização $\mathbf{P}$ definida na expressão (2.102). A soma $\mathbf{P}_1 + \mathbf{P}_2$, é conhecida como soma Kronecker, portanto a decomposição em valor singular dessa soma de penalidades permite a diagonalização simultânea de $\mathbf{P}_1$ e $\mathbf{P}_2$; isto é, $\mathbf{P}_1 + \mathbf{P}_2 = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}'$, (HORN e JOHNSON, 1991). O uso dessa decomposição em valor singular, permite particionar essa soma em duas partes, uma penalidade nula que contém os efeitos fixos do modelo misto e uma penalidade diagonal, não nula, que contém os efeitos aleatórios do modelo. As penalidades individuais, $\mathbf{P}_1 = \mathbf{I}_{c_2} \otimes \mathbf{D}'_1 \mathbf{D}_1$, e $\mathbf{P}_2 = \mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_1}$, estão em função da diagonalização individual definidas pela decomposição em valor singular, $\mathbf{D}'_1 \mathbf{D}_1 = \mathbf{U}_1 \boldsymbol{\Sigma}_1 \mathbf{U}'_1$ e $\mathbf{D}'_2 \mathbf{D}_2 = \mathbf{U}_2 \boldsymbol{\Sigma}_2 \mathbf{U}'_2$, respectivamente, com matrizes de autovetores expressas por, $\mathbf{U}_1 = [\mathbf{U}_0^1 : \mathbf{U}_+^1]$ e $\mathbf{U}_2 = [\mathbf{U}_0^2 : \mathbf{U}_+^2]$, definidas em (2.74) e com matrizes de autovalores $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$, respectivamente.

Seguindo a abordagem de LEE e DURBAN (2011), com os mesmos procedimentos do caso unidimensional, precisa-se encontrar uma matriz de transformação, $\mathbf{T}$, para calcular a nova base e coeficientes do modelo misto no caso de duas dimensões.

Lema 2.10.1 A matriz de transformação $\mathbf{T}$ para reparametrizar a base e coeficientes do modelo como representação de P-splines como modelo misto no caso de duas dimensões, é dado por

$$\mathbf{T} = [\mathbf{T}_0 : \mathbf{T}_+] = [\mathbf{U}_0 : \mathbf{U}_+] \quad , \quad (2.105)$$

em que, $\mathbf{T}$ é uma matriz ortogonal de tamanho $n_1 n_2 \times c_1 c_2$, com sub matrizes: $\mathbf{U}_0 = [\mathbf{U}_0^2 \otimes \mathbf{U}_0^1]$, que representa o bloco da parte fixa, e $\mathbf{U}_+ = [\mathbf{U}_+^2 \otimes \mathbf{U}_0^1 : \mathbf{U}_0^2 \otimes \mathbf{U}_+^1 : \mathbf{U}_+^2 \otimes \mathbf{U}_+^1]$, que são os blocos da parte aleatória.

Demonstração de ortogonalidade de $\mathbf{T}$

Sejam as matrizes marginais: $\mathbf{T}_1 = [\mathbf{U}_0^1 : \mathbf{U}_+^1]$, com dimensão, $\dim(\mathbf{T}_1) = n_1 \times c_1$, e $\mathbf{T}_2 = [\mathbf{U}_0^2 : \mathbf{U}_+^2]$, de dimensão, $\dim(\mathbf{T}_2) = n_2 \times c_2$, aplicando produto Kronecker a $\mathbf{T}_1$ e $\mathbf{T}_2$ em função das submatrizes $\mathbf{U}_0$ e $\mathbf{U}_+$, obtém-se

$$\begin{aligned} \mathbf{T} &= \mathbf{T}_2 \otimes \mathbf{T}_1 \\ &= (\mathbf{U}_0^2 \otimes [\mathbf{U}_0^1 : \mathbf{U}_+^1] : \mathbf{U}_+^2 \otimes [\mathbf{U}_0^1 : \mathbf{U}_+^1]) \quad . \end{aligned} \quad (2.106)$$

Para demonstrar que $\mathbf{T}$ é ortogonal, utiliza-se a definição $\mathbf{T}\mathbf{T}'$, para obter a seguinte expressão

$$\begin{aligned} \mathbf{T}\mathbf{T}' &= (\mathbf{U}_0^2 \mathbf{U}_0^{2'} + \mathbf{U}_+^2 \mathbf{U}_+^{2'}) \otimes (\mathbf{U}_0^1 \mathbf{U}_0^{1'} + \mathbf{U}_+^1 \mathbf{U}_+^{1'}) \\ &= \mathbf{I}_{c_1} \otimes \mathbf{I}_{c_2} \\ &= \mathbf{I}_{c_1 c_2} \quad . \end{aligned}$$

e portanto, $\mathbf{T}$ é ortogonal.

Lema 2.10.2 Seja a matriz de transformação $\mathbf{T}$, como definida na expressão (2.105) do Lema (2.10.1); tal que, $\mathbf{B}\mathbf{T} = [\mathbf{X} : \mathbf{Z}]$, em que $\mathbf{X}$ e $\mathbf{Z}$, são as matrizes de efeitos fixos e aleatórios respectivamente do modelo misto, expressas por

$$\mathbf{X} = \mathbf{X}_2 \otimes \mathbf{X}_1 \quad \text{e} \quad \mathbf{Z} = \mathbf{Z}_2 \otimes \mathbf{X}_1 : \mathbf{X}_2 \otimes \mathbf{Z}_1 : \mathbf{Z}_2 \otimes \mathbf{Z}_1,$$

Demonstração

Utilizando o Lema (2.7.2) do caso univariado, com matrizes definidas

$$\mathbf{X} = \mathbf{B}\mathbf{U}_0 \quad \text{e} \quad \mathbf{Z} = \mathbf{B}\mathbf{U}_+,$$

e sendo a matriz $\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1$, tal que a matriz de efeitos fixos é dada por

$$\begin{aligned} \mathbf{X} &= \mathbf{B}\mathbf{U}_0 \\ &= (\mathbf{B}_2 \otimes \mathbf{B}_1)(\mathbf{U}_0^2 \otimes \mathbf{U}_0^1) \\ &= \mathbf{X}_2 \otimes \mathbf{X}_1. \end{aligned} \quad (2.107)$$

A parte aleatória é definida por

$$\begin{aligned} \mathbf{Z} &= \mathbf{B}\mathbf{U}_+ \\ &= (\mathbf{B}_2 \otimes \mathbf{B}_1)(\mathbf{U}_+^2 \otimes \mathbf{U}_0^1 : \mathbf{U}_0^2 \otimes \mathbf{U}_+^1 : \mathbf{U}_+^2 \otimes \mathbf{U}_+^1) \\ &= (\mathbf{Z}_2 \otimes \mathbf{X}_1 : \mathbf{X}_2 \otimes \mathbf{Z}_1 : \mathbf{Z}_2 \otimes \mathbf{Z}_1). \end{aligned} \quad (2.108)$$

Como $\mathbf{B}\mathbf{T} = [\mathbf{X} : \mathbf{Z}]$ e a matriz de transformação $\mathbf{T}$ é ortogonal, obtém-se

$$\begin{aligned} \mathbf{B}\mathbf{T}\mathbf{T}' &= [\mathbf{X} : \mathbf{Z}]\mathbf{T}' \\ \mathbf{B}\mathbf{T}\mathbf{T}'\mathbf{a} &= [\mathbf{X} : \mathbf{Z}]\mathbf{T}'\mathbf{a} \quad , \end{aligned}$$

onde o termo $\mathbf{T}'\mathbf{a}$ é expressa por

$$\mathbf{T}'\mathbf{a} = \begin{bmatrix} \mathbf{U}_0'\mathbf{a} \\ \mathbf{U}_+'\mathbf{a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} \quad ,$$

logo, os novos coeficientes são:

$$\beta = U_0' a \quad \text{e} \quad \alpha = U_+' a.$$

Portanto, obtém-se a nova reparametrização do modelo

$$BTT'a = [X : Z] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = X\beta + Z\alpha = Ba. \quad (2.109)$$

Teorema 2.10.1 Seja a matriz de transformação ortogonal T como definida na expressão (2.105) do Lema (2.10.1) e seja a matriz de penalidade P , expressa como a soma Kronecker: $P_1 + P_2$. Utilizando a definição do Corolário (2.7.2), a matriz bloco diagonal da penalidade do modelo misto é $F = T_+'PT_+$, para o caso de duas dimensões é dada por,

$$F = \begin{bmatrix} \lambda_2 \tilde{\Sigma}_2 \otimes I_{q_1} & & \\ & \lambda_1 I_{q_2} \otimes \tilde{\Sigma}_1 & \\ & & \lambda_1 I_{c_2-q_2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes I_{c_1-q_1} \end{bmatrix},$$

em que F , é o bloco diagonal da matriz de penalidade sobre os efeitos aleatórios dos coeficientes de α .

Demonstração

Seja a matriz de transformação ortogonal $T = [U_0 : U_+] = [U_0^2 \otimes U_0^1 : U_+^2 \otimes U_0^1 : U_0^2 \otimes U_+^1 : U_+^2 \otimes U_+^1]$, como definida no Lema (2.10.1), e utilizando a definição do Teorema 2.7.1, tem-se

$$\Phi = T'PT = \text{blockdiag}(\mathbf{0}_{q_1 q_2}, F), \quad (2.110)$$

em que a matriz de penalização P , é definida como na expressão em 2.102 é considerada como soma Kronecker de P_1 e P_2 e utilizando (2.110), obtém-se

$$\begin{aligned} \Phi &= T'PT \\ &= T'[\lambda_1(I_{c_2} \otimes D_1' D_1) + \lambda_2(D_2' D_2 \otimes I_{c_1})]T, \end{aligned} \quad (2.111)$$

logo

$$\begin{aligned} \Phi &= \lambda_1 \begin{bmatrix} U_0^{2'} \otimes U_0^{1'} \\ U_+^{2'} \otimes U_0^{1'} \\ U_0^{2'} \otimes U_+^{1'} \\ U_+^{2'} \otimes U_+^{1'} \end{bmatrix} (I_{c_2} \otimes D_1' D_1) [U_0^2 \otimes U_0^1 : U_+^2 \otimes U_0^1 : U_0^2 \otimes U_+^1 : U_+^2 \otimes U_+^1] + \\ &\quad \lambda_2 \begin{bmatrix} U_0^{2'} \otimes U_0^{1'} \\ U_+^{2'} \otimes U_0^{1'} \\ U_0^{2'} \otimes U_+^{1'} \\ U_+^{2'} \otimes U_+^{1'} \end{bmatrix} (D_2' D_2 \otimes I_{c_1}) [U_0^2 \otimes U_0^1 : U_+^2 \otimes U_0^1 : U_0^2 \otimes U_+^1 : U_+^2 \otimes U_+^1] \end{aligned}$$

$$\begin{aligned} \Phi &= \lambda_1 \begin{bmatrix} I_{c_2} U_0^{2'} U_0^2 \otimes U_0^{1'} D_1' D_1 U_0^1 & & & \\ & U_+^{2'} I_{c_2} U_+^2 \otimes U_0^{1'} D_1' D_1 U_0^1 & & \\ & & U_0^{2'} I_{c_2} U_0^2 \otimes U_+^{1'} D_1' D_1 U_+^1 & \\ & & & U_+^{2'} I_{c_2} U_+^2 \otimes U_+^{1'} D_1' D_1 U_+^1 \end{bmatrix} + \\ &\quad \lambda_2 \begin{bmatrix} U_0^{2'} D_2' D_2 U_0^2 \otimes U_0^{1'} I_{c_1} U_0^1 & & & \\ & U_+^{2'} D_2' D_2 U_+^2 \otimes U_0^{1'} I_{c_1} U_0^1 & & \\ & & U_0^{2'} D_2' D_2 U_0^2 \otimes U_+^{1'} I_{c_1} U_+^1 & \\ & & & U_+^{2'} D_2' D_2 U_+^2 \otimes U_+^{1'} I_{c_1} U_+^1 \end{bmatrix}, \end{aligned}$$

dado que

$$\begin{aligned} U_0^{1'} U_0^1 &= I_{q_1} & \text{e} & & U_+^{1'} U_+^1 &= I_{c_1-q_1} & , \\ U_0^{2'} U_0^2 &= I_{q_2} & \text{e} & & U_+^{2'} U_+^2 &= I_{c_2-q_2} & , \\ U_0^{2'} D_2' D_2 U_0^2 &= \mathbf{0}_{q_2} & \text{e} & & U_0^{1'} D_1' D_1 U_0^1 &= \mathbf{0}_{q_1} & , \\ U_+^{2'} D_2' D_2 U_+^2 &= \tilde{\Sigma}_2 & \text{e} & & U_+^{1'} D_1' D_1 U_+^1 &= \tilde{\Sigma}_1 & , \end{aligned}$$

portanto,

$$\Phi = \lambda_1 \begin{bmatrix} I_{q_2} \otimes \mathbf{0}_{q_1} & & & \\ & I_{c_2-q_2} \otimes \mathbf{0}_{q_1} & & \\ & & I_{c_2-q_2} \otimes \tilde{\Sigma}_1 & \\ & & & I_{c_2-q_2} \otimes \tilde{\Sigma}_1 \end{bmatrix} +$$

$$\lambda_2 \begin{bmatrix} \mathbf{0}_{q_2} \otimes I_{q_1} & & & \\ & \tilde{\Sigma}_2 \otimes I_{q_1} & & \\ & & \mathbf{0}_{q_2} \otimes I_{c_1-q_1} & \\ & & & \tilde{\Sigma}_2 \otimes I_{c_1-q_1} \end{bmatrix}$$

$$\Phi = \left[\begin{array}{c|c} \mathbf{0}_{q_1 q_2} & \\ \hline & \lambda_2 \tilde{\Sigma}_2 \otimes I_{q_2} \\ & \lambda_1 I_{q_2} \otimes \tilde{\Sigma}_1 \\ & \lambda_1 I_{c_2-q_2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes I_{c_1-q_1} \end{array} \right],$$

logo, a matriz Φ é definida como o bloco diagonal

$$\Phi = \text{blockdiag}(\mathbf{0}_{q_1 q_2}, \mathbf{F}),$$

em que a submatriz $\mathbf{F}$ está expressa por

$$\mathbf{F} = \begin{bmatrix} \lambda_2 \tilde{\Sigma}_2 \otimes I_{q_1} & & \\ & \lambda_1 I_{q_2} \otimes \tilde{\Sigma}_1 & \\ & & \lambda_1 I_{c_2-q_2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes I_{c_1-q_1} \end{bmatrix}.$$

Portanto, dadas as novas bases e a nova penalidade, a soma dos quadrados penalizados resulta em

$$S = (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\alpha)'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\alpha) + \alpha' \mathbf{F} \alpha \quad . \quad (2.112)$$

2.11 Modelo Misto Aditivo Generalizado

Uma alternativa ao uso de modelos de regressão linear, é usar os modelos aditivos, entre eles temos o Modelo Misto Aditivo Generalizado o que será tratado no presente trabalho. LIN e ZHANG (1999a), propuseram o Modelo Misto Aditivo Generalizado (*Generalized Additive Mixed Models* - GAMM), como uma classe de modelos que são uma extensão de modelos mistos lineares generalizados e do modelo aditivo generalizado (*Generalized Additive Models* - GAM) proposto por HASTIE e TIBSHIRANI (1990). Essa classe de modelos usa funções não-paramétricas aditivas para modelar efeito das covariáveis sobre a variável resposta assim como, também leva em consideração o efeito da presença de superdispersão e correlação, acrescentando efeito aleatório ao preditor aditivo do modelo. Os GAMM podem ser utilizados em estudos com delineamentos experimentais sejam eles aninhados ou cruzados, assim como aplicados a dados espaciais, dados agrupados ou cluster e hierárquicos. Para definir um GAMM, precisa-se de algumas definições básicas, que serão discutidas a seguir.

2.11.1 Modelo linear generalizado

O modelo linear generalizado (*Generalized Linear Models* - GLM) foi proposto por McCULLAGH e NELDER (1989). Um GLM consiste de um componente aleatório, um componente sistemático e uma função de ligação que relaciona os dois componentes.

1. Componente aleatório: Expresso pela variável resposta $\mathbf{y}$ do modelo, cuja média é $\mu = E(\mathbf{y})$, relacionado ao conjunto de covariáveis $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ por meio do componente sistemático,

2. Componente sistemático: As variáveis explanatórias entram na forma de estrutura linear do modelo, chamado preditor linear, definido por $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \mathbf{x}_i'\boldsymbol{\beta}$ ou por $\boldsymbol{\eta}_i = \sum_{j=1}^p x_{ij}\beta_j$, com matriz do modelo $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)'$, em que $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$; vetor de parâmetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$, e vetor de preditor linear $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_p)'$,
3. Função de ligação: denotada por $g(\cdot)$, é uma função monótona e derivável, o que permite a relação entre a média $\boldsymbol{\mu}$ e o preditor linear, $\boldsymbol{\eta} = g(\boldsymbol{\mu})$.

Com base nesses três componentes, e sendo o vetor da variável resposta $\mathbf{y} = (y_1, y_2, \dots, y_n)'$, procedente de uma distribuição que pertence à família exponencial definida por

$$f(\mathbf{y}; \boldsymbol{\theta}, \phi) = \exp \left\{ \frac{\mathbf{y}\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(\mathbf{y}; \phi) \right\} \quad , \quad (2.113)$$

em que $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$, são funções específicas; o parâmetro $\boldsymbol{\theta}$, é chamado parâmetro canônico de locação e ϕ parâmetro de dispersão. A média e a variância de $\mathbf{y}$ são $E(\mathbf{y}) = \boldsymbol{\mu} = b'(\boldsymbol{\theta})$ e variância $Var(\mathbf{y}) = a(\phi)V(\boldsymbol{\mu})$, em que $V(\boldsymbol{\mu}) = b''(\boldsymbol{\theta})$, é a função de variância que depende da média $\boldsymbol{\mu}$ por meio do parâmetro canônico $\boldsymbol{\theta}$.

2.11.2 Modelos mistos lineares generalizados

Os modelos mistos lineares generalizados (*Generalized Linear Mixed Model* - GLMM) propostos por BRESLOW e CLAYTON (1993), são uma função dos GLM e dos modelos mistos, adaptando as propriedades dessas duas abordagens. Segundo DEAN e NIELSEN (2007), os GLMM incorporam uma variedade de efeitos aleatórios como é tratado nos modelos mistos lineares, esses modelos com efeitos aleatórios ajudam a controlar a heterogeneidade não observada, por exemplo a heterogeneidade é constante no tempo e correlacionadas com as variáveis explanatórias.

Uma característica de GLMM é que eles utilizam a função da média para modelar os efeitos da covariável, levando em consideração superdispersão e correlação acrescentando os efeitos aleatórios ao preditor linear de um GLM.

O preditor linear de um GLMM é definido por

$$\boldsymbol{\eta}|\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} = \sum_{j=1}^p x_{ij}\beta_j + \sum_{k=1}^q z_{ik}\alpha_k \quad , \quad (2.114)$$

em que $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_q)'$, é o vetor de efeitos aleatórios com $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G})$, a matriz de variâncias é $\mathbf{G} = \sigma_\alpha^2 \mathbf{I}_\alpha$, e a matriz de efeitos aleatórios do modelo $\mathbf{Z} = (z_{1k}, z_{2k}, \dots, z_{ik})'$.

Observa-se pela expressão (2.114) que o GLMM para famílias exponenciais é basicamente um modelo hierárquico, com a seguinte função,

$$f(\mathbf{y}|\boldsymbol{\eta}) = \exp \{ \mathbf{y}'\boldsymbol{\eta} - \mathbf{1}'b(\boldsymbol{\eta}) + \mathbf{1}'c(\mathbf{y}) \} \quad , \quad (2.115)$$

mas o preditor linear de (2.114) é substituído em (2.115), escrevendo essa última expressão

$$f(\mathbf{y}|\boldsymbol{\alpha}) = \exp \{ \mathbf{y}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}) - \mathbf{1}'b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}) + \mathbf{1}'\log(\Gamma(\mathbf{y} + 1)) \} \quad , \quad (2.116)$$

em que $f(\mathbf{y}|\boldsymbol{\alpha})$ pertence a uma distribuição da família exponencial. Portanto, as respostas observadas y_i são condicionalmente independentes dado o vetor q -dimensional $\boldsymbol{\alpha}$ de efeitos aleatórios com média condicional $E[y_i|\boldsymbol{\alpha}] = \mu_i^\alpha$ e variância condicional $Var(y_i | \boldsymbol{\alpha}) = \phi(m_i)^{-1}V(\mu_i^\alpha)$, nota-se que $V(\cdot)$ é a função de variância da média condicional, m_i é um peso *priori* conhecido e ϕ é o parâmetro de dispersão como no caso do modelo linear generalizado. Observa-se que a média condicional está ligada ao preditor linear, dado por $\eta_i^\alpha = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\alpha}$, por meio da função de ligação $g(\mu_i^\alpha) = \eta_i^\alpha$.

2.11.3 Modelo misto aditivo generalizado

LIN e ZHANG (1999a), introduziram os Modelos Mistos Aditivos Generalizados (*Generalized Additive Mixed Models* - GAMM); no entanto CHEN (2007) desenvolveu o processo de estimação dos parâmetros utilizando dois algoritmos, sendo o primeiro a aproximação por Newton-Raphson e o segundo uma extensão do algoritmo Esperança-Maximização (EM) por Monte Carlo, onde foi utilizada a log-verossimilhança marginal penalizada. WOOD (2006) e ZUUR *et al.* (2009) utilizaram os GAMM cujas funções suaves estão em termos do produto tensor utilizando como bases de baixo rango os *thin plate splines*.

Conforme a proposta e notação de LIN e ZHANG (1999a), os GAMM distingue-se do GLMM, em que o preditor linear é substituído por um preditor aditivo no componente sistemático do modelo. O preditor aditivo envolve uma soma de funções suaves das variáveis explanatórias definidas no componente sistemático, reescrevendo a expressão (2.114), tem-se

$$\boldsymbol{\eta}|\boldsymbol{\alpha} = \sum_{j=1}^p f_j(x_{ij}) + \sum_{k=1}^q z_{ik}\alpha_k \quad . \quad (2.117)$$

Seguindo a mesma metodologia dos modelos lineares generalizados, tem-se

- Componente aleatório:

$$f(\mathbf{y}|\boldsymbol{\eta}) = \exp \{ \mathbf{y}'\boldsymbol{\eta} - \mathbf{1}'b(\boldsymbol{\eta}) + \mathbf{1}'c(\mathbf{y}) \} \quad ,$$

- Componente sistemático:

$$\boldsymbol{\eta}|\boldsymbol{\alpha} = \boldsymbol{\eta}_i^\alpha = \sum_{j=1}^p f_j(x_{ij}) + \sum_{k=1}^q z_{ik}\alpha_k \quad ,$$

- Função de ligação:

$$\eta_i^\alpha = g(\mu_i^\alpha) \quad ,$$

- Condições: $f''(\cdot)$ é contínua e $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$.

Como no caso do GLMM, a função $g(\cdot)$ é uma função de ligação monótona diferenciável, $f_j(\cdot)$, é a função suave, $\boldsymbol{\alpha}$ vetor de efeitos aleatórios, supondo que são distribuídos normalmente com variância $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$, e $\boldsymbol{\theta}$ é o vetor de componentes de variância de ordem $c \times 1$, (LIN e ZHANG, 1999a).

O objetivo principal do GAMM expresso em (2.117), é que as funções suaves aditivas são usadas para modelar os efeitos das covariáveis e os efeitos aleatórios são usados para modelar correlação entre observações, por possuir uma estrutura de covariância flexível dos efeitos aleatórios $\boldsymbol{\alpha}$.

Entre algumas referências onde foram utilizados modelos como casos especiais do GAMM, tem-se ZEGER e DIGGLE (1994) e ZHANG *et al.* (1998) que consideram a expressão (2.117), como modelo semiparamétrico misto. Os autores assumiram uma simples função não-paramétrica em função do tempo e das observações longitudinais, consideradas distribuídas normalmente. ZUUR *et al.* (2009), utiliza o GAMM em dados espaciais e estudos ecológicos para modelar a correlação espacial.

Considerando o seguinte modelo

$$y_i = f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + \dots + \mathbf{Z}_i\boldsymbol{\alpha} + \epsilon_i, \quad (2.118)$$

em que y_i é a variável resposta; os $f_j(\cdot)$ são as funções suaves das covariáveis x_{ki} ; $\mathbf{Z}_i$ é a linha da matriz do modelo de efeitos aleatórios; $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G})$, é o vetor de coeficientes de efeitos aleatórios com matriz de covariância definida positiva desconhecida $\mathbf{G}$, o vetor dos resíduos com i -ésimo elemento ϵ_i dado por $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, com matriz de variância-covariância $\boldsymbol{\Lambda}$, a qual tem algum tipo de estrutura de correlação.

2.11.4 Suavizado dos componentes com representação P-spline

WOOD (2006) utiliza para suavizar as funções suaves a suavização por *splines* utilizando como base os *thin plate splines*. Usando a abordagem de WOOD (2006) para o suavizado em duas dimensões em um GAMM, explica-se como uma função suave pode ser reformulada em componentes convencionais de um modelo misto linear como visto na seção (2.6.4). Primeiramente é considerado, o suavizado com um simples parâmetro de suavização; em seguida, considerando o suavizado pelo produto tensor no caso bidimensional utilizando P-splines.

1. Considerando uma função unidimensional: $f(x) = \mathbf{B}\mathbf{a}$, conforme visto em (2.7.5), utiliza-se toda a teoria da representação P-spline como modelo misto; ou seja, a reparametrização é necessária na função suave em duas partes: a primeira parte para ser tratada como efeito fixo e a outra para ser tratada como efeito aleatório como visto no Lema (2.7.2). Em geral, isso pode ser obtido usando a decomposição em valores singulares da matriz de penalidade, $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}'$, em que $\mathbf{U}$ é a matriz ortogonal como definida em (2.74), cujas colunas são os autovalores de $\mathbf{P}$, e $\mathbf{\Sigma}$ é uma matriz diagonal com os autovalores correspondentes dispostos em forma decrescente na diagonal principal, onde $\tilde{\mathbf{\Sigma}}_{c-d}$, denota a sub matriz de autovalores positivos de $\mathbf{\Sigma}$, em que $c = k + 1 + p$, é o número de colunas da base B-splines e d , é a ordem da penalidade. Reparametrizando a nova matriz e o novo vetor de coeficientes tem-se

$$\mathbf{X} = \mathbf{B}\mathbf{U}_0 \quad \text{e} \quad (2.119)$$

$$\mathbf{Z} = \mathbf{B}\mathbf{U}_+ \quad (2.120)$$

com coeficientes

$$\boldsymbol{\beta} = \mathbf{U}_0' \mathbf{a} \quad \text{e}$$

$$\boldsymbol{\alpha} = \mathbf{U}_+' \mathbf{a} \quad .$$

Utilizando o Corolário (2.7.2), obtém-se a nova penalidade $\mathbf{a}'\mathbf{P}\mathbf{a} = \boldsymbol{\alpha}'\mathbf{F}\boldsymbol{\alpha}$. Logo como representação em modelo misto de uma função suave em termos de um preditor linear e uma distribuição de efeitos aleatórios, está dado por

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}, \quad \text{em que} \quad \boldsymbol{\alpha} \sim N\left(\mathbf{0}, \frac{\tilde{\mathbf{\Sigma}}^{-1}}{\lambda}\right), \quad (2.121)$$

sendo $\boldsymbol{\beta}$ e λ parâmetros fixos a serem estimados.

2. Considerando uma função bidimensional $f(x_1, x_2)$: Nesta abordagem utiliza-se todo o referencial apresentado em (2.10), que consiste em ter uma função suave de duas dimensões e buscar expressar em termos de uma representação P-splines como modelos mistos conforme visto em (2.103) e reparametrizando como dado na Equação (2.104); objetivo é buscar o suavizado de cada dimensão expresso em termos do produto tensor com dois parâmetros de suavização em cada dimensão a ser suavizada. Novamente, reparametrizando a expressão (2.103) é necessário expressar a soma de duas penalizações $\mathbf{P}_1 + \mathbf{P}_2$ em termos da decomposição em valor singular, $\mathbf{P}_1 + \mathbf{P}_2 = \mathbf{U}\mathbf{\Sigma}\mathbf{U}'$, o que permite particionar essa soma em duas partes, uma penalidade nula que contém os efeitos fixos do modelo misto e uma penalidade diagonal, não nula, que contém os efeitos aleatórios do modelo. A matriz $\mathbf{U}$ é a matriz ortogonal de autovetores e $\mathbf{\Sigma}$ é a matriz diagonal dos autovalores dessa soma de penalizações. Pode-se observar que não existe nenhum parâmetro de suavização nessa soma a ser decomposta, então utilizando o Teorema (2.10.1) consegue-se a submatriz $\mathbf{F}$, que é o bloco diagonal da matriz de penalidade sobre os efeitos aleatórios dos coeficientes de $\boldsymbol{\alpha}$ e é expresso por

$$\mathbf{F} = \begin{bmatrix} \lambda_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_{q_1} & & \\ & \lambda_1 \mathbf{I}_{q_2} \otimes \tilde{\Sigma}_1 & \\ & & \lambda_1 \mathbf{I}_{c_2-q_2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_{c_1-q_1} \end{bmatrix},$$

em que $\tilde{\Sigma}_1$ e $\tilde{\Sigma}_2$, denotam a submatriz de autovalores positivos de Σ que corresponde a parte de efeitos aleatórios do modelo.

Portanto, a representação como modelo misto dos termos suavizados utilizando produto tensor, é dada pelo preditor linear e uma distribuição de efeitos aleatórios

$$\mathbf{X}\beta + \mathbf{Z}\alpha, \quad \text{em que } \alpha \sim N(\mathbf{0}, \mathbf{F}^{-1}) \quad . \quad (2.122)$$

em que β , λ_1 e λ_2 são parâmetros fixos a serem estimados.

3 MATERIAL E MÉTODOS

Os dados utilizados correspondem às produções de cana-de-açúcar em toneladas/*ha*, de cada mesorregião do estado de São Paulo - Brasil, com o objetivo de aplicar toda a metodologia de *P-splines* descrita no capítulo anterior.

A agricultura é a atividade econômica mais dependente das condições climáticas (MONTEIRO, 2009); isto significa que a variabilidade natural das chuvas, temperatura e outras condições meteorológicas afetam não só o processo metabólico das plantas, bem como as produções de diferentes culturas, portanto, as condições climáticas são consideradas como fatores que explicam a variabilidade das produções agrícolas.

No decorrer dos anos, vários estudos verificaram que as variáveis climáticas, tais como os níveis de temperatura e precipitação, são fundamentais para explicar o desenvolvimento e níveis da produção agrícola, o padrão do uso do solo e até mesmo os níveis das lavouras (DESCHÊNES e GREENSTONE, 2012).

No Brasil, a produção de cana-de-açúcar teve início no período colonial e é uma das principais culturas agrícolas na economia brasileira, concentrando-se essa produção nas regiões Nordeste e Centro-Sul do Brasil, tendo maior destaque no estado de São Paulo (UNICA, 2017).

O presente estudo de aplicação, teve como objetivo estudar os níveis médios da produção de cana-de-açúcar de cada mesorregião do estado de São Paulo, sob a influência das alterações nos níveis de temperatura e precipitação, com a finalidade de mostrar e avaliar o uso da metodologia *P-splines* nas variáveis climáticas sobre o desempenho da produção da cana-de-açúcar.

3.1 Materiais

- Fonte de dados e variáveis utilizadas: Os dados utilizados neste trabalho, foram coletados de várias fontes estatísticas disponíveis. As informações referentes a coleta de dados de produção de cana-de-açúcar em toneladas/*ha*, foram obtidas do anuário de Produção Agrícola Municipal (PAM) do Instituto Brasileiro de Geografia Estatística (IBGE, 2017), dados que correspondem por mesorregião do estado de São Paulo desde o ano 2006 até 2015. As informações relacionadas aos dados meteorológicos como precipitação, temperatura máxima, temperatura média e temperatura mínima, foram obtidas das estações meteorológicas do Instituto Nacional de Meteorologia (INMET, 2017) localizadas no estado de São Paulo, do Centro Integrado de Informações Agrometeorológicas (CIIAGRO, 2017), assim como do posto meteorológico “Professor Dr. Jesus Marden dos Santos” da Escola Superior de Agricultura “Luiz de Queiroz- Universidade de São Paulo (ESALQ, 2017).

Segundo o PAM as mesorregiões onde há produção de cana-de-açúcar e que fazem parte do estado de São Paulo são: Araçatuba, Araraquara, Assis, Bauru, Campinas, Itapetininga, Litoral Sul Paulista, Marco Metropolitana Paulista, Marília, Piracicaba, Presidente Prudente, Ribeirão Preto, São José do Rio Preto e Vale do Paraíba Paulista. Na Figura 3.1 mostra-se o mapa de zoneamento agroambiental para o setor sucroalcooleiro aptidão edafoclimática para a produção de cana-de-açúcar.

3.2 Métodos

- Relação entre a produção e as variáveis climáticas: Buscou-se relacionar a produção de cana-de-açúcar de cada mesorregião com os níveis médios de temperatura ($^{\circ}C$) e precipitação (*mm*), registrados anualmente desde o período 2006 a 2015. O foco principal é explorar o comportamento dos efeitos anuais da mudança de temperatura e precipitação e como influenciam na produção

Figura 3.1: Mapa edafoclimático da produção de cana-de-açúcar do Estado de São Paulo.

Fonte: Instituto Agrônomo de Campinas - Governo do Estado de São Paulo.

de cana-de-açúcar em cada mesorregião do estado de São Paulo, utilizando e avaliando com a metodologia proposta do uso de *P-splines* em cada uma das variáveis climáticas na variável ano.

Como tem-se observações múltiplas de cada uma das mesorregiões, as observações da mesma mesorregião são mais similares que as observações de diferentes mesorregiões, no entanto, não há interesse particular no efeito das mesorregiões, mas essa variável será considerada como fator de agrupamento já que permitirá uma possível correlação entre as observações de uma mesorregião individual. O modelo geral descrito a considerar-se é expressado por

$$\begin{aligned} \text{Produção} = & \text{Efeito Ano} + \text{Efeito Temperatura máxima} + \text{Efeito Temperatura mínima} + \\ & \text{Efeito Precipitação} + \text{Erro}, \end{aligned} \quad (3.1)$$

utilizando notação estatística, a expressão de (3.1) é dada por

$$\begin{aligned} \text{Produção}_{ij} = & \beta_0 + \beta_1 \times \text{Ano}_{ij} + \beta_2 \times \text{Temperatura máxima}_{ij} + \beta_3 \times \text{Temperatura mínima}_{ij} + \\ & \beta_4 \times \text{Precipitação}_{ij} + \epsilon_{ij}, \end{aligned} \quad (3.2)$$

em que: Produção_{ij} é a j -ésima observação na i -ésima mesorregião. As variáveis Ano_{ij} , $\text{Temperatura máxima}_{ij}$, $\text{Temperatura mínima}_{ij}$ e Precipitação_{ij} são contínuas, $\beta_0, \beta_1, \beta_2, \beta_3$ e β_4 são os parâmetros da regressão; e os resíduos ϵ_{ij} tem distribuição Normal com média 0 e variância σ_ϵ^2 . Com $i = 1, \dots, 14$ mesorregiões e $j = 1, \dots, 10$ observações que correspondem a cada covariável.

3.2.1 Análise exploratória dos dados

Fazendo primeiro uma análise prévia dos dados com o modelo da expressão (3.2), a Figura 3.2 mostra a matriz de dispersão das observações das variáveis produção, precipitação, temperatura

máxima, temperatura mínima e ano. Também observa-se o histograma de cada uma dessas variáveis o que evidencia que cada uma delas apresenta o comportamento de alguma distribuição assimétrica aparentemente não conhecida. Observa-se a relação e o comportamento da variável produção em função das outras covariáveis bem como mostra-se uma curva de ajuste em cada uma dessas relações além disso observa-se um pequeno grupo de observações que estão fora dessa curva de ajuste, ou seja existe a possibilidade de existir uma variabilidade entre essas observações; portanto com base nesse fato, pode-se utilizar algum tipo de suavização que leve em consideração a variabilidade que apresenta-se.

Figura 3.2: Matriz de dispersão das variáveis produção, precipitação, temperatura máxima, temperatura mínima e ano.

Como passo seguinte, utilizou-se um gráfico semi-normal de probabilidades com envelope simulado da distribuição dos resíduos do modelo (ATKINSON, 1985), para a variável produção em função das outras covariáveis utilizando a expressão (3.2), e na Figura 3.3 observa-se a presença de pontos externos ao envelope representando quase um 20% dos resíduos não contidos nesse envelope, o que indica possível existência de falta de ajuste do modelo devido à existência de correlação entre as observações no sentido temporal que não foi incorporada, o que indica a necessidade de ajustes de modelos com tais características.

Como parte da análise prévia dos dados, na Tabela 3.1, mostra-se as estimativas dos parâmetros da regressão considerando os termos do preditor do modelo como efeitos fixos. As estimativas dos parâmetros das variáveis climáticas: temperatura máxima e precipitação foram altamente significativas o que indica que essas variáveis apresentam uma influência sobre a produção de cana-de-açúcar. No entanto a variável Ano que é considerada uma variável importante nesta análise não foi significativa, por não ser considerado uma possível correlação ao longo do tempo.

Na Figura 3.4, mostra-se o comportamento da produção de cana-de-açúcar em cada mesorregião num período de 10 anos (2006 até 2015) e pode-se ver que a taxa de câmbio da produção em cada mesorregião apresenta muita variação entre o intercepto e inclinação das retas que foram ajustadas, isso leva a considerar modelos mistos com efeitos aleatórios sejam eles independentes ou correlacionados entre intercepto e inclinação em relação ao tempo, neste caso para a covariável Ano.

Figura 3.3: Gráfico semi-normal de probabilidades com envelope simulado para a distribuição dos resíduos do modelo.

Tabela 3.1: Estimativas e erros padrão dos parâmetros da regressão.

Coeficientes	Estimativas	Erros padrão	<i>p</i> -valor
Intercepto	-53677402	1524888769	0.971972
Temperatura máxima	4921015	1394425	0.000571
Temperatura mínima	1689357	1188102	0.157363
Precipitação	-246355	64870	0.000220
Ano	-34137	765464	0.964495

Com base nessa análise prévia dos dados e o desconhecimento do comportamento de cada uma das covariáveis que influenciam a variável produção, procede-se o uso das seguintes abordagens:

3.2.2 Abordagem 1: Modelos mistos

Sabe-se que os modelos mistos permitem modelar a variância quando ela não é homogênea ou quando deseja-se incluir efeitos mais complexos como uma correlação serial quando tem-se amostras com dependência temporal (PINHEIRO e BATES, 2000). Considerando o modelo da expressão (2.54) e utilizando todo o referencial teórico da seção (2.7.1), procede-se a metodologia dos modelos mistos para o caso de produção de cana-de-açúcar.

Logo, o modelo linear de efeitos mistos que é proposto para este caso de estudo na produção de cana-de-açúcar é da forma

$$\begin{aligned} \text{Produção}_{ij} = & \beta_0 + a_{i2} \times \text{Ano}_{ij} + \beta_2 \times \text{Temperatura máxima}_{ij} + \beta_3 \times \text{Temperatura mínima}_{ij} + \\ & \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}, \end{aligned} \quad (3.3)$$

em que a_{i1} é o intercepto aleatório individual para cada curva diferente, assumindo com distribuição normal com média 0 e variância $\sigma_{\text{mesorreg}}^2$; a_{i2} permite as inclinações aleatórias individuais para cada curva diferente, com distribuição normal de média 0 e variância σ_{Ano}^2 ; logo $(a_{i1}, a_{i2})'$ tem distribuição normal com vetor de médias $\mathbf{0}$ e matriz de variância-covariância para os interceptos e inclinações aleatórias denotada por $\mathbf{G}$; os resíduos ϵ_{ij} tem distribuição Normal com média 0 e variância σ_{ϵ}^2 . Os termos aleatórios são assumidos independentes.

Figura 3.4: Gráfico do comportamento da produção de cana-de-açúcar num período de 10 anos.

No caso de supor somente modelos com intercepto aleatório, estaríamos dizendo que todas as curvas individuais são paralelas com a mesma forma da curva; ou seja, esses modelos não permitiriam a possibilidade de modelar cada curva com a sua própria forma e de maneira individual; é por essa razão de propor e utilizar esse modelo com diferentes curvas individuais, isto é, que cada curva apresente uma determinada inclinação o que permite que cada curva individual seja diferente.

Na Tabela 3.2 mostra-se os modelos a serem ajustados, considerando de efeito aleatório na variável Ano, assim como o intercepto a_{i1} , os quais são tratados como modelos com intercepto e inclinação aleatórios independentes, daí que apresenta-se a possibilidade de modelar heterogeneidade individual tanto para os interceptos como também ao longo do tempo. Em cada nível do fator de agrupamento mesorregião, modela-se o efeito aleatório tanto para o intercepto como para a inclinação em relação ao Ano, isto permite mostrar que a relação que existe entre a produção com as demais covariáveis incluindo Ano sejam diferentes em cada mesorregião e que essas variáveis aleatórias estejam correlacionadas.

Tabela 3.2: Modelos mistos (MM) considerando efeito aleatório no intercepto e no tempo (Ano).

Modelos	Expressões
MM1	$\beta_0 + a_{i2} \times \text{Ano}_{ij} + \beta_2 \times \text{Temp. máxima}_{ij} + \beta_3 \times \text{Temp. mínima}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$
MM2	$\beta_0 + a_{i2} \times \text{Ano}_{ij} + \beta_2 \times \text{Temp. máxima}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$
MM3	$\beta_0 + a_{i2} \times \text{Ano}_{ij} + \beta_3 \times \text{Temp. mínima}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$
MM4	$\beta_0 + a_{i2} \times \text{Ano}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$

Para comparar esses modelos serão utilizados os critérios de informação AIC e BIC obtidos por meio da função *lme* (*Linear Mixed-Effects Model*) do pacote *nlme* (*Nonlinear Mixed-Effects Model*) do R CORE TEAM (2017). As estimativas dos parâmetros dos efeitos fixos e das componentes de variância serão obtidas utilizando o método REML que está incluído dentro da função *lme*.

3.2.3 Abordagem 2: Representação P-splines como modelos mistos

O modelo apresentado a seguir está expresso por

$$\begin{aligned} \text{Produção}_{ij} = & \beta_0 + f(\text{Ano}_{ij}) + \beta_2 \times \text{Temperatura máxima}_{ij} + \beta_3 \times \text{Temperatura mínima}_{ij} + \\ & \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}, \end{aligned} \quad (3.4)$$

em que $f(\text{Ano}_{ij})$ é a função desconhecida ou função suave da variável Ano, que mostra a tendência do comportamento dessa variável para a produção de cana-de-açúcar. Estima-se f por meio de *P-splines* utilizando a representação como modelos mistos apresentado na seção (2.7.5), com a finalidade de unificar a estrutura do modelo; a_i é o intercepto aleatório com distribuição Normal com média 0 e variância $\sigma_{\text{mesorreg}}^2$; os resíduos ϵ_{ij} tem distribuição Normal com média 0 e variância σ_{ϵ}^2 .

Mostra-se na Tabela 3.3 os modelos com a função suave na variável Ano a serem ajustados, em que essa variável será considerada de efeito aleatório, assim como o intercepto a . Como no caso de modelos mistos expressados em (3.3), em cada nível do fator de agrupamento mesorregião, modela-se o efeito aleatório tanto para o intercepto como para a inclinação em relação a Ano, o que permite mostrar que a relação existente entre produção e as demais covariáveis incluindo Ano sejam diferentes em cada mesorregião e que essas variabilidades aleatórias estejam correlacionados.

Tabela 3.3: Modelos com função suave na variável Ano, efeito aleatório no intercepto, utilizando *P-splines* com representação como modelos mistos - MPS

Modelos	Expressões
MPS1	$\beta_0 + f(\text{Ano}_{ij}) + \beta_2 \times \text{Temp. máxima}_{ij} + \beta_3 \times \text{Temp. mínima}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$
MPS2	$\beta_0 + f(\text{Ano}_{ij}) + \beta_2 \times \text{Temp. máxima}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$
MPS3	$\beta_0 + f(\text{Ano}_{ij}) + \beta_3 \times \text{Temp. mínima}_{ij} + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$
MPS4	$\beta_0 + f(\text{Ano}_{ij}) + \beta_4 \times \text{Precipitação}_{ij} + a_{i1} + \epsilon_{ij}$

MPS-Modelo *P-spline*

No caso da função suave, utiliza-se como função de base as *B-splines* de terceiro grau polinomial, com 40 nós e uma penalização discreta de segunda ordem. Para testar os modelos utiliza-se os critérios de informação AIC e BIC para comparar esses modelos ajustados utilizando a função *lme* (*Linear Mixed-Effects Model*) do pacote *nlme* (*Nonlinear Mixed-Effects Model*) do R CORE TEAM (2017). As estimativas dos parâmetros dos efeitos fixos e das componentes de variância serão obtidos utilizando o método REML.

3.2.4 Abordagem 3: Modelo misto aditivo generalizado com *P-splines*

Nesta abordagem utiliza-se GAMM para buscar a relação entre as variáveis climáticas e ano com a variável resposta produção de cana-de-açúcar das 14 mesorregiões do estado de São Paulo. A finalidade de utilizar essa abordagem é que muitas vezes não é levada em consideração a possível existência de heterocedasticidade e uma das possíveis causas é que a verdadeira relação entre a variável resposta e as covariáveis seja não linear. Consideram-se dois casos: 1) Realizar um suavizado a cada covariável do modelo e 2) suavizar bidimensionalmente duas covariáveis do modelo a ser utilizado. Para essa abordagem utiliza-se o referencial teórico de (2.11.4), em que a variável resposta produção tem distribuição Gaussiana que pertence a família exponencial de distribuições, com função de ligação identidade e sempre levando em consideração que para cada covariável a ser suavizada, utiliza-se representação *P-splines* como modelos mistos com função de base *B-splines* de terceiro grau polinomial, e uma penalização discreta de segunda ordem.

O modelo completo a ser utilizado é da forma ,

$$\begin{aligned} \text{Produção}_{ij} = & \beta_0 + f(\text{Ano}_{ij}) + f(\text{Temperatura máxima}_{ij}) + f(\text{Temperatura mínima}_{ij}) \\ & f(\text{Precipitação}_{ij}) + a_{i1} + \epsilon_{ij}, \end{aligned} \quad (3.5)$$

com matrizes do modelo misto expressas por,

$$\mathbf{X} = \begin{bmatrix} 1 & \text{Ano}_{11} & \text{Temperatura máxima}_{11} & \text{Temperatura mínima}_{11} & \text{Precipitação}_{11} \\ 1 & \text{Ano}_{12} & \text{Temperatura máxima}_{12} & \text{Temperatura mínima}_{12} & \text{Precipitação}_{12} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \text{Ano}_{1410} & \text{Temperatura máxima}_{1410} & \text{Temperatura mínima}_{1410} & \text{Precipitação}_{1410} \end{bmatrix}$$

$$Z = \begin{bmatrix} Z_{\text{Ano}} : Z_{\text{Temperatura máxima}} : Z_{\text{Temperatura mínima}} : Z_{\text{Precipitação}} : Z_a \end{bmatrix}$$

$$\alpha \sim N \left[0, \begin{pmatrix} I\sigma_{\text{Ano}}^2 & 0 & 0 & 0 & 0 \\ 0 & I\sigma_{\text{Temperatura máxima}}^2 & 0 & 0 & 0 \\ 0 & 0 & I\sigma_{\text{Temperatura mínima}}^2 & 0 & 0 \\ 0 & 0 & 0 & I\sigma_{\text{Precipitação}}^2 & 0 \\ 0 & 0 & 0 & 0 & I\sigma_{\text{Mesorregião}}^2 \end{pmatrix} \right]$$

Nas duas primeiras abordagens, os resíduos são considerados com variância constante; isto significa, que não é levada em conta uma estrutura de correlação. Nesta abordagem é considerado o processo autoregressivo na estrutura de correlação dos resíduos e assim pode-se estimar as curvas suavizando cada uma das covariáveis. Em cada uma das funções suaves do modelo 3.5 é utilizado *P-splines*, já que com essa representação é possível estimar a tendência de Ano e flexibiliza a estrutura dos erros de forma simultânea dos dois efeitos, já que se não se utiliza a representação de *P-splines* como modelos mistos, seria difícil estimar ambos os efeitos.

3.2.4.1 Modelo aditivo e bidimensional suave para temperatura máxima e precipitação

O modelo aditivo a ser ajustado considera duas funções suaves, que ajusta uma mesma curva para uma primeira covariável, e a mesma curva para uma segunda covariável. O modelo aditivo de duas funções para produção de cana-de-açúcar considerando efeitos aleatórios individuais é expresso por

$$\text{Produção}_{ij} = \beta_0 + f(\text{Temp. máxima}_{ij}) + f(\text{Precipitação}_{ij}) + a_{i1} + \epsilon_{ij} \quad (3.6)$$

O modelo bidimensional a ser ajustado considera as duas variáveis climáticas em uma mesma função suave, logo o modelo é expresso por

$$\text{Produção}_{ij} = \beta_0 + f(\text{Temp. máxima}_{ij}, \text{Precipitação}_{ij}) + a_{i1} + \epsilon_{ij} \quad (3.7)$$

e a função suave dessa curva é ajustada utilizando o produto tensor para duas variáveis. Para ajustar esse tipo de modelos utiliza-se a função `gamm` (*Generalized Additive Mixed Models*) do pacote `mgcv` do R CORE TEAM (2017).

4 RESULTADOS

Os resultados obtidos para cada uma das abordagens citadas anteriormente são:

4.1 Abordagem 1: Modelos mistos

Levando em consideração os modelos ajustados descritos na Tabela 3.2, para escolher o melhor modelo ajustado, foram utilizados os critérios de informação AIC e BIC, cujos resultados são mostrados na Tabela (4.1) e segundo o AIC e o BIC, o modelo completo MM1, apresenta o menor valor em relação aos demais modelos, indicando que esse modelo é o mais parsimonioso.

Tabela 4.1: Critérios de informação AIC e BIC para os modelos mistos propostos.

Modelos	AIC	BIC
MM1	4614.818	4640.966
MM2	4642.780	4666.081
MM3	4644.729	4668.030
MM2	4674.804	4695.244

Com base na escolha do modelo MM1, na Tabela 4.2, apresenta-se as estimativas dos efeitos fixos e os componentes de variância para esse modelo, bem como seu erro padrão e p-valor correspondente. Observando o p-valor para Ano, há evidência altamente significativa que este influencia a produção de cana-de-açúcar, seguido da temperatura máxima. No entanto, segundo o p-valor para precipitação e temperatura mínima não mostram nenhuma influência significativa na produção. As estimativas das variâncias dos efeitos aleatórios são bastante altas, especialmente a estimativa da variância do intercepto aleatório que mostra a variabilidade entre as 14 mesoregiões que são bem diferentes entre si. A correlação entre as estimativas das variâncias do intercepto aleatório como para o efeito em relação ao Ano foi de 0.42.

Tabela 4.2: Estimativas dos parâmetros dos efeitos fixos e componentes de variância para o modelo MM1, erro padrão, p-valor.

Parâmetro	Estimativas	Erro padrão	P-valor
<i>Intercepto</i>	61452773	15833531	0.0002
<i>Ano</i>	1018547	315130	0.0016
<i>Precipitação</i>	1316	17549	0.9404
<i>Temp. máxima</i>	-927761	455810	0.0440
<i>Temp. mínima</i>	-699687	501466	0.1655
$\hat{\sigma}_{Ano}^2$	1.180880e+12		
$\hat{\sigma}_{mesorreg}^2$	7.370867e+14		
σ_{ϵ}^2	1.274786e+13		

4.2 Abordagem 2: Representação P-splines como modelos mistos

Considerando os modelos ajustados da Tabela 3.3 com função suave na variável Ano e considerando a representação de P-splines como modelos mistos, mostra-se os valores de AIC e BIC desses modelos ajustados na Tabela 4.3 e o modelo selecionado corresponde ao MPS1 que é o modelo completo.

Com base nesse modelo mostra-se na Tabela 4.4, as estimativas dos parâmetros fixos e dos componentes de variância do modelo e os respectivos erros padrão das estimativas dos efeitos fixos.

Tabela 4.3: Critérios de informação AIC e BIC para os modelos P-*splines* propostos.

Modelos	AIC	BIC
MPS1	4578.998	4608.051
MPS2	4606.370	4632.584
MPS3	4605.393	4631.606
MPS4	4633.523	4656.883

Observando o p-valor para Ano, há evidência altamente significativa que influência a produção de cana-de-açúcar, no entanto, temperatura máxima, temperatura mínima e precipitação não apresentam nenhuma evidência significativa ou seja, o comportamento de cada uma dessas variáveis climáticas não é linear, o que também faz supor que estamos violando a pressuposição de independência dos resíduos, uma vez que somente foi suavizada a variável Ano nesta abordagem, o que indica que o ajuste desse modelo não é bom. As estimativas das variâncias dos efeitos aleatórios são bastante altos, especialmente a estimativa da variância do intercepto aleatório que mostra a variabilidade entre as 14 mesorregiões. A correlação entre as estimativas das variâncias do intercepto aleatório como para a inclinação em relação ao efeito aleatório Ano foi de 0.41.

Tabela 4.4: Estimativas dos parâmetros dos efeitos fixos e componentes de variância para o modelo MPS1, erros padrão, p-valor.

Parâmetros	Estimativas	Erros padrão	P-valor
<i>Intercepto</i>	43077828	14167241	0.0029
<i>Ano</i>	995598	320726	0.0024
<i>Precipitação</i>	-11967	15140	0.4308
<i>Temp. máxima</i>	-353422	394341	0.3719
<i>Temp. mínima</i>	-553048	430363	0.2012
$\hat{\sigma}_{Ano}^2$	1.277233e+12		
$\hat{\sigma}_{mesorreg}^2$	6.942348e+14		
σ_{ϵ}^2	8.474859e+12		

Na Figura 4.1 pode-se observar a média de produção de cada mesorregião ao longo dos 10 anos, utilizando modelo misto sob abordagem P-*splines*, suavizando somente a variável Ano, utilizando base B-*splines* de grau 3, e ordem de penalização 2 com 5 nós, observa-se a forma dessa função suave da variável Ano, não é uma linha reta. A curva ajustada da média (curva de cor azul), encontra-se em um intervalo de confiança de 95% e observa-se também que a produção de cana teve um crescimento até o ano 2011 quando parece estabilizar.

4.2.1 Abordagem 3: Modelo misto aditivo generalizado com P-*splines*

4.2.1.1 Suavizado unidimensional

A seguir, mostram-se os resultados da abordagem 3, considerando como função suave cada uma das variáveis climáticas bem como a variável temporal Ano dentro do modelo aditivo, a vantagem de usar essas funções suaves, é que permitem uma tendência não linear da produção ao longo do tempo bem como o comportamento não linear dos efeitos de cada uma das variáveis climáticas. Além disso, nas abordagens utilizadas anteriormente não foram consideradas a estrutura de correlação nos resíduos, no entanto, nessa abordagem considerou-se a existência dessa correlação utilizando-se o processo autoregressivo de ordem AR(1) como expresso em (2.92), com a representação de P-*splines* como modelos mistos o que permite estimar a tendência de Ano e flexibiliza a estrutura dos erros de forma simultânea para os dois efeitos.

Figura 4.1: Intervalo de confiança para a média da produção de cana-de-açúcar ao longo do tempo.

Na Figura 4.2, mostra-se a tendência da covariável Ano em duas situações: 1) Sem considerar uma estrutura de correlação nos resíduos, o que causa dois efeitos: Figura 4.2.(a) uma curva extremamente suave que é uma reta com quantidade de suavização de um grau de liberdade efetivos (g.l.e.) e na Figura 4.2.(b) com a ajuda de um correlograma, é possível verificar que os resíduos são correlacionados, ou seja, mostra-se a violação clara da pressuposição de independência dos resíduos. 2) Para resolver esse problema, foi introduzida no modelo uma estrutura de autocorrelação AR(1) nos resíduos, logo observa-se que na Figura 4.2.(c), apresenta uma curva suave com 3.55 graus de liberdade efetivos e na Figura 4.2.(d) apresenta-se um correlograma com estrutura autoregressiva AR(1) nos resíduos. Também pode-se observar que na Figura 4.2.(c) que a forma do suavizador indica que a produção de cana-de-açúcar teve um crescimento até 2010 e a partir desse ano apresentou uma estabilidade até o ano 2015. Essa mesma Figura pode ser comparada com a Figura 4.1 que mostra o intervalo de confiança de 95% da produção de cana-de-açúcar ao longo do tempo; isto significa que as bandas que formam esses intervalos de confiança da curva, são exatamente as bandas dos suavizadores obtidos utilizando o modelo GAMM.

Na Tabela 4.5, mostra-se a comparação dos modelos ajustados, um modelo sem considerar uma estrutura de correlação nos resíduos e outros dois modelos utilizando estrutura autoregressiva AR(1) e AR(2), considerando função suave na covariável Ano, a escolha do modelo é feita usando como critério de seleção AIC e BIC. Segundo os dois critérios utilizados, o modelo escolhido é o modelo que considera um processo autoregressiva AR(1).

Tabela 4.5: Comparação de modelos considerando estrutura de correlação autoregressiva nos resíduos para a covariável Ano.

Modelo	AIC	BIC
Sem correlação	5213.269	5225.035
AR(1)	4662.327	4677.036
AR(2)	4664.420	4682.070

Com base na escolha do modelo GAMM considerando estrutura de correlação com processo autoregressiva AR(1) nos resíduos, procedeu-se o ajuste individual das covariáveis. A seguir mostra-se a Tabela 4.6 com as funções suaves estimadas de cada covariável, com seus respectivos graus de liberdade efetivos (g.l.e), parâmetros de suavização, testes F e o p-valor correspondentes. Pode-se observar que as

Figura 4.2: Formas das funções suaves estimadas para a covariável Ano na produção de cana-de-açúcar utilizando modelo generalizado aditivo misto em que (a) Curva estimada sem considerar estrutura de correlação nos resíduos e (b) Função de autocorrelação sem considerar o efeito dos resíduos. (c) Curva estimada considerando estrutura de correlação nos resíduos e (d) Função de autocorrelação considerando estrutura autoregressiva AR(1).

variáveis de interesse foram altamente significativos à produção de cana-de-açúcar.

Tabela 4.6: Aproximação das funções suaves estimadas individuais das covariáveis no GAMM.

Função suave	g.l.e	λ	F	p-valor
s(Temperatura máxima)	4.709	1.226891	16.81	2.25e-12
s(Temperatura mínima)	4.266	1.809065	9.27	1.61e-06
s(Precipitação)	4.887	0.746584	17.4	2.62e-12
s(Ano)	3.555	1317.997	9.884	1.51e-06

A seguir, na Figura 4.3 mostram-se as formas das funções suaves estimadas ou suavizadores de cada covariável climática sobre a produção de cana-de-açúcar para o GAMM e mostra claramente evidências do efeito de cada uma dessas covariáveis que segue um padrão não linear. Assim, pode-se observar que a linha sólida é a função suave estimada e as linhas pontilhadas são as bandas ao 95% de confiança. Em cada gráfico da Figura 4.3, o eixo horizontal mostra os valores de temperatura mínima, temperatura máxima e precipitação e no eixo vertical, a contribuição da função suave para os valores ajustados dessa produção de cana-de-açúcar. Como o efeito da temperatura mínima, temperatura máxima e precipitação não é linear, tais funções suaves estimadas são significativos com 4.27, 4.71 e 4.89 graus de liberdade efetivos respectivamente, uma linha reta teria somente um grau de liberdade efetivos. Na Figura

4.3a observa-se que o suavizador para temperatura mínima apresenta dois periodos nos quais a produção de cana-de-açúcar teve uma maior produção relativa. Na Figura 4.3b a forma da função suave estimada da variável temperatura máxima sugere que existe um período alto onde a produção de cana-de-açúcar teve maior produtividade a medida em que a temperatura foi maior e na Figura 4.3c o suavizador da variável precipitação mostra um período de produção bastante alto quando os valores dessa precipitação são valores baixos.

Figura 4.3: Forma dos suavizadores para temperatura mínima, máxima e precipitação utilizando modelo generalizado aditivo misto.

4.2.1.2 Suavizado do modelo aditivo e bidimensional para temperatura máxima e precipitação

Considerando o ajuste do modelo aditivo expresso por (3.6), os resultados são mostrados na Tabela 4.7 onde mostra-se as aproximações dos graus de liberdade efetivos, parâmetros de suavização, os testes F e o p-valor das funções suaves estimadas de cada variável do modelo aditivo. Nessa tabela pode-se observar que os graus de liberdade efetivos das funções suaves estimadas são diferentes dos obtidos de forma unidimensional (Tabela 4.6); o que se deve pelo fato que os graus de liberdade efetivos associados a uma covariável, em geral, mudará de valor em presença de outra covariável no mesmo modelo aditivo ajustado, pois as funções suaves desses modelos GAMM não estão aninhados. Também pode-se observar que as funções suaves estimadas do modelo aditivo são altamente significativas.

Com base nos resultados obtidos da Tabela 4.7, pode-se obter as Figuras 4.4a) e b) que mostram de forma individual as curvas suaves estimadas para cada covariável do modelo aditivo; no entanto a Figura 4.4c) mostra a superfície aditiva dos dois suavizadores que deram lugar a esse modelo. Comparando as curvas do modelo aditivo para temperatura máxima e precipitação em relação as curvas ajustadas

Tabela 4.7: Funções suaves estimadas do modelo aditivo para as covariáveis temperatura máxima e precipitação no GAMM.

Função suave	g.l.e	λ	F	p-valor
s(Temperatura máxima)	5.067	0.6040863	11.71	1.09e-09
s(Precipitação)	2.088	109.8897028	11.03	3.26e-05

obtidas individualmente da Figura 4.3, pode-se observar que a forma do suavizador para temperatura máxima não apresentou evidências de mudar a forma da curva, somente mudou os graus de liberdade efetivos para 5.07, no entanto em relação a precipitação observa-se que mudou por completo a tendência dessa curva, isso deve-se ao fato de acrescentar uma função suave de uma covariável ao modelo aditivo, portanto, os graus de liberdade efetivos mudarão pela presença do outro termo suave.

Figura 4.4: Curvas ajustadas com um modelo aditivo para a) temperatura máxima, b) precipitação e c) superfície aditiva considerando as duas variáveis climáticas, temperatura máxima e precipitação.

Considerando o modelo (3.7) para a função das duas variáveis, ajusta-se essa curva utilizando o produto tensor com *P-splines* no modelo GAMM. Os resultados são apresentados na Tabela 4.8, onde a função suave bidimensional estimada mostra aproximadamente 11.69 graus de liberdade efetivos, com um p-valor ($<2e-16$) altamente significativo. Mostra-se também os parâmetros de suavização de cada variável da função suave bidimensional o que mostra efetivamente que os suavizadores de cada eixo da superfície são não isotrópicos, também chega-se a conclusão que pode-se utilizar uma função bidimensional com duas variáveis de diferentes medidas e dimensões.

Tabela 4.8: Função suave estimada do modelo bidimensional das covariáveis temperatura máxima e precipitação no GAMM.

Função suave	g.l.e	λ	F	p-valor
te(T.máxima,Precipitação)	11.69		15.2	<2e-16
Temp. máxima $_{\lambda}$	-	0.4334616	-	-
Precipitação $_{\lambda}$	-	363.0988183	-	-

te: produto tensor de duas variáveis

A Figura 4.5 mostra a superfície bidimensional que foi formada pelo modelo (3.7) com a função suave considerando as variáveis temperatura máxima e precipitação. Nessa figura há dois gráficos: a) na forma de curvas de nível de superfície e b) na forma de uma superfície, mostrando em cada eixo a variável climática em estudo. Observa-se que nessa superfície ajustada, há diferenças na tendência da temperatura máxima. Também pode-se ver que o eixo da variável precipitação da superfície, apresenta a tendência de uma reta o que é mostrado em valor numérico do parâmetro de suavização da Tabela 4.8 que é bastante grande (363.10).

Figura 4.5: Curvas ajustadas em um modelo bidimensional para temperatura máxima e precipitação: a) curvas de nível de superfície e b) superfície bidimensional.

Na Tabela 4.9 mostra-se os critérios de informação AIC e BIC dos modelos aditivo e bidimensional considerando as variáveis climáticas temperatura máxima e precipitação. Segundo o AIC e BIC o modelo bidimensional é o melhor modelo que ajusta essa curva bidimensional.

Tabela 4.9: AIC e BIC dos modelos aditivo e bidimensional de temperatura máxima e precipitação.

Modelo	AIC	BIC
Modelo aditivo	5154.744	5178.277
Modelo bidimensional	5138.258	5164.732

Na Tabela 4.10 mostra-se as funções suaves estimadas por variável climática e ano, como também por mesorregião, e pode-se observar as diferenças significativas com diferentes níveis de significância para as mesorregiões: Araçatuba, Araraquara, Assis, Bauru, Marília, Piracicaba, Presidente Prudente, Ribeirão Preto e São José do Rio Preto.

Com base em todos os resultados obtidos, obteve-se a Figura 4.6, a qual mostra a predição de valores ajustados do modelo para a produção de cana-de-açúcar ao longo dos anos. Observando essa

Tabela 4.10: Funções suaves estimadas das covariáveis por mesorregião.

Função suave	g.l.e	F	p-valor	
s(Ano):Mesorregião Araçatuba	2.889	53.511	< 2e-16	***
s(Ano):Mesorregião Araraquara	2.157	16.489	7.59e-08	***
s(Ano):Mesorregião Assis	1.218	10.252	0.00161	**
s(Ano):Mesorregião Bauru	1.000	3.724	0.05651	.
s(Ano):Mesorregião Itapetininga	1.000	0.716	0.39953	
s(Ano):Mesorregião Litoral Sul Paulista	1.000	0.122	0.72722	
s(Ano):Mesorregião Macro Metropolitana Paulista	1.000	0.088	0.76734	
s(Ano):Mesorregião Marília	1.000	9.985	0.00209	**
s(Ano):Mesorregião Piracicaba	1.924	3.072	0.04137	*
s(Ano):Mesorregião Presidente Prudente	2.205	100.435	< 2e-16	***
s(Ano):Mesorregião Ribeirão Preto	2.912	47.568	< 2e-16	***
s(Ano):Mesorregião São José do Rio Preto	2.845	251.287	< 2e-16	***
s(Ano):Mesorregião Vale do Paraíba Paulista	1.000	2.125	0.14812	
te(Temperatura máxima,Precipitação)	4.021	3.338	0.00922	**
s(Temperatura máxima):Mesorregião Araçatuba	1.002	0.571	0.44651	
s(Temperatura máxima):Mesorregião Araraquara	1.000	9.783	0.00256	**
s(Temperatura máxima):Mesorregião Assis	1.712	3.290	0.03397	*
s(Temperatura máxima):Mesorregião Bauru	2.638	5.471	0.00162	**
s(Temperatura máxima):Mesorregião Campinas	2.099	2.079	0.10571	
s(Temperatura máxima):Mesorregião Itapetininga	1.000	0.565	0.45498	
s(Temperatura máxima):Mesorregião Litoral Sul Paulista	1.000	0.000	0.99969	
s(Temperatura máxima):Mesorregião Macro Metropolitana Paulista	1.000	0.205	0.65204	
s(Temperatura máxima):Mesorregião Marília	1.001	0.447	0.50591	
s(Temperatura máxima):Mesorregião Piracicaba	2.489	12.733	3.60e-06	***
s(Temperatura máxima):Mesorregião Presidente Prudente	1.000	7.572	0.00754	**
s(Temperatura máxima):Mesorregião Ribeirão Preto	2.600	14.290	0.00122	**
s(Temperatura máxima):Mesorregião São José do Rio Preto	2.117	1.697	0.09828	.
s(Temperatura máxima):Mesorregião Vale do Paraíba Paulista	1.000	0.002	0.96240	
s(Precipitação):Mesorregião Araçatuba	1.000	0.340	0.561686	
s(Precipitação):Mesorregião Araraquara	1.000	0.149	0.700496	
s(Precipitação):Mesorregião Assis	1.000	0.018	0.894174	
s(Precipitação):Mesorregião Bauru	2.117	5.922	0.012787	*
s(Precipitação):Mesorregião Campinas	1.000	0.384	0.537593	
s(Precipitação):Mesorregião Itapetininga	1.000	0.823	0.367702	
s(Precipitação):Mesorregião Litoral Sul Paulista	1.000	0.000	0.999806	
s(Precipitação):Mesorregião Macro Metropolitana Paulista	1.000	0.026	0.872698	
s(Precipitação):Mesorregião Marília	1.000	0.627	0.431276	
s(Precipitação):Mesorregião Piracicaba	1.889	7.707	0.005101	**
s(Precipitação):Mesorregião Presidente Prudente	1.003	0.756	0.385981	
s(Precipitação):Mesorregião Ribeirão Preto	1.990	7.425	0.001057	**
s(Precipitação):Mesorregião São José do Rio Preto	1.668	2.090	0.259902	
s(Precipitação):Mesorregião Vale do Paraíba Paulista	1.000	0.000	0.989416	

Nível de significância: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

te: Produto tensor de duas variáveis.

figura pode-se concluir que as mesorregiões mencionadas são aquelas que foram significativas na Tabela 4.10 e apresentam maiores produções de cana-de-açúcar.

Figura 4.6: Valores preditos para produção de cana-de-açúcar ao longo dos anos.

5 DISCUSSÃO

Neste trabalho apresentou-se toda a abordagem e metodologia dos *P-splines* baseados em modelos mistos e modelos mistos aditivos generalizados com estruturas suavizadoras expressas com variáveis unidimensional e bidimensionalmente com aplicação sobre a produção de cana-de-açúcar em função de variáveis climáticas como temperatura e precipitação e variável temporal Ano. O motivo de usar essa abordagem como método de suavização é que na maioria das vezes não é conhecido o comportamento ou tendência dessas covariáveis climáticas mas sabe-se que elas influenciam diretamente sobre a variável resposta. Na parte do referencial teórico, apresentaram-se várias formas de suavizar as curvas utilizando diferentes métodos de suavização com *splines* (*Splines* de regressão, suavização por *splines*, *P-splines* e representação de *P-splines* como modelos mistos) motivando seu uso em relação a algumas vantagens. A seguinte função

$$f(x) = \cos(5\pi x),$$

foi utilizada para a simulação dos dados, considerando-se 200 observações que estão no intervalo $[0, 1]$. Seguidamente avaliou-se a função $f(x)$, em cada um dos valores e acrescentou-se um erro aleatório (valor simulado de uma Normal $(0, 1)$), logo os dados simulados são da forma

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, 1).$$

Com essa simulação de dados, pode-se observar a melhoria nos ajustes das curvas quando utiliza-se suavização por *splines*, *P-splines* ou representação de *P-splines* como modelos mistos; no entanto, deve-se destacar o uso do ajuste da curva por representação de *P-splines* como modelos mistos, por proporcionar melhores ajustes nas diferentes curvas para os dados simulados assim como apresentar de forma automática a estimação do parâmetro de suavização λ . Durante o processo de simulação dos dados foram utilizados diferentes números de nós (5, 15 e 40) e diferentes valores nos parâmetros de suavização, considerando como base *B-splines* de terceiro grau polinomial tanto para *splines* de regressão, *P-splines* e representação *P-splines* como modelos mistos, no caso de suavização por *splines* foi utilizado como base os *splines* cúbicos.

Para ajustar as curvas com os *splines* de regressão e *P-splines* foram utilizados os códigos feitos no R que utilizados pela Professora Dra. Maria Durbán¹, ministrados no minicurso “*Splines con penalizaciones: Teoría y aplicaciones*” . Para montar as bases *B-splines* no R, foi utilizado a fórmula recursiva de DE BOOR (1979). Para o ajuste de suavização por *splines*, utilizou-se uma função disponível no pacote *smooth.splines* do R, esta função internamente executa uma validação cruzada generalizada, que fornece o parâmetro λ mais adequado para tal ajuste. Para o ajuste de curvas utilizando *P-splines* como modelos mistos, foram utilizadas ferramentas de modelos mistos disponíveis na livreria *nlme* do R. Pode-se observar que o ajuste utilizando *P-splines* como modelos mistos foi mais preciso, independente do número de nós utilizados. No entanto, no caso de *splines* de regressão, acrescentou-se o número de nós básicos, as curvas obtidas com esse método de suavização apresentam muitas oscilações em relação à função original.

Uma característica em destaque das *P-splines* é o uso da matriz de penalização, por estar definida como a matriz discreta baseada nas diferenças entre os coeficientes básicos adjacentes dos *B-splines*. Como argumentado por EILERS e MARX (1996), esta penalização é uma “boa aproximação discreta da segunda derivada do quadrado integrável de uma função”. As bases *B-splines* de grau 3 e uma penalidade de segunda diferença de ordem, chegaram a ser uma escolha comum. O tipo de penalidade dependerá do tipo de base a ser utilizada EILERS e MARX (1996), assim como RUPPERT e CARROLL (2000) utilizaram para o ajuste de curvas os *P-splines*, uma penalização de segunda ordem de diferenças e bases com funções de polinômios truncados.

¹Universidad Carlos III Madrid-España.

A representação de *P-splines* como modelos mistos também foi estendida para o caso bidimensional de uma função suave com duas variáveis, onde utilizou-se o produto tensor, cuja metodologia é facilitadora para o ajuste das curvas em duas dimensões, obtendo como resultado uma superfície suavizada.

Também vimos que os modelos mistos podem ser estendidos a modelos lineares generalizados mistos (efeitos lineares mas com distribuições não normais) e modelos generalizados aditivos mistos (GAMM) com efeitos fixos não lineares e não normais. Esse último modelo, por possuir estruturas aditivas suaves, permite a combinação de vários modelos, e para suavizar estruturas aditivas utiliza a base *B-splines* e uma penalidade discreta baseada nas diferenças entre os coeficientes básicos adjacentes dos *B-splines* como são os *P-splines*. Dentre dessas estruturas aditivas também foi considerado o suavizado bidimensional onde foi utilizada toda a representação de *P-splines* como modelos mistos e a combinação do uso de produto tensor para as bases *B-splines* e penalizações discretas.

Na aplicação, utilizou-se toda a abordagem dos *P-splines* para relacionar a variável resposta produção de cana-de-açúcar em função da variável tempo e variáveis climáticas, considerando essas co-variáveis como funções desconhecidas a serem estimadas, utilizando diversas abordagens: como simples modelos mistos, representação de *P-splines* como modelos mistos, e modelos GAMM onde as estruturas aditivas desse modelo utiliza representação de *P-splines* como modelos mistos tanto unidimensional como bidimensional. Durante as análises feitas, mesmo considerando efeito aleatório na variável temporal Ano, os efeitos aleatórios dos resíduos foram considerados como independentes, mas quando foi utilizado modelos GAMM foi levada em consideração a estrutura de correlação para os resíduos (essas classes de modelos permitem incluir esses tipos de correlação nos resíduos) neste caso utilizou-se processo autoregressivo AR(1), que com essa estrutura melhora o ajuste das curvas tanto para as variáveis climáticas como para a variável resposta, mostrando de modo prático o uso desse tipo de modelo GAMM. Cabe fazer destaque que nesses modelos para o caso de aplicação do presente trabalho, a resposta foi considerada com distribuição normal e função de ligação identidade.

Os resultados obtidos usando o modelo GAMM e a Figura 4.6, que corresponde aos valores preditos cujas curvas mostram o comportamento da produção de cana-de-açúcar por mesorregião do estado de São Paulo ao longo dos anos, coincidem efetivamente com as mesorregiões que apresentam maior produtividade de cana como mostra o mapa de zoneamento edafoclimático para a produção de cana-de-açúcar (ver Figura 3.1).

Espera-se com o presente trabalho ter mostrado como os *P-splines* e a representação dos *P-splines* como modelos mistos são usados, e que para estimar as curvas com os *P-splines* não são necessários de métodos iterativos de estimação como *backfitting algorithm*. Com o REML consegue-se obter as estimativas dos parâmetros fixos, componentes de variância e parâmetro de suavização e tendo resultados úteis pode-se obter as predições do modelo.

Por último, decidiu-se aplicar a abordagem de *P-splines* na área agroambiental, como foi na produção de cana-de-açúcar em função de variáveis climáticas, sabendo que estas últimas apresentam comportamentos que não tem uma distribuição conhecida, portanto a maioria de trabalhos que estudam essa relação de produção de cana-de-açúcar e suas variáveis climáticas utilizam nas suas análises estatísticas, modelos lineares simples ou mistos, sem levar em consideração o comportamento de tais variáveis o que muitas vezes leva a conclusões erradas. Como exemplo, se somente consideramos como resultados obtidos nas Tabelas 4.2 e 4.4 utilizando modelos mistos e representação *P-splines* como modelos mistos respectivamente, neste último em que foi considerado somente como função suave a variável Ano, observa-se que as variáveis precipitação, temperatura máxima e temperatura mínima não foram significativas, no entanto, utilizando modelo GAMM com os *P-splines* nas funções suaves aditivas, os resultados são diferentes, já que nesse caso leva-se em consideração o comportamento dessas variáveis climáticas, assim como também do ano.

6 CONCLUSÃO

Com base em todo o trabalho desenvolvido no presente estudo, pode-se concluir:

- Ao analisar dados ou observações que apresentam ruído ou não seguem um determinado padrão ou comportamento probabilístico, a estimação das funções suaves com a abordagem dos *P-splines* melhora o ajuste das curvas, o que representa soluções sensíveis aos problemas complexos,
- O uso de *P-splines* depende do tipo de base a ser utilizada e o mais recomendado é o uso de bases *B-splines*, o que permite um baixo custo computacional. Alguns autores WOOD (2003), WOOD (2006), WOOD *et al.* (2017) e ZUUR *et al.* (2009) utilizam outras bases que não foram estudadas no presente trabalho, tais como os *thin plate splines*,
- Os GAMM são mais utilizados nas áreas de ecologia e biologia por possuírem delineamentos e estruturas complexas e o comportamento ou tendência dos dados não seguem um padrão linear. Além disso, as estruturas dos componentes aditivos desses modelos utilizam outros métodos de suavização para as covariáveis tais como suavização por *splines*, com bases *splines* cúbicos ou *thin plate splines*, o que muitas vezes gera um alto custo computacional,
- Visando um trabalho futuro, pretende-se utilizar os mesmos dados da produção de cana-de-açúcar, considerando neste caso variáveis geográficas como latitude e longitude das localizações de cada mesorregião e de cada posto meteorológico considerado e assim, utilizar a abordagem dos *P-splines*.

REFERÊNCIAS

- ABDELKADER, N. B., 2009 *Regresión no Paramétrica en R*. Máster en Estadística Aplicada, España.
- AKAIKE, H., 1973 Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika* **60**: 255–265.
- ATKINSON, A., 1985 *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford Statistical Science Series, Clarendon Press.
- BIEZUNER, R. J., 2014 Notas de seminário de iniciação científica. pp. 1–14, Minas Gerais, Departamento de Matemática - UFMG.
- BRESLOW, N. E. e D. G. CLAYTON, 1993 Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9–25.
- BRUMBACK, B. A. e J. A. RICE, 1998 Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**: 961–976.
- CHEN, C., 2007 Generalized additive mixed models. *Communications in Statistics - Theory and Methods* **29**: 1257–1271.
- CIIAGRO, 2017 Centro Integrado de Informações Agrometeorológicas. <http://www.ciiagro.sp.gov.br>, Acesso em: fev. 2017.
- CRAVEN, P. e G. WAHBA, 1978 Smoothing noisy data with spline functions. *Numerische Mathematik* **31**: 377–403.
- CURRIE, I. D. e M. DURBAN, 2002 Flexible smoothing with p-splines: a unified approach. *Statistical Modelling* **2**: 333–349.
- CURRIE, I. D., M. DURBÁN, e P. H. C. EILERS, 2004 Smoothing and forecasting mortality rates. *Statistical Modelling* **4**: 279–298.
- CURRIE, I. D., M. DURBÁN, P. H. C. EILERS, M. DURBAN, e P. H. C. EILERS, 2006 Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B* **68**: 1–22.
- DE BOOR, C., 1979 A practical guide to splines. *Advances in Mathematics* **32**: 81.
- DEAN, C. B. e J. D. NIELSEN, 2007 Generalized linear mixed models: A review and some extensions. *Lifetime Data Analysis* **13**: 497–512.
- DELICADO, P., 2008 Curso de Modelos no Paramétricos p. 200.
- DESCHÊNES, O. e M. GREENSTONE, 2012 The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather: Reply. *The American Economic Review* **102**: 3761–3773.
- DIAS, R., 1997 O uso de splines em regressão não paramétrica. pp. 1–13, São Paulo, Coloquio do departamento de Estatística, IME-USP.
- DIAS, R., 2001 Regressão não paramétrica. pp. 1–15, São Paulo, Universidade Estadual de Campinas.
- DURBAN, M., 2002 Flexible smoothing with P -splines: a uni ed approach pp. 333–349.

- DURBÁN, M., 2014 Splines con penalizaciones: Teoría y aplicaciones. pp. 1–55, Universidad Carlos III de Madrid.
- EILERS, P. H. e B. D. MARX, 2003 Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems* **66**: 159 – 174.
- EILERS, P. H. C. e B. D. MARX, 1996 Flexible smoothing with B -splines and penalties. *Statistical Science* **11**: 89–121.
- EILERS, P. H. C. e B. D. MARX, 2010 Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**: 637–653.
- EILERS, P. H. C., B. D. MARX, e M. DURBÁN, 2015 Twenty years of P-splines. *Sort* **39**: 149–186.
- ESALQ, 2017 Posto Meteorológico “Professor Jesús Marden dos Santos”. Escola Superior de Agricultura Luiz de Queiroz - Universidad de São Paulo. <http://www.leb.esalq.usp.br/posto/>, Acesso em: fev. 2017.
- EUBANK, R., 1988 *Spline Smoothing and Nonparametric Regression*. Statistics (Marcel Dekker), M. Dekker.
- GREEN, P. J. e B. W. SILVERMAN, 1994 *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, volume 58.
- GU, C. e G. WAHBA, 1991 Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing* **12**: 383–398.
- HARVILLE, D. A., 1977 Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**: 320–338.
- HASTIE, T., R. TIBSHIRANI, e J. FRIEDMAN, 2009 The Elements of Statistical Learning. *Elements* **1**: 337–387.
- HASTIE, T. J. e R. J. TIBSHIRANI, 1990 Generalized Additive Models.
- HORN, R. A. e C. R. JOHNSON, 1991 *Topics in Matrix Analysis*. Cambridge University Press.
- IBGE, 2017 Instituto Brasileiro de Geografia e Estatística. Produção agrícola municipal 2006 e 2015. <http://www.ibge.gov.br>, Acesso em: fev. 2017.
- INMET, 2017 Instituto Nacional de Meteorologia. Ministerio da Agricultura, Pecuária e Abastecimento. <http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas>, Acesso em: fev. 2017.
- LEE, D.-J. e M. DURBAN, 2011 P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling* **11**: 49–69.
- LIN, X. e D. ZHANG, 1999a Inference in Generalized Additive Mixed Models by Using Smoothing Splines. *Source Journal of the Royal Statistical Society. Series B (Statistical Methodology)* *J. R. Statist. Soc. B* **61**: 381–400.
- LIN, X. e D. ZHANG, 1999b Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **61**: 381–400.
- MA, S. e J. S. RACINE, 2013 Additive regression splines with irrelevant categorical and continuous regressors. *Statistica Sinica* **23**: 515–541.

- MCCULLAGH, P. e J. A. NELDER, 1989 *Generalized linear models*. Chapman and Hall-CRC Monographs on Statistics and Applied Probability, Chapman and Hall-CRC, second edition.
- MONTEIRO, J., 2009 *Agrometeorologia dos cultivos: o fator meteorológico na produção agrícola*. Ministério da agricultura, pecuária e abastecimento (MAPA). Instituto nacional de meteorologia (INMET).
- O'SULLIVAN, F., 1986 A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1**: 502–518.
- PATTERSON, H. D. e R. THOMPSON, 1971 Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.
- PINHEIRO, J. C. e D. M. BATES, 2000 *Mixed-effects models in S and S-PLUS*. Statistics and computing, Springer, first edition.
- R CORE TEAM, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RAO, C. e J. KLEFFE, 1988 *Estimation of variance components and applications*. North-Holland series in statistics and probability, North-Holland.
- ROBINSON, G. K., 1991 That blup is a good thing: The estimation of random effects. *Statistical Science* **6**: 15–32.
- RUPPERT, D. e R. J. CARROLL, 2000 Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**: 205–223.
- RUPPERT, D., M. WAND, e R. CARROLL, 2003 *Semiparametric regression*, volume 54.
- SCHALL, R., 1991 Estimation in generalized linear models with random effects. *Biometrika* **78**: 719–727.
- SCHWARZ, G., 1978 Estimating the dimension of a model. *The Annals of Statistics* **6**: 461–464.
- SEARLE, S., G. CASELLA, e C. MCCULLOCH, 2009 *Variance Components*. Wiley Series in Probability and Statistics, Wiley.
- SPEED, T., 1991 [that blup is a good thing: The estimation of random effects]: Comment. *Statist. Sci.* **6**: 42–44.
- STONE, M., 1974 Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society* **36**: 111–147.
- UNICA, 2017 União da Indústria de Cana de Açúcar. <http://www.unica.com.br/>, Acesso em: Jun. 2017.
- VERBYLA, A. P., B. R. CULLIS, M. G. KENWARD, e S. J. WELHAM, 1999 The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48**: 269–300, Part 3.
- WAHBA, G., 1990 *Spline models for observational data*. CBMS-NSF Regional Conference series in applied mathematics 59, Society for Industrial and Applied Mathematics, illustrated edition edition.
- WAND, M. P., 2003 Smoothing and mixed models. *Computational Statistics* **18**: 223–249.
- WANG, Y., 1998a Mixed effect smoothing spline analysis of variance **60**: 159–174.
- WANG, Y., 1998b Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* **93**: 341–348.

- WASSERMANN, L., 2006 All of nonparametric statistics.
- WATSON, G. S., 1964 Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* (1961-2002) **26**: 359–372.
- WOOD, A. S., F. SCHEIPL, e M. S. WOOD, 2015 Package ‘*gamm4*’ **4**: 1–9.
- WOOD, S. N., 2003 Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**: 95–114.
- WOOD, S. N., 2006 Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62**: 1025–1036.
- WOOD, S. N., Z. LI, G. SHADDICK, e N. H. AUGUSTIN, 2017 Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data. *Journal of the American Statistical Association* **0**: 1–12.
- ZEGER, S. e P. DIGGLE, 1994 Semi-parametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics* **50**: 689–699.
- ZHANG, D., X. LIN, J. RAZ, e M. SOWERS, 1998 Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**: 710–719.
- ZUUR, A. F., E. N. IENO, N. J. WALKER, A. A. SAVELIEV, e G. M. SMITH, 2009 *Mixed Effects Models and Extensions in Ecology with R*, volume 1.