UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Introdução à Análise Descritiva de Dados Funcionais

Autor: Maicom Frozza
Orientador: Professor Dr. Flávio Augusto Ziegelmann

Porto Alegre, 06 de Julho de 2010.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Introdução à Análise Descritiva de Dados Funcionais

Autor: Maicom Frozza

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Flávio Augusto Ziegelmann
Professor Dr. Danilo Marcondes Filho

Porto Alegre, 06 de Julho de 2010.

Dedico este trabalho a todos que contribuíram na minha educação (meus pais, meu irmão, minha família e os professores que passaram por mim nesses anos de estudo) a fim de que um dia eu pudesse concluir esta etapa!

Agradecimentos

Ao mesmo tempo em que essa etapa está se concluindo, outra começará em seguida, com novas responsabilidades e novos desafios. Mas, para que esse dia chegasse, muitas foram as superações que eu tive que enfrentar: a vida em uma cidade grande, longe de casa, da minha família e da tranquilidade; a diferença do nível de aplicação aos estudos que as disciplinas ministradas na UFRGS exigem, principalmente no início do curso; entre outras coisas que eu passei nos primeiros anos que morei em Porto Alegre.

Tudo isso eu não consegui sozinho. Foram várias as pessoas que me ajudaram a enfrentar e transpor esses obstáculos. Assim, eu gostaria de agradecer em primeiro lugar a toda minha família, em especial meus pais Leopoldo e Fiorinda, que não mediram esforços para que eu pudesse realizar este meu sonho. Agradecer meu irmão Michel, que esteve junto comigo em todos os meus anos de vida. Agradecer meus nonos Rodolpho e Irma e a tia Luiza, que de uma maneira especial estiveram ao meu lado, onde quer que eu fosse.

Agradecer, também, aos meus amigos. Todos que passaram por mim, desde a minha infância até agora na faculdade. Entre eles, o Junior, o Rafael, o André, o Ivan, o Rodrigo, o Maico, de Boa Vista do Sul/RS; o Postal, o Rodrigo, o Erechim, o Siqueira e a Adri, o Jucelino e a Nice, a Juci, o Émerson e o pessoal da portaria: Marcelo, Paulo, Lessa, Emerson, Everaldo, Pedro, Angelita, Bete, Silvio, da Casa do Estudante da UFRGS; ao Renan, Rodrigo, Jeferson, Julianderson, Julio e aos demais colegas, em especial, aos formandos: Gilberto, Andréia, Fernanda e Carol; e a todos que eu conheci nestes anos.

Agradecer a todos professores que foram fundamentais para que eu conseguisse chegar ao que eu sou hoje. Desde aqueles que estiveram comigo nos níveis iniciais na Escola de Ensino Médio Marcelino Champagnat até aqueles que passaram por mim na UFRGS.

Agradecer ao Marcos, a Bel e ao Rafael, quando estive ainda prestando vestibular para ingressar nessa longa jornada que agora está chegando ao fim. Agradecer ao Egídio e a Helena, porque eles foram minha segunda família aqui em Porto Alegre.

Enfim, muito obrigado a todos que de uma forma ou de outra tornaram esse meu sonho realidade.

Resumo

A Análise Estatística de Dados Funcionais é bastante recente e fornece uma nova forma de visualização, análise e interpretação dos dados. Essa nova maneira de representação está baseada na ideia de que cada observação é uma função real e não simplesmente um escalar ou vetor. Os motivos para analisarmos os dados como funções são, entre outros: i) hoje em dia, há uma maior disponibilidade ao acesso a dados funcionais; ii) em alguns casos pode ser problemático tratar uma observação funcional como não sendo uma; e iii) podemos utilizar a informação das derivadas das funções, uso incompleto ou mesmo inexistente nas técnicas não funcionais. Este trabalho irá tratar da Análise Descritiva de Dados Funcionais, a fim de oferecer um texto introdutório voltado para uma apresentação inicial do assunto, motivando as ideias fundamentais e exemplificando com uma aplicação a dados reais. Assim, abordaremos os conceitos básicos da Análise Descritiva de Dados Funcionais, notações usuais nesta área e as estatísticas descritivas funcionais, analisando os dados do Estudo de Crescimento de Berkeley. Esse estudo, realizado em 1954, investigou problemas relacionados ao desenvolvimento físico, motor e mental, no início da vida de algumas pessoas.

Palavras-chave: Dados Funcionais. Estatística Descritiva. Suavização. *Splines*.

Abstract

The Statistical Analysis of Functional Data is fairly recent and provides a new way of viewing, analyzing and interpreting data. This new way of representation is based on the idea that each observation is a real function, not simply a scalar or vector. Reasons for analyzing the data as functions are, i) nowadays, there is a greater accessibility to functional data, ii) in some cases may be problematic to treat some data as not being a functional and iii) we can use information from the functions derivatives, which is not so common in non-functional techniques. This paper will address the Descriptive Analysis of Functional Data in order to offer an introductory text. It provides fundamental ideas, motivating and illustrating them with an application to real data. Thus, we will focus on basic concepts of Descriptive Analysis of Functional Data, usual notations in this area and the functional descriptive statistics, analyzing data from the Berkeley Growth Study. This study, conducted in 1954, investigates issues related to physical, mental and motor development in the early life of some people.

Keywords: Functional Data Analysis. Descriptive statistics. Smoothing. Splines.

Sumário

Lista de Figuras

1. Introdução

Ao longo do tempo a humanidade vem aprimorando as técnicas e os recursos que utiliza no seu cotidiano. Com o advento de tecnologias inovadoras, tem-se conseguido aperfeiçoar essas técnicas, que ajudam na realização das diversas atividades de uma forma mais eficaz, desde as mais simples até a resolução de problemas bastante complexos.

Após o aparecimento dos computadores, que trouxeram consigo grande economia de tempo e de mão-de-obra, a Estatística chegou ao nível que está hoje. Entretanto, o progresso das técnicas estatísticas continua em constante desenvolvimento. Na última década, tem havido também uma mudança, por exemplo, quanto à forma de análise de determinados tipos de dados. Aqueles oriundos de medições que teoricamente poderiam ser feitas continuamente, ao invés de serem tratados como escalares ou vetores de dimensão finita, tem a possibilidade de serem tratados como uma função analítica (ou seja, como curvas) e, por isso, são chamados de dados funcionais.

Por exemplo, ao estudarmos o crescimento de crianças podemos estar interessados em, além de estimar a curva de crescimento, simultaneamente estimar a velocidade de crescimento e/ou aceleração como função do tempo para cada indivíduo. Desenvolveu-se, portanto, uma nova metodologia para nos auxiliar em problemas deste tipo, denominada de Análise de Dados Funcionais (*Functional Data Analysis*). Termo, este, delimitado para o conjunto de métodos usados para realizar a análise de dados que são apresentados sob a forma de funções. Note que, neste caso, o termo funcional se refere à estrutura dos dados e não à sua forma explícita, pois na prática os dados são observados de maneira discreta.

A Análise de Dados Funcionais (ADF) ainda é um tema recente, aos poucos vem começando a ser divulgada no Brasil, contudo está em crescente desenvolvimento visto que incitou grande interesse da comunidade estatística mundial. A maioria das técnicas iniciais desenvolvidas para a ADF foi introduzida por Ramsay and Dalzell (1991) e Ramsay and Silverman (1997). Nesses estudos foram apresentados resultados de desenvolvimento na área através de trabalhos centrados na adaptação das ferramentas e técnicas usualmente empregadas na Estatística que tornam possível o estudo de dados funcionais. Muitas destas

adaptações incidem sobre técnicas de análise de dados multivariados.

Através do surgimento de novas tecnologias, os dados funcionais estão sendo observados e investigados com uma frequência cada vez maior em diversos campos do conhecimento, tais como Arqueologia, Economia, Meteorologia, Ciências Biológicas, Engenharia, etc. Em muitos destes casos, o interesse não está somente na estimação das curvas, mas na estimação das derivadas e integrais destas curvas, além de outros funcionais. Assim, mesmo que não exista uma exigência para que os dados sejam suaves, a suavidade ou outra regularidade será um aspecto chave da análise quando estivermos interessados nesses funcionais.

Mas por que analisar os dados como funções? Quais são as vantagens que isso nos traz? As razões práticas para nós analisarmos os dados de um ponto de vista funcional são diversas. Entre elas podemos citar as seguintes:

- As observações funcionais ocorrem cada vez mais frequentemente em contextos aplicados, à medida que aumenta a facilidade de coleta automatizada e quase-contínua de dados. Além disso, procedimentos de suavização e interpolação podem produzir representações funcionais de conjuntos finitos de observações.
- Alguns problemas de modelagem são mais naturais quando se pensa em termos funcionais, embora somente um número finito de observações esteja disponível.
- Os objetivos de uma análise podem ser funcionais de forma natural, por exemplo, faz mais sentido usar um número finito de variáveis (através de suas observações) para estimar toda uma função, suas derivadas ou valores de outras operações funcionais ou seria mais adequado ver essas mesmas variáveis como pontos de avaliação de um funcional?
- Levar em consideração certos aspectos, tais como a suavidade, para dados multivariados ocasionados por processos funcionais.
- Imagens, bem como curvas, de grande interesse em dias atuais, podem ser representadas como dados funcionais ou como parâmetros funcionais em modelos.

Os objetivos da análise de dados funcionais são essencialmente os mesmos que os de outros ramos da estatística. Eles são os seguintes: representar os dados de forma a auxiliar futuras análises; exibir os dados de maneira que várias características sejam destacadas; estudar fontes importantes de padrão e variação entre os dados; explicar a variação em um resultado ou variável dependente usando a informação da variável independente; comparar

dois ou mais conjuntos de dados com respeito a certos tipos de variação, onde dois conjuntos de dados podem conter diferentes conjuntos de replicações das mesmas funções, ou diferentes funções para um conjunto comum de replicações; reduzir a dimensionalidade do problema; estudar a dinâmica de séries temporais funcionais; entre outros.

Já o objetivo desse trabalho é oferecer um texto introdutório, motivando o assunto e exemplificando com dados reais. Assim, abordaremos principalmente os conceitos básicos da ADF, suas notações usuais e estatísticas descritivas funcionais.

O Capítulo 2 faz uma breve descrição sobre as notações usualmente utilizadas na literatura de dados funcionais. Além disso, comenta alguns itens fundamentais para o desenvolvimento das técnicas básicas da ADF que estão relacionadas com características e propriedades funcionais.

A ADF é indicada quando um conjunto de observações funcionais forma a amostra, ou seja, quando existe uma função aleatória suave subjacente responsável pela geração da amostra de funcionais. Entretanto, se a função não tem um comportamento suave, sugere-se a realização de uma pré-suavização dos dados antes da ADF ser posta em prática. Assim, o Capítulo 3 aborda o método de suavização via *splines*. Nele está descrito o procedimento para realizar esse método de suavização, as propriedades relacionadas e os detalhes sobre o parâmetro de suavização.

O Capítulo 4 enfatiza os principais aspectos da Análise de Dados Funcionais. Dentre os quais, podemos citar, por exemplo, as propriedades desse tipo de análise, o conceito formal e, também, prático de dados funcionais, além de algumas estatísticas funcionais básicas.

Enfim, no último capítulo, o Capítulo 5, apresentamos alguns resultados da Análise Descritiva de Dados Funcionais aplicada a dados reais referentes ao Estudo de Crescimento de Berkeley (Tuddenham and Snyder, 1954). Esses resultados são relativos aos principais conceitos descritos no Capítulo 4. Os códigos de programação para a ADF deste trabalho foram escritos utilizando as funções disponíveis no pacote *fda* do *software* R (versão 2.11.1), e encontram-se no Anexo A.

2. A suavização via *splines*

O objetivo desse capítulo é estudar como a penalização da rugosidade (do inglês *roughness penalty*) funciona no caso funcional mais simples. Nesse caso, nosso intuito é estimar uma função não periódica *x* com base em observações discretas embebidas em ruído, dispostas em um vetor *y*.

Estudos assintóticos sobre o estimador obtido utilizando o método de suavização por *splines* podem ser encontrados em Eubank (1988), onde é possível, entre outras coisas, verificar a consistência desse estimador. Silverman (1984) apresenta também um importante resultado. Ele mostra que, sob certas condições, a suavização *spline* corresponde aproximadamente à suavização por *kernel* com a janela *h* dependendo da densidade local dos pontos de observação. Green and Silverman (1994) discutem uma variedade de problemas estatísticos que podem ser abordados usando penalizações por falta de suavidade.

2.1. As Funções *Splines* e B-*splines*

2.1.1. Funções *Splines*

Devido a sua estrutura simples e às boas propriedades de aproximação, os polinômios são amplamente utilizados na prática para aproximar funções. Desse modo, o intervalo $T = [a,b]$, onde está definida a função, é dividido em subintervalos menores da forma $[x_0, x_1],\ldots,[x_k, x_{k+1}]$ e então um polinômio p_i (de grau menor que função que será suavizada) é usado para aproximação em cada intervalo subintervalo.

Esse procedimento produz uma função de aproximação polinomial por partes $s(\cdot)$, onde $s(t) = p_i(t)$ em $[x_i, x_{i+1}]$, $i = 0,\ldots,k$. Os valores $x_0, x_1,\ldots, x_k, x_{k+1}$ são chamados de nós (do inglês *knots*), sendo que x_0 e x_{k+1} são os nós exteriores e os demais $x_1,\ldots, x_k$, os nós interiores.

No caso geral, as partes do polinômio $p_i(t)$, usadas na aproximação de cada subintervalo, são independentes umas das outras e não formam uma função contínua em $[a,b]$. Isso não pode ser aceito para aproximar uma função suave. Portanto, é necessário que as partes do polinômio sejam unidas suavemente e também que sejam deriváveis um certo número de vezes. Como resultado, obtém-se uma função polinomial por partes, suave, chamada função *spline*.

Um *spline* de ordem m ($ordem = grau + 1$) com k nós interiores em $x_1,\ldots,x_k$ é qualquer função da forma

$$s(t) = \sum_{i=0}^{m-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i \left(t - x_i\right)^{m-1}, \tag{2.1}$$

onde os coeficientes $\theta_0,\ldots,\theta_{m-1},\delta_1,\ldots,\delta_k$ são números reais, $\{x_1,\ldots,x_k\}$ são os nós interiores e $\left\{1,t,t^2,\ldots,t^{m-1},\left(t-x_1\right)^{m-1},\ldots,\left(t-x_k\right)^{m-1}\right\}$ são as funções bases. Assim, pode-se concluir que qualquer função *spline* é uma combinação linear de $m+k$ funções base.

A escolha do número de nós tem sido um assunto de muita pesquisa. Muitos nós produzem uma curva sem quase nenhuma suavidade (interpolação), já poucos nós suavizam demais os dados.

O conjunto de funções *spline* de ordem *m* e nós interiores $x_1,\ldots,x_k$ é chamado de espaço *spline* e é denotado por $\boldsymbol{S}_m\left(x_1,\ldots,x_k\right)$. Mais ainda, o espaço *spline* é um espaço linear de dimensão $m+k$ (Schumaker, 1981).

Uma extensão dos *splines*, os chamados B-*splines* formam uma base de espaços spline (Schumaker, 1981).

2.1.2. B-*splines*

Os B-*splines*, assim como os *splines*, são constituídos de pedaços de polinômios unidos de forma especial em certos valores chamados nós. Esse tipo de base possui duas vantagens: a computação muito rápida e a grande flexibilidade.

Por exemplo, um B-*spline* de grau 1 consiste de dois pedaços lineares, um pedaço de x_0 a x_1, e outro de x_1 a x_2. Os nós são x_0, x_1 e x_2. É claro que é possível construir um conjunto tão amplo de B-*splines* quanto se queira, basta introduzir mais nós.

Assim, seja o intervalo T = [a, b] dividido em k+1 subintervalos menores da forma $[x_0, x_1], \ldots, [x_k, x_{k+1}]$. Como em cada intervalo m B-*splines* de ordem m são não nulos, o número total de nós para a construção dos B-*splines* deve ser $k+2m$. Portanto, alguns nós adicionais precisam ser incluídos (ver Schumaker, 1981). Para isso, $m-1$ nós são adicionados no início e no final da sequência de tal forma que $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_{m-1} \leq x_0$ e $x_{k+1} \leq \tau_{m+k+2} \leq \tau_{m+k+3} \leq \ldots \leq \tau_{k+2m}$. Os valores desses nós adicionais são arbitrários. É comum fazer com que $\tau_1 = \ldots = \tau_{m-1} = x_0$ e $x_{k+1} = \tau_{m+k+2} = \ldots = \tau_{k+2m}$.

A partir disso, De Boor (1978) desenvolveu um algoritmo para calcular B-*splines* de qualquer ordem através de B-*splines* de ordem menor, ou seja, é possível calcular os B-*splines* através de uma relação de recorrência. Devido ao fato de um B-*spline* de ordem 1 ser uma constante em um intervalo entre dois nós, o cálculo de B-*splines* de qualquer ordem é facilitado.

Algoritmo de De Boor (1978): O *i*-ésimo B-*spline* de ordem *m* para uma sequência crescente de nós $\tau = \{\tau_i\}$ pode ser calculado como

$$B_{i,m}(t) = \frac{t - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(t) + \frac{\tau_{i+m} - t}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(t), \tag{2.2}$$

onde

$$B_{i,1}(t) = \begin{cases} 1 & \text{se } \tau_i \leq t \leq \tau_{i+1} \\ 0 & \text{caso contrário.} \end{cases} \tag{2.3}$$

Os B-*splines* podem ter também múltiplos nós e, assim, é necessário ter certo cuidado ao usar a relação de recorrência (3.3) para evitar divisão por zero (ver De Boor, 1978).

Assim, considerando B-*splines* de ordem *m* com *k* nós interiores, é possível escrever a função *x* como

$$x(t) = \sum_{i=1}^{K=m+k} c_i B_{i,m}(t) \, . \tag{2.4}$$

Por sua vez, a estimativa suave de *x* calculada a partir dos dados observados é dada por

$$\hat{x}(t) = \sum_{i=1}^{K} \hat{c}_i B_i(t) \ . \tag{2.5}$$

Para mais detalhes sobre os polinômios *splines* e B-*splines* podem ser encontrados em De Boor (1978), Eubank (1988) e Green and Silverman (1994).

2.2. Dois objetivos na estimação de uma função

O método de suavização *spline* estima uma curva *x* através das observações $y_i = x(t_j) + \varepsilon_j$ e deixa explícitos dois objetivos conflitantes na estimação de curvas em geral. Por um lado, deseja-se assegurar que a curva estimada produza um bom ajuste aos dados. Por outro lado, não se quer um ajuste tão bom ao ponto do resultado ser uma curva totalmente não suave ou localmente variável.

Esses dois objetivos, que competem entre si, correspondem ao Erro Quadrático Médio (EQM):

$$\text{Erro Quadrático Médio} = \text{Variância} + \text{Vício}^2.$$

Outras funções perda podem ser escolhidas em certas situações. O balanço entre viés e variância se aplica também a esses casos, embora não com a mesma decomposição acima.

O EQM pode ser frequentemente reduzido através da introdução de um pouco de vício com o objetivo de diminuir a variância, e essa é uma razão chave pela qual se impõe suavidade à curva estimada. Quando se requer que a estimativa varie suavemente de um valor para o outro, o que se faz é buscar a informação dos dados vizinhos. Desse modo fica expresso que uma certa regularidade na função *x* é esperada. Ao compartilhar essa informação, a curva estimada se torna mais estável, ao preço de um aumento no vício. A penalização da não suavidade torna explícito o que é sacrificado no vício para que o EQM ou outra função perda seja melhorada.

2.3. Quantificando a suavidade

Uma maneira comum de quantificar a noção de suavidade de uma função será apresentada nesta seção. O quadrado da segunda derivada de uma função *x* em *t*, $\left[D^2x(t)\right]^2$, é frequentemente chamado de curvatura da função em *t*. Assim, uma medida natural da suavidade de uma função é a integral do quadrado da sua segunda derivada, ou seja,

$$PEN_2(x) = \int \left[D^2x(s)\right]^2 ds \ . \tag{2.6}$$

Isso avalia a curvatura total presente na função *x*, ou alternativamente, o grau em que *x* se distancia de uma reta. Consequentemente espera-se que funções que oscilam muito apresentem valores altos de $PEN_2(x)$, isso porque suas segundas derivadas, em módulo, são elevadas ao longo de grande parte do intervalo de interesse.

Muitas análises de dados funcionais requerem a estimação de derivadas, ou porque elas são o interesse direto, ou porque elas são importantes de alguma forma na análise. A penalização (2.6) não é adequada, uma vez que ela controla a curvatura de *x* propriamente dita e somente a inclinação de Dx. Isso não requer nem ao menos que a segunda derivada seja contínua, quanto mais suave.

Se a derivada de ordem *m* for a mais alta a ser utilizada, deve-se, na verdade, usar na penalização as derivadas de ordem $m+2$, a fim de controlar a curvatura de $D^m x$. Por exemplo, a estimativa da aceleração é melhor se for utilizada a seguinte penalização:

$$PEN_4(x) = \int \left[D^4x(s)\right]^2 ds, \tag{2.7}$$

uma vez que ela controla a curvatura em D^2x.

2.4. A soma dos quadrados dos erros penalizada

Seja $x(\mathbf{t})$ o vetor $n\times 1$ que resulta da função *x* sendo avaliada no vetor $n\times 1$ de argumentos **t**. A soma dos quadrados dos erros penalizada é definida como

$$SQEPEN_\lambda = \left[\mathbf{y} - x(\mathbf{t})\right]' \mathbf{W}\left[\mathbf{y} - x(\mathbf{t})\right] + \lambda \cdot PEN_2(x) \ . \tag{2.8}$$

Se $\mathbf{W} = \mathbf{I}$, ou seja, se a suposição padrão para os erros é assumida, tem-se

$$SQEPEN_{\lambda} = \left[\mathbf{y} - x(\mathbf{t})\right]' \left[\mathbf{y} - x(\mathbf{t})\right] + \lambda \cdot PEN_2(x) \ . \tag{2.9}$$

A estimativa da função é obtida encontrando a função *x* que minimiza $SQEPEN_{\lambda}$ sob o espaço de funções para o qual $PEN_2(x)$ está definido.

O parâmetro λ é um parâmetro de suavização que mede o equilíbrio entre o ajuste aos dados, medido pela soma dos quadrados dos erros no primeiro termo, e a variabilidade da função *x*, quantificada por $PEN_2(x)$ no segundo termo. Na prática, se o parâmetro de suavização λ é zero, não há nenhuma penalidade por falta de suavidade. Conforme λ aumenta, aumenta a penalização sobre a falta de suavidade e as curvas estimadas tornam-se mais suaves.

Assim, conforme λ se torna cada vez maior, funções que não são lineares sofrem uma penalização da não suavidade também cada vez maior através do termo $PEN_2(x)$ e, consequentemente, o critério $SQEPEN_{\lambda}$ dará maior ênfase à suavidade de *x* e menor ênfase para o ajuste aos dados. Por essa razão, conforme $\lambda \to \infty$, a curva ajustada $\hat{x}$ deve-se aproximar da regressão linear padrão dos dados observados.

Por outro lado, para λ pequeno a curva tende a se tornar mais e mais variável uma vez que há menos penalização para a sua não suavidade e, conforme $\lambda \to 0$, a curva ajustada se aproxima de uma interpolação dos dados, satisfazendo $\hat{x}(t_j) = y_j$ para todo *j*. Contudo, mesmo nesse caso limite, a curva obtida não é arbitrariamente variável. Ao invés disso, ela é a curva mais suave duas vezes diferenciável que ajusta exatamente os dados.

2.5. A estrutura da suavização spline

Suponha que não se faça nenhuma suposição sobre a função *x*, exceto que ela tem segunda derivada. Assuma também que os pontos amostrais de observação t_j, $j = 1,\ldots,n$ são distintos. Que tipo de função poderia minimizar a soma dos quadrados dos erros penalizada?

Um notável teorema, apresentado por De Boor (2001), diz que a curva *x* que minimiza $SQEPEN_\lambda$ é um *spline* cúbico com nós nos pontos de observação t_j. Note que não foi feita nenhuma suposição sobre como *x* é construída. A estrutura *spline* de *x* é uma consequência desse teorema, em que uma função objetivo é otimizada com respeito a toda uma função (ver também Schoenberg, 1964a e Schoenberg, 1964b).

A suavização *spline* se adapta naturalmente aos pontos de observação não igualmente espaçados, e, assim, leva vantagem em regiões onde a densidade dos dados é alta e, ao mesmo tempo, produz uma estimativa suave nas regiões onde existem poucas observações. A técnica computacional mais comum para suavização *spline* é usar uma expansão por B-*splines* com nós nos pontos de observação e, assim, minimizar o critério (2.8) com respeito aos coeficientes da expansão.

2.6. Como a suavização spline é calculada?

Uma função pode ser descrita como uma combinação linear de funções base, ou seja,

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}(t), \tag{2.10}$$

onde **c** é o vetor dos coeficientes c_k de tamanho *K* e ϕ é o vetor de funções base, também de tamanho *K*.

Sem a penalização da não suavidade, o vetor **c** que minimiza a soma dos quadrados dos erros é dado por

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y}, \tag{2.11}$$

onde $\boldsymbol{\Phi}$ é uma matriz $n \times K$ contendo os valores das *K* funções base calculadas nos *n* pontos de observação $t_1,\ldots,t_n$, **W** é uma matriz de pesos que permite uma possível estrutura de covariância entre os erros e **y** é o vetor de dados discretos a serem suavizados, onde $\mathbf{y} = x(\mathbf{t}) + \boldsymbol{\varepsilon}$. A expressão correspondente para o vetor de valores ajustados é

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y} = \mathbf{S}\mathbf{y}, \tag{2.12}$$

onde $\mathbf{S} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}$ é o operador de projeção correspondente ao sistema de base ϕ.

É possível reescrever a penalização da não suavidade $PEN_m(x)$ em termos matriciais

$$PEN_m(x) = \int \left[D^m x(s)\right]^2 ds = \mathbf{c}'\mathbf{R}\mathbf{c}, \tag{2.13}$$

onde

$$\mathbf{R} = \int D^m \phi(s) D^m \phi'(s) ds, \tag{2.14}$$

é uma matriz quadrada de dimensão K cujas entradas são

$$R_{ij} = \int D^m \phi_i(s) D^m \phi_j(s) ds \ . \tag{2.15}$$

Dessa forma, a soma dos quadrados dos erros penalizada pode ser reescrita como

$$SQEPEN_\lambda = (\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})' \mathbf{W} (\mathbf{y} - \boldsymbol{\Phi}\mathbf{c}) + \lambda \mathbf{c}'\mathbf{R}\mathbf{c} \ . \tag{2.16}$$

Agora, o vetor de coeficientes estimado que minimiza (2.16) é dado por

$$\hat{\mathbf{c}} = \left(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \lambda \mathbf{R}\right)^{-1} \boldsymbol{\Phi}'\mathbf{W}\mathbf{y} \ . \tag{2.17}$$

Por sua vez, como $\hat{\mathbf{y}} = \hat{x}(\mathbf{t}) = \boldsymbol{\Phi}\hat{\mathbf{c}}$, tem-se que a matriz "chapéu" usando a penalização da não suavidade é dada por

$$\mathbf{S}_\lambda = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \lambda \mathbf{R})^{-1} \boldsymbol{\Phi}'\mathbf{W} \ . \tag{2.18}$$

Esse operador mais geral (2.18) pode ser chamado de operador de sub-projeção, porque, ao contrário do operador de projeção, $\mathbf{S}_\lambda$ não satisfaz a relação de idempotência. Outra aplicação importante de $\mathbf{S}_\lambda$ está no cálculo dos graus de liberdade da suavização *spline*,

$$gl_\lambda = \mathrm{tr}\left(\mathbf{S}_\lambda\right) \ . \tag{2.19}$$

onde para qualquer matriz quadrada $\mathbf{A}$, $\mathrm{tr}(\mathbf{A})$ é o traço de $\mathbf{A}$, ou seja, a soma dos elementos de sua diagonal.

O cálculo da matriz **R** geralmente requer a aproximação numérica da integral em (2.14), embora as expressões exatas possam ser calculadas quando funções base de Fourier ou B-*splines* estão envolvidas (seção 5.2.6, Ramsay and Silverman, 2005).

2.7. A escolha do parâmetro de suavização

Existem duas abordagens distintas em relação à escolha do parâmetro de suavização λ. A primeira abordagem considera livre a escolha do parâmetro de suavização como uma importante característica do procedimento. O que se faz é utilizar diferentes parâmetros e, assim, escolher aquele que, de certa forma, produz a estimativa que melhor se ajusta aos dados. Isso faz com que esse método seja subjetivo, porém muito utilizado na prática porque ele é uma ótima opção quando se tem que ajustar uma única curva.

A outra abordagem lida com a necessidade de se ter um procedimento automático para a escolha de λ com base nos dados. Pode-se dizer que condicionado na escolha do método automático a ser usado, essa é uma forma objetiva de escolha de λ.

Os métodos automáticos não precisam ser utilizados de forma decisiva. Podem, por exemplo, serem usados para a escolha de um valor inicial para um possível refinamento. Esses métodos são essenciais quando a curva estimada é usada como parte integrante de um outro procedimento mais complexo ou se o método é usado frequentemente em muitos conjuntos de dados.

Existem diferentes procedimentos automáticos de escolha do parâmetro de suavização. O mais conhecido de todos é o método de validação cruzada, outro também estudado é o método da validação cruzada generalizada. Ambos os métodos serão apresentados nas duas próximas seções.

2.7.1. Validação Cruzada

A motivação básica para o método de validação cruzada (VC) está relacionada com a predição. Assumindo que o erro aleatório possui média zero, a curva verdadeira x tem a

seguinte propriedade: se uma observação *y* é tomada no ponto *t*, o valor $x(t)$ é a melhor predição de *y* em termos de erro quadrático médio. Então, um bom estimador $\hat{x}(t)$ para $x(t)$ seria aquele que produzisse um pequeno valor de $\{y - \hat{x}(t)\}^2$ para uma nova observação *y* no ponto *t*.

É claro que, na prática, quando o método de suavização é aplicado em um simples conjunto de dados, novas observações não estão disponíveis. A técnica de validação cruzada produz "novas observações" através dos dados como é descrito a seguir.

Segundo Green and Silverman (1994), considere um dado valor λ para o parâmetro de suavização. Tome a observação y_i em t_i como sendo uma nova observação, omitindo-a do resto dos dados. Denote a curva estimada sem a *i*-ésima observação usando λ como parâmetro de suavização como $\hat{x}^{(-i)}(t;\lambda)$. Sabe-se que $\hat{x}^{(-i)}(t;\lambda)$ minimiza

$$\sum_{j \neq i} \{y_j - \hat{x}(t_j)\}^2 + \lambda \int [D^2 x(s)]^2 \, ds \, . \tag{2.20}$$

A qualidade de predição de $\hat{x}^{(-i)}$ pode ser julgada através de quão bem o valor $\hat{x}^{(-i)}(t_j)$ se aproxima de y_i. Uma vez que a escolha da observação a ser omitida é arbitrária, a eficácia do procedimento de suavização com o parâmetro λ pode ser quantificada através do critério de validação cruzada

$$VC(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{x}^{(-i)}(t_i;\lambda)\}^2 \, . \tag{2.21}$$

A idéia básica da validação cruzada é escolher o valor de λ que minimiza $VC(\lambda)$. Não se pode garantir que a função de validação cruzada tenha um único mínimo, então, deve-se tomar cuidado com essa minimização. Uma boa ideia para esse caso é buscar o mínimo num certo intervalo de valores de λ e depois refinar essa busca.

À primeira vista, olhando (2.21), parece que para obter $VC(\lambda)$ é necessário resolver *n* problemas de suavização separadamente, de forma a encontrar as *n* curvas $\hat{x}^{(-i)}$. Porém, será visto a seguir que isso não é necessário.

É importante lembrar que as estimativas encontradas através da suavização *spline* dependem linearmente dos dados y_i através da equação

$$\hat{x}(\mathbf{t}) = \mathbf{S}_\lambda \mathbf{y} \ , \tag{2.22}$$

onde a matriz "chapéu" $\mathbf{S}_\lambda$ está definida na equação (2.18).

Usando a matriz $\mathbf{S}_\lambda$ é possível obter uma forma mais simples de calcular o critério de validação cruzada. A fórmula é dada por

$$VC(\lambda) = n^{-1} \sum_{i=1}^{n} \left(\frac{y_i - \hat{x}(t_i)}{1 - S_{\lambda_{ii}}} \right)^2 , \tag{2.23}$$

onde $\hat{x}$ é a estimativa por suavização *spline* calculada para todo o conjunto de dados $\{(t_i, y_i)\}$, com parâmetro de suavização λ. Por sua vez, $S_{\lambda_{ii}}$ é o *i*-ésimo valor da diagonal de $\mathbf{S}_\lambda$.

A equação (2.23) mostra que, conhecendo as entradas da diagonal de $\mathbf{S}_\lambda$, o critério de validação cruzada pode ser calculado através dos resíduos $\{y_i - \hat{x}(t_i)\}$ quando a suavização *spline* é aplicada em todo o conjunto de dados. Portanto, não existem problemas de suavização adicionais a serem resolvidos.

A prova desse resultado pode ser encontrada em Green and Silverman (Lema 3.1, seção 3.2.1, 1994). Ela está baseada na demonstração de um lema apresentado por Craven e Wahba (1979), que produz a seguinte identidade matemática:

$$y_i - \hat{x}^{(-i)}(t_i) = \frac{y_i - \hat{x}(t_i)}{1 - S_{\lambda_{ii}}} \ . \tag{2.24}$$

2.7.2. Validação Cruzada Generalizada

A validação cruzada generalizada (VCG), uma forma modificada de validação cruzada, é um método comum para a escolha do parâmetro de suavização desenvolvido por Craven e Wahba (1979).

A equação (2.23) mostra os resíduos divididos pelos fatores $1 - S_{\lambda_{ii}}$. A ideia básica da validação cruzada generalizada é substituir esses fatores pelo valor médio deles $1 - n^{-1}\mathrm{tr}(\mathbf{S}_\lambda)$. Uma vez que agora tem-se o mesmo fator para todo *i*, o seguinte critério é obtido:

$$VCG(\lambda) = n^{-1} \frac{\sum_{i=1}^{n} \left[y_i - \hat{x}(t_i) \right]^2}{\left[1 - n^{-1} \mathrm{tr}(\mathbf{S}_\lambda) \right]^2}, \tag{2.25}$$

onde $\mathrm{tr}(\mathbf{S}_\lambda)$ é o traço da matriz $\mathbf{S}_\lambda$.

Da mesma forma como para a validação cruzada, a escolha do parâmetro de suavização é feita encontrando o valor de λ que minimiza o critério $VCG(\lambda)$.

Há muitas razões para não colocar completamente nossa confiança em qualquer método automático para seleção de λ, como o método VCG. Por exemplo, se as derivadas são necessárias, então o nível de suavização que é automaticamente selecionado pode dar derivadas complicadas. Além disso, estes métodos também são altamente sensíveis à suposição de erros independentes. Uma sugestão é começar com o valor que minimizar a medida GCV, e depois explorar os valores de λ próximos (acima e abaixo) deste valor para ver como se comporta (Ramsay *et al*, 2010). Gu (2002) oferece uma discussão detalhada da teoria e de questões computacionais associados aos métodos $VC(\lambda)$, $VCG(\lambda)$ e outros métodos para escolher λ.

3. Análise Descritiva de Dados Funcionais

Uma nova maneira de analisar os dados, na qual não se observa simplesmente mais um conjunto de escalares ou vetores, mas sim um conjunto de funções, está ganhando cada vez mais espaço na comunidade científica mundial da Estatística. À modelagem deste tipo de dados se denomina Análise de Dados Funcionais (ADF).

A ADF está focada em representar os dados através de funções a fim de simplificar as manipulações e a evidenciar as principais características que esses dados apresentam, permitindo a análise de padrões e variações nessas representações. Alguns dos seus objetivos são

1. Possibilitar a manipulação da função por métodos conhecidos da estatística. Note que as funções devem ser suaves (derivadas contínuas); se não forem é necessário que os dados sejam submetidos a um método de suavização antes que a ADF seja aplicada.
2. Destacar aspectos expressivos dos dados analisados.
3. Identificar a presença de padrões dentro do conjunto de dados: em diversas situações, aplicamos alguma operação sobre os nossos dados como, por exemplo, derivadas, integrais ou outros funcionais, para facilitar a visualização.
4. Explicar o comportamento das variáveis de saída em questão.
5. Estimar a dimensionalidade finita do problema.
6. Estudar a dinâmica no caso de séries temporais funcionais.

Uma estratégia para analisar dados funcionais de uma forma genérica é por meio das três etapas básicas: exploratória, confirmatória e preditiva. Exploratória para identificar as características inerentes aos dados, bem como revelar informações ocultas, a fim de auxiliar na análise das técnicas que podem ser empregadas de uma forma eficiente. Confirmatória para realizar inferências, supondo uma estrutura presente e aferindo algumas informações específicas ou hipóteses que podem ser confirmadas pelos dados. Preditiva para realizar afirmações, através dos padrões encontrados, sobre dados não observados, habitualmente para dados futuros.

Neste capítulo serão abordados os fundamentos da Análise Descritiva de Dados Funcionais. Os principais conceitos envolvidos nesse tipo de análise, suas propriedades e algumas estatísticas funcionais básicas.

3.1. Características Funcionais

Uma característica funcional pode ser considerada como um evento associado a um determinado valor *t*. Ou seja, a maioria dessas características é definida por uma localização, uma amplitude e/ou uma medida da sua dimensão. Por exemplo, o tamanho de um pico ou vale é uma questão de amplitude e, também, de largura. Desse modo, os níveis (valores da função que consideramos importantes) são eventos unidimensionais, cruzamentos são bidimensionais, e os picos e vales são tridimensionais.

A dimensionalidade de uma característica funcional nos diz muito sobre a quantidade de informação que será necessária para estimá-la. Por exemplo, mesmo uma quantidade pequena de erro observacional dos dados nos obrigará a fornecer cinco em vez de três valores da função em locais dentro de um pico, e nos dados com níveis de erro comum na Análise de Dados Funcionais seria aceitável sete a onze valores para cada pico (seção 2.4.1, Ramsay and Silverman, 2005).

Pode-se dizer que a dimensionalidade prática de uma função é a quantidade total de informação necessária para defini-la. Além disso, muitas vezes ignora-se o fato de o funcional aleatório não ser observável diretamente, ou seja, ignora-se o erro de observação e/ou de aproximação (oriundo da suavização dos dados originais). Contudo, em alguns estudos como Bathia, Yao and Ziegelmann (2010) e Hall and Vial (2006), esse erro é levado em consideração.

A dimensionalidade tem uma grande importância como um indicador de complexidade na ADF. As funções são potencialmente infinito-dimensionais, ou seja, uma especificação completa de uma função *x* pode exigir o conhecimento do seu valor $x(t)$ para cada *t*, logo a dimensionalidade de uma função pode ser arbitrariamente grande. Isso implica que nunca podemos coletar dados suficientes para estimar exatamente estas funções. No entanto, na prática, trabalha-se com funções que não possuem tanta complexidade.

3.2. Dados Funcionais

Por ser ainda um campo recente na Estatística, a ADF não possui uma definição delimitada. Entretanto, alguns aspectos comuns dos dados funcionais que aparecem frequentemente na literatura devem ser destacados.

- Conceitualmente, dados funcionais são continuamente definidos. Porém, na prática são, em geral, observados discretamente e registrados computacionalmente de maneira finito-dimensional, mas isso não afeta as interpretações e nem as análises.

- O dado individual é toda a função. Os dados funcionais podem ser independentes uns dos outros, entretanto para os valores dentro do mesmo dado funcional obviamente não existem suposições sobre a independência.

- Os dados necessariamente não precisam ser suaves, mas a suavidade ou outra regularidade será importante para a aplicação de algumas técnicas da análise. Além disso, algumas suposições de suavidade serão apropriadas para funções envolvidas na modelagem dos dados observados, como as funções médias.

Em geral, a análise de dados funcionais trabalha com dados nos quais a *i*-ésima observação é uma função real (definida num espaço de funções $\mathfrak{L}_2$), $x(t_i)$, $i = 1,\ldots,n$, $t \in T$, onde *T* é um intervalo real. Uma descrição formal de dados funcionais será dada a seguir.

Definição 3.1 Uma variável aleatória X é chamada variável funcional (v.f.) se assume valores num espaço de dimensão infinita (ou espaço funcional). Uma observação *x* de X é chamada de dado funcional (Ferraty and Vieu, 2002).

Definição 3.2 Um conjunto de dados funcionais $x_1,\ldots,x_n$ é a observação de *n* variáveis funcionais $X_1,\ldots,X_n$ como *X* (Ferraty and Vieu, 2002).

Existem muitas formas de origem para dados funcionais. Nos casos mais comuns, as observações originais são interpoladas de dados longitudinais ou obtidas através da densidade estimada de observações numéricas independentes para cada indivíduo em estudo. Imagens, bem como outras formas, podem aparecer como dados funcionais. Num exemplo de arqueologia, a forma de uma imagem bidimensional de cada osso é o dado funcional, ou então como curvas traçadas numa superfície ou espaço.

Para que os dados funcionais sejam estudados computacionalmente uma representação adequada está fundamentada no uso da combinação linear de um conjunto de funções-bases previamente escolhidas. A utilização dessas funções bases acarreta vantagens na representação computacional e, assim, dos cálculos que poderão ser executados. Além disso, os dados podem possuir os mais diversos padrões de variação. Neste caso, para proceder a conversão de um conjunto de valores para dados funcionais requer apenas dois passos: o primeiro é escolher e definir um conjunto de funções para formar uma base (funções base) e o outro está em calcular a melhor combinação linear para cada conjunto de valores.

Quando se diz que uma observação é um dado funcional está se referindo a existência de uma função suave que gera os valores daquela observação. A presença desta suavidade é um forte indício para a aplicação da ADF em contraposição as demais técnicas estatísticas como, por exemplo, a análise multivariada clássica. Entretanto, se a função não tem um comportamento suave, realiza-se um processo de suavização para que a ADF possa ser usada.

Mesmo que a função deva apresentar um caráter suave, isto não é uma condição necessária para a observação original $y = (y_1, \ldots, y_n)$, visto que está sujeita a ruídos intrínsecos à observação. De fato, pode-se escrever:

$$y_j = x(t_j) + \varepsilon_j, \tag{3.26}$$

onde *x* é a função suave e ε é o ruído.

Como vimos anteriormente, as funções são representadas por combinações lineares de funções base. Funções ϕ_k conhecidas, independentes entre si e que satisfazem a propriedade em que a combinação linear destas funções é capaz de representar arbitrariamente bem a função original. As funções originais, denotadas por *x*, são representadas então por

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t), \tag{3.27}$$

sendo *K* o número de funções base usadas na representação e c_k números reais chamados de coeficientes da representação. O número *K* de bases determina o grau de suavidade da função obtida. Quanto maior o valor de *K*, mais sua será a função obtida.

O objetivo essencial em ADF é a obtenção de uma função suave com determinadas características semelhantes às encontradas na função original. Não há necessidade de obter

uma equivalência exata entre os valores da função, assim, na prática, é usado um número relativamente pequeno de funções base. Isso tem como vantagem um número maior de graus de liberdade, importante no cálculo de testes de hipóteses e intervalos de confianças, e um menor custo computacional.

A função *x* não é conhecida previamente, assim é necessário fazer aproximações para esta função e uma das formas é por meio da representação de *x* em um subespaço qualquer de funções base. Essa representação de *x* destaca-se pelas seguintes vantagens:

1. Facilidade na manipulação de um conjunto de valores que possui amostragem irregular ou dados faltantes.

2. Possibilidade de se aplicarem operações funcionais como a integração e derivada a estas funções.

3. Processo computacional simples, visto que várias estratégias de análise por ADF usam apenas o vetor de coeficientes.

4. Não há, na teoria, nenhuma restrição quanto ao uso da base, a escolha depende do comportamento do dado original. As bases mais comuns são as de Fourier, *splines*, *wavelets*, logarítmica, exponencial, polinomial.

As funções abordadas em ADF possuem dimensão infinita, pelo fato que *a priori* os valores não podem ser definidos previamente e deste modo é necessário que todos os valores da função, para todo *t*, sejam conhecidos, obtendo um conjunto infinito de valores. Entretanto, na prática a utilização de um conjunto limitado de valores é suficiente para caracterizar a função acrescida de um erro. Assim, para cada caso temos um valor ideal para a dimensão de cada função.

3.3. Algumas propriedades dos dados funcionais

A ideia básica da Análise de Dados Funcionais é considerar que as funções dos dados observados são como entidades individuais, ou seja, cada entidade é representada por uma curva. As relações da literatura clássica que existem para as observações escalares ou vetoriais são substituídas pelas relações entre as curvas ou funções, e são objeto de investigação.

O termo funcional se refere à estrutura intrínseca dos dados ao invés de sua forma explícita, apesar de, na prática, os dados funcionais serem geralmente observados e registrados discretamente. Cada observação funcional *x* consiste de *n* pares (t_j, y_j), em que y_j corresponde ao valor da função *x* em t_j, possivelmente acrescido de um erro de medição.

3.3.1. O que torna os dados discretos em dados funcionais?

O que significa para uma observação funcional conhecer a forma funcional de *x*? Não há a necessidade de *x* ser gravado para cada valor de *t*, pois isso, além de envolver um alto custo, envolveria um número incontável de valores! Pelo contrário, assume-se a existência de uma função *x*, baseada nos dados observados, implicando, a princípio, que é possível avaliar *x* em qualquer ponto *t*, e avaliar também qualquer uma das suas derivadas $D^m x$ que existam em *t*. Mas, se apenas valores discretos estão disponíveis, avaliar $x(t)$ e $D^m x(t)$ em qualquer valor arbitrário *t* envolverá métodos de suavização.

Geralmente, a função *x* é suave, de modo que um par de valores de dados adjacentes, y_j e y_{j+1} são necessariamente ligados de alguma forma e dificilmente serão muito diferentes. Se essa propriedade de suavidade não se aplica, os dados não seriam tratados de uma forma funcional e sim através de técnicas multivariadas.

Pela suavidade, costuma-se dizer que a função *x* possua uma ou mais derivadas. Normalmente, utilizam-se os dados discretos y_j, $j = 1,\ldots,n$, para estimar a função *x* e, ao mesmo tempo, certo número de suas derivadas. Por exemplo, se formos seguir a altura *x* em função do tempo *t* que é descrito por um objeto que se move como um foguete, nós queremos, também, estimar a sua velocidade (Dx) e sua aceleração ($D^2 x$).

Existem várias técnicas na literatura para converter dados brutos em dados funcionais. Portanto, como os dados podem ser considerados como soma da forma funcional com algum ruído, a representação funcional envolve suavização, que é aplicável não apenas às curvas em si, como também às suas derivadas.

3.3.2. Amostras de dados funcionais

O problema da ADF está definido em torno de uma amostra de dados funcionais. Especificamente, o registro ou a observação da função x_i pode consistir de n_i pares $\left(t_{ij}, y_{ij}\right)$, $j = 1,\ldots,n_i$. Os valores de t_{ij} e do intervalo *T*, durante o qual os dados são coletados, podem ser os mesmos para cada registro, mas também podem variar de registro para outro. Sempre é assumido que a amplitude dos valores de interesse para o argumento *t* é um intervalo limitado *T* e que *x* satisfaz condições de continuidade e suavidade em *T*. Sem essas condições, é impossível fazer qualquer inferência sobre os valores $x(t)$ em qualquer ponto *t*, com exceção dos atuais pontos de observação.

Normalmente, a construção da observação funcional x_i usando um dado discreto y_{ij} ocorre separadamente ou de forma independente para cada registro *i*. Por simplicidade, será considerado que uma única função *x* é observada.

Enfim, uma grande quantidade de dados funcionais é distribuída por domínios de argumentos multidimensionais. Podem-se ter dados observados ao longo de uma ou mais dimensões do espaço, bem como ao longo do tempo, por exemplo. Uma fotografia ou uma imagem do cérebro é uma observação funcional, onde a composição da cor e, eventualmente, a intensidade é uma função da localização espacial.

3.4. Estatísticas descritivas para dados funcionais

3.4.1. Média e variância funcionais

As estatísticas descritivas básicas que conhecemos para os dados univariados são aplicadas igualmente para os dados funcionais.

A função média para uma amostra de *n* dados funcionais (ou seja, *n* funções) é dada por

$$\overline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) \,. \tag{3.28}$$

Da mesma forma, a função variância é definida por

$$Var_X(t) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t) - \overline{x}(t)\right)^2, \tag{3.29}$$

enquanto a função desvio padrão é a raiz quadrada da função variância.

3.4.2. Funções de covariância e correlação

A função covariância resume a dependência das observações ao longo de valores de argumentos diferentes e é calculada para todo t_1 e t_2 através da fórmula

$$Cov_X(t_1,t_2) = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t_1) - \overline{x}(t_1)\right]\left[x_i(t_2) - \overline{x}(t_2)\right]. \tag{3.30}$$

A função de correlação associada é dada por

$$Corr_X(t_1,t_2) = \frac{Cov_X(t_1,t_2)}{\sqrt{Var_X(t_1)Var_X(t_2)}}. \tag{3.31}$$

De forma análoga a análise multivariada, podem ser obtidas as matrizes de correlação, variância e covariância.

3.4.3. Funções de covariância e correlação cruzadas

Em geral, quando são observados pares de funções (x_i, y_i), o modo em que estas dependem uma da outra pode ser quantificado pela função de covariância cruzada

$$Cov_{X,Y}(t_1,t_2) = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t_1) - \overline{x}(t_1)\right]\left[y_i(t_2) - \overline{y}(t_2)\right]. \tag{3.32}$$

ou pela função de correlação cruzada

$$Corr_{X,Y}(t_1,t_2) = \frac{Cov_{X,Y}(t_1,t_2)}{\sqrt{Var_X(t_1)Var_Y(t_2)}}. \tag{3.33}$$

4. Aplicação a dados reais

Estudos sobre o crescimento humano, ao contrário do que muitos imaginam à primeira vista, não constituem um processo simples de ser analisado; pode-se ter uma idéia disso nas próprias experiências pessoais de crescimento! Esses estudos trabalham cuidadosamente com o padrão de crescimento durante a infância e adolescência ao longo de décadas. Um exemplo típico de dados referentes ao crescimento é mostrado na Figura 1.

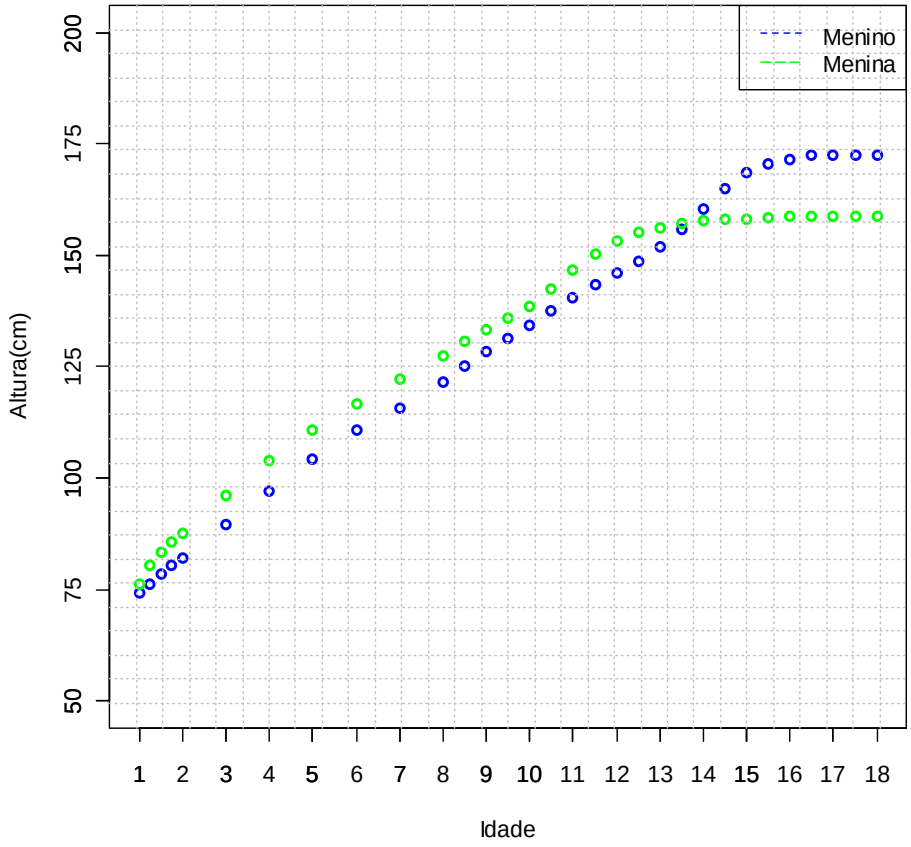

Figura 1: Dados brutos para dois determinados indivíduos em um estudo de crescimento.

A coleta de dados como estes é demorada e demanda um alto custo, porque as crianças têm de ser medidas com precisão e monitoradas por um longo período de suas vidas. Uma vez que as crianças devem ser levadas para o laboratório em idades pré-definidas, elas passam por uma bateria de medições durante um período de cerca de 20 anos, acarretando assim um alto custo para coletar esse tipo de dado. Este regime de observação exige muita dedicação e

persistência dos pais e a taxa de abandono é compreensivelmente elevada. Além disso, requer um treinamento considerável para obter uma medida exata da altura, mas ainda assim é muito complicado. Os procedimentos mais cuidadosos exibem desvios padrões sobre as medidas repetidas de cerca de três milímetros (Ramsay and Silverman, 2005).

Os registros da altura de uma criança durante 20 anos mostram certas características que são difíceis para um analista modelar. A abordagem clássica tem usado funções matemáticas que dependem de um número limitado de constantes desconhecidas. Segundo Ramsay and Silverman (2005), os melhores modelos paramétricos apresentam muitos parâmetros e, ainda assim, alguns aspectos do crescimento não são detectados.

Nesse capítulo, vamos trabalhar com o desenvolvimento na Análise Descritiva de Dados Funcionais sobre o crescimento humano. Será dada uma visão introdutória a fim de apresentar uma idéia básica de uma aplicação de dados reais. Esses dados correspondem a uma parte dos dados do Estudo de Crescimento de Berkeley (Tuddenham and Snyder, 1954). Eles estão publicados e, portanto, disponíveis gratuitamente. Todas as análises foram realizadas no *software* R versão 2.11.1, com auxílio do pacote *fda*.

4.1. Os dados funcionais propriamente ditos

A documentação cuidadosa do crescimento humano é essencial para definir o que chamamos de crescimento normal, para que, através desse estudo, possamos detectar, o mais cedo possível, quando algo está errado com o processo de crescimento. Os cientistas envolvidos no estudo do crescimento precisam de dados de alta qualidade para melhorar o entendimento sobre como o corpo regula o seu próprio crescimento. Assim, para auxiliar esses estudos de uma maneira mais dinâmica, aplica-se a Análise de Dados Funcionais as alturas relativas a cada criança e/ou adolescente.

Para exemplificar este estudo, foram analisadas as alturas medidas em 31 idades que variam entre 1 e 18 anos, separadas por sexo. As idades desempenham um papel claro na nossa análise, porque elas não são igualmente espaçadas. Um objetivo pode ser separar a variação do tempo dos eventos mais significantes de crescimento da variação da intensidade de crescimento.

A Figura 2 mostra os dados brutos em dois cenários distintos. Enquanto que os gráficos da esquerda correspondem aos dados originais das 54 meninas e 39 meninos (respectivamente, nessa ordem) analisados no estudo, os da direita são referentes as 10 primeiras curvas dessas observações, tanto das meninas como dos meninos.

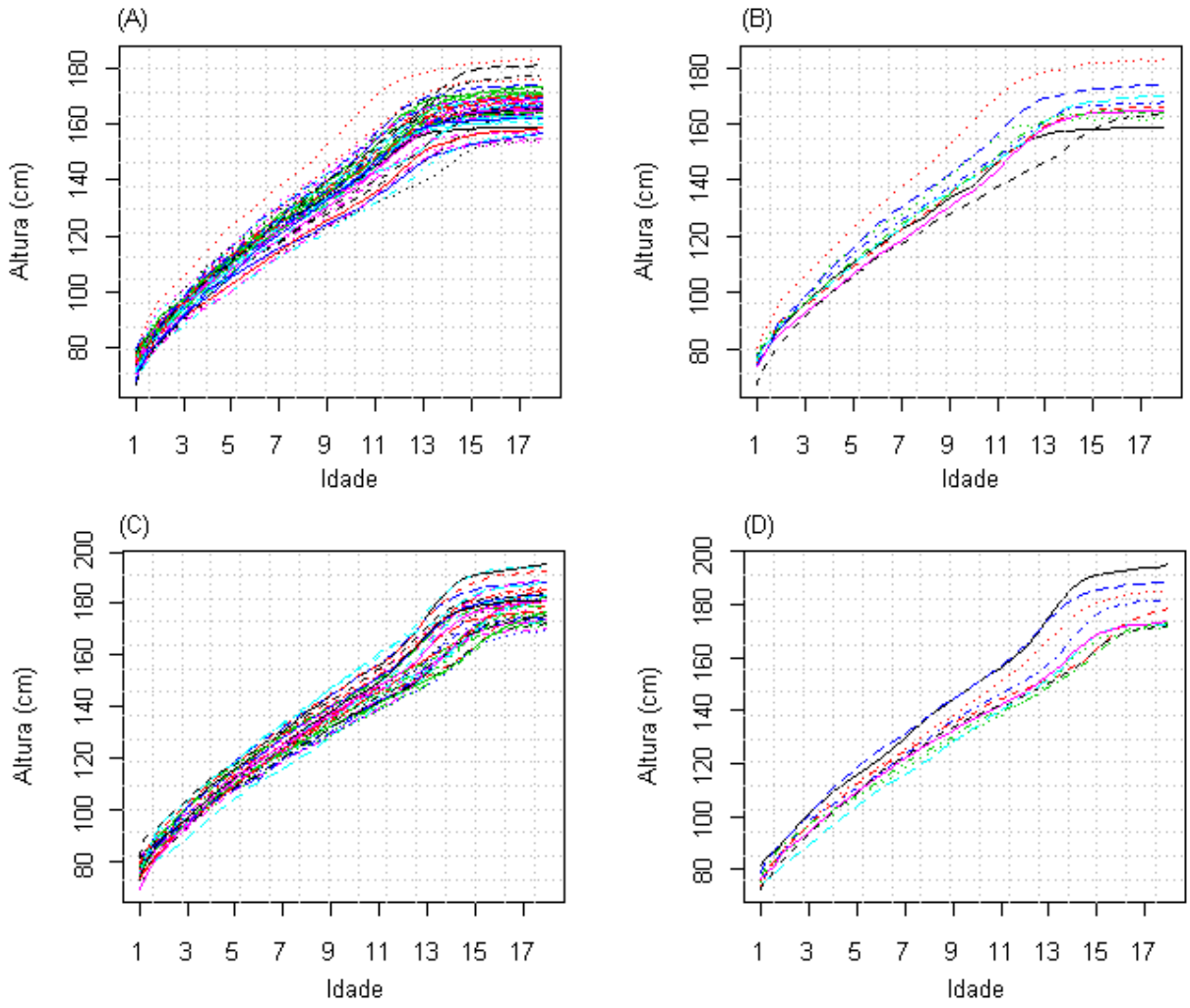

Figura 2: Dados brutos do Estudo de Crescimento de Berkeley. (A) e (B) são as curvas que correspondem as alturas (em cm) das meninas; (C) e (D) as alturas dos meninos.

Consideramos como fazer esse tipo de registro em um dado funcional para incorporar análises posteriores. Uma curva regular que passa pelos pontos na figura é comumente chamada de curva de crescimento, mas o crescimento é realmente a "taxa de aumento" da altura da criança. Nas crianças, esta taxa é necessariamente positiva, pois só muito mais tarde na vida que as pessoas começam a perder estatura.

4.2. A suavização dos dados

O objetivo essencial e um dos aspectos chave em ADF é a obtenção de uma função suave com determinadas características semelhantes às encontradas na função original. Assim, a presença desta suavidade é uma das principais características para a aplicação da ADF. Se as funções não são suaves é necessário que os dados sejam submetidos a um método de suavização antes que a ADF seja aplicada. Esses métodos de suavização não são aplicáveis apenas às curvas em si como também às suas derivadas. Aos dados do estudo de crescimento de Berkeley, aplicamos o método de suavização por *splines*.

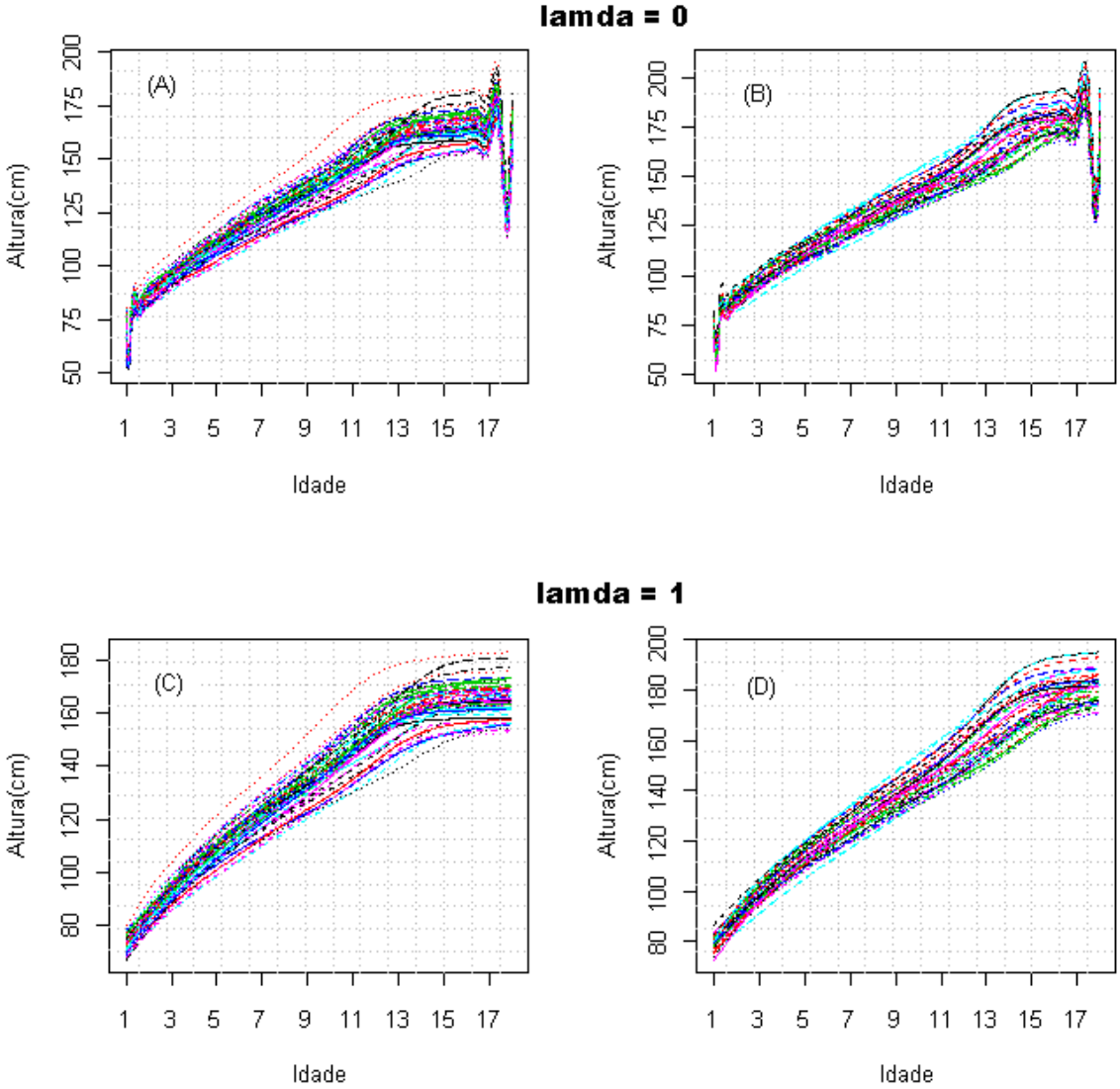

Figura 3: Suavização *spline* dos dados brutos do Estudo de Crescimento de Berkeley.

Na Figura 3, fazemos uma comparação da suavização *spline* com dois distintos parâmetros de suavização λ. Os gráficos (A) e (C) correspondem aos dados das meninas e os

gráficos (B) e (D) aos dados dos meninos. Podemos notar que, tanto no exemplo das meninas como também dos meninos, a suavização *spline* apresenta um comportamento estranho nas suas extremidades quando o parâmetro de suavização é zero. Quando analisamos os mesmos dados, mas com $\lambda = 1$, observamos um comportamento mais suave e tende a se aproximar dos dados originais.

A Figura 4, com o intuito de proporcionar uma melhor visualização, representa graficamente os dados funcionais das dez primeiras meninas do Estudo do Crescimento de Berkeley. Entre as dez curvas ou dados funcionais, podemos destacar algumas informações. Observa-se que o crescimento é mais rápido nos primeiros anos de vida, mas nota-se o aumento na inclinação durante o estirão de crescimento pubertal, que ocorre em idades variando de aproximadamente 9 aos 13 anos. Uma garota é alta em todas as idades, mas algumas meninas podem ser altas durante a infância e terminarem com uma estatura adulta pequena. Os intervalos entre as medidas são de 6 meses ou mais, e através dessa perspectiva a longo prazo tem-se a impressão de um processo de crescimento suave.

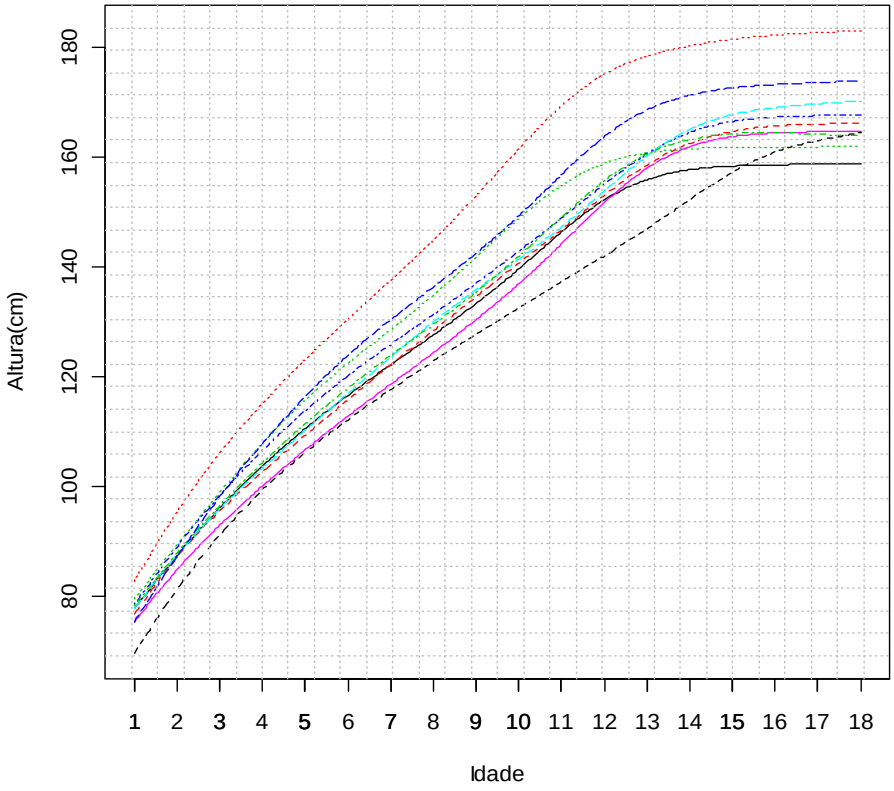

Figura 4: Dados funcionais referentes as alturas das 10 primeiras meninas do Estudo de Crescimento de Berkeley.

Após obter os dados funcionais uma questão importante é a distorção de tempo ou de registro. Deve ser levado em conta em análises e conclusões futuras que nem todas as crianças passam por eventos, tais como a puberdade, na mesma idade. Somente quando são submetidas todas as crianças com um relógio biológico comum, então se pode falar de um padrão de crescimento médio ou investigar a variabilidade da amostra.

4.3. A velocidade e a aceleração do crescimento

Embora os dados e as curvas mostrados anteriormente sejam comumente referidos como "curvas de crescimento", o termo crescimento, na verdade, significa mudança. Então, é a função velocidade $V(t)$, a taxa instantânea de mudança na altura no tempo *t*, que é a real curva de crescimento.

Assim, o termo "crescimento" deve ser usado quando se trata de $V(t)$. Devido à altura não decrescer (pelo menos durante os anos de crescimento iniciais), a velocidade ou crescimento é necessariamente positiva. Os dados de altura refletem o crescimento somente de forma indireta, porque eles são medidas das consequências do crescimento.

Se as observações são tomadas em unidades do tempo t_i, pode-se considerar a estimação da velocidade pela razão de diferenças

$$V(t_i) = \left[H(t_{i+1}) - H(t_i) \right] / (t_{i+1} - t_i) \ . \tag{4.1}$$

Entretanto, isso não é uma boa ideia do ponto de vista estatístico, uma vez que uma pequena quantidade de ruído nas medidas de altura terá um grande efeito na razão, e esse problema torna-se pior conforme os pontos de tempo ficam próximos. É melhor ajustar os dados de altura com uma curva suave apropriada e, então, estimar a velocidade (Ramsay and Silverman, 2002).

Pode-se obter mais informação do processo de crescimento estudando a taxa de mudança na velocidade, ou seja, a aceleração, denotada pela segunda derivada. Os dados transformados – aplicados a 2ª derivada – representam a aceleração estimada da altura.

A Figura 5(A) mostra as curvas estimadas da velocidade para as dez primeiras garotas no Estudo de Crescimento de Berkeley. Agora é possível ver mais claramente o que está acontecendo. O estirão de crescimento na Figura 5(A) é certamente mais óbvio do que, por exemplo, na Figura 4. Nota-se que por volta dos nove anos a velocidade de crescimento começa a aumentar consideravelmente. Agora também sabe-se que é necessário trabalhar muito para encontrar bons métodos para estimar a velocidade.

As curvas de aceleração estimadas para as dez garotas do Estudo de Crescimento de Berkeley podem ser observadas na Figura 5(B) juntamente com a curva média. Agora, pode-se ver ainda mais claramente o que acontece no estirão de crescimento pubertal. Observa-se um grande aumento na aceleração no início do estirão de crescimento pubertal, o que era esperado, seguido por um retorno a zero quando a velocidade não é mais crescente, e finalmente a aceleração se torna negativa na fase final aumentando até chegar em torno de zero novamente quanto a altura se estabiliza.

Outras conclusões da análise gráfica podem ser afirmadas facilmente. É possível observar que o momento do estirão de crescimento pubertal varia bastante de garota para garota. Pode-se também notar que existe uma ou mais oscilações na aceleração antes do estirão de crescimento pubertal. A capacidade de detectar essas oscilações foi um dos importantes avanços recentes da tecnologia não paramétrica de estimação nessa área.

Os resultados são bastante reveladores, demonstrando que o crescimento não ocorre sem problemas, mas é composto de curtos períodos de rápido crescimento entremeados por períodos de relativa estabilidade. O tamanho e o espaçamento desses “saltos” podem ser muito curtos, especialmente em bebês, onde os nossos resultados sugerem que os ciclos de crescimento do comprimento de apenas alguns dias.

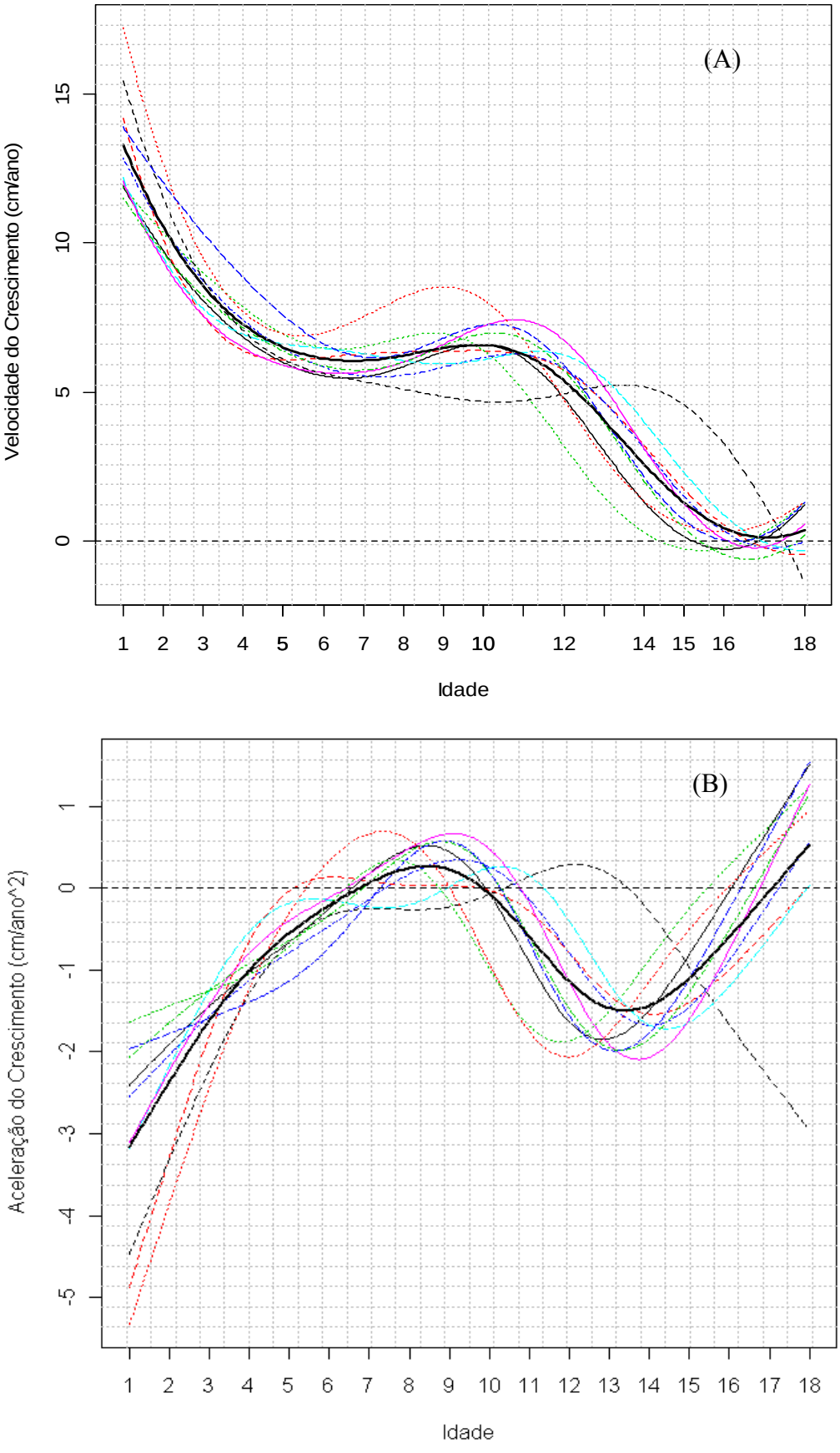

Figura 5: Curvas estimadas de velocidade (A) e de aceleração (B) do crescimento para as 10 primeiras garotas.

4.4. Estatísticas descritivas dos dados funcionais

Nessa seção, vamos realizar uma comparação entre os dados de crescimento do estudo de Berkeley dos meninos e das meninas. Uma forma básica é através das estatísticas descritivas desses dados.

Na Figura 6, em que mostramos as funções média dos dados funcionais das meninas e dos meninos do Estudo de Crescimento de Berkeley, observa-se que por volta dos 13 anos é o período em que os meninos começam a ter um crescimento maior do que as meninas, estas por sua vez estão quase com um crescimento constante. Uma das principais causas disso deve ser a puberdade, período em que ocorrem mudanças biológicas e fisiológicas no corpo de meninos e meninas. Ao ponto de que a função média dos dados funcionais dos meninos se afasta da função média dos dados funcionais das meninas durante a idade em que eles começam a entrar na puberdade, que inicia por volta dos 11 ou 12 anos.

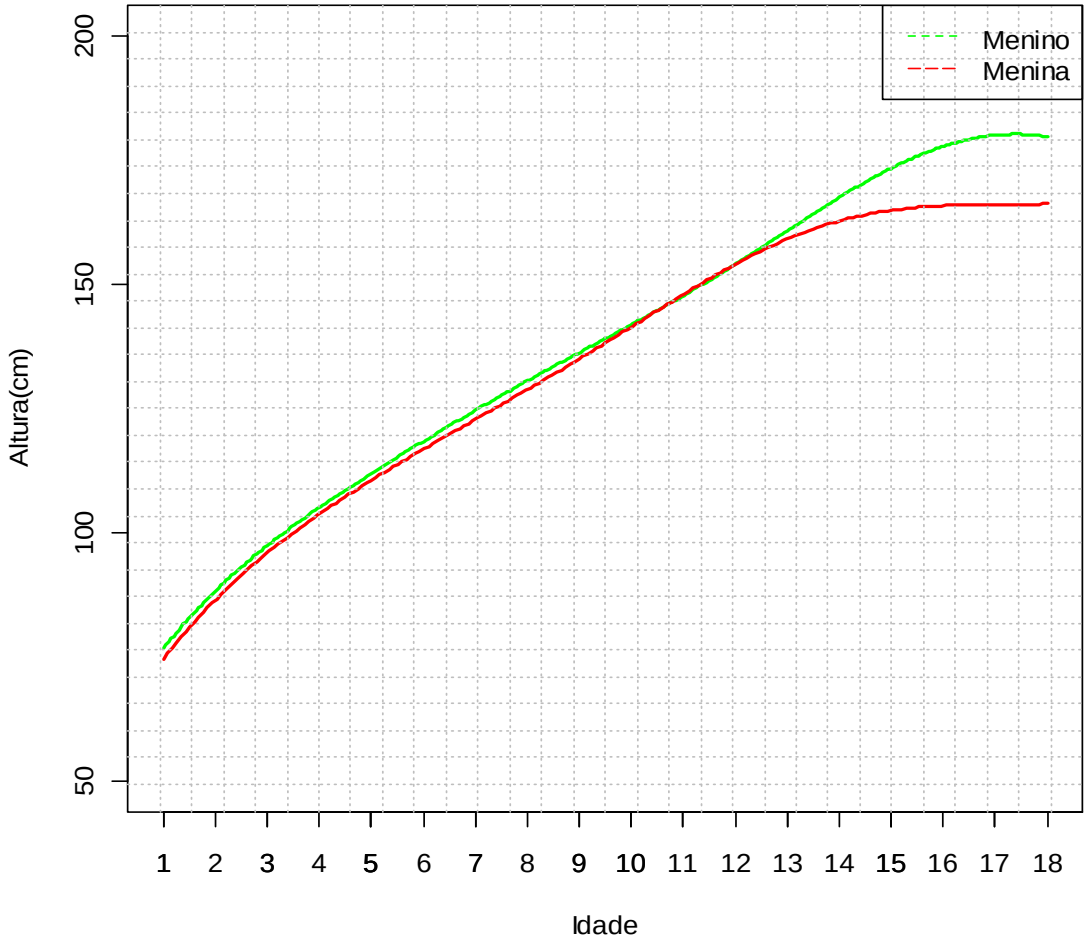

Figura 6: Função média dos dados funcionais do Estudo de Crescimento de Berkeley

Outra forma de visualizar as variações dos dados funcionais entre meninos e meninas é através da função desvio padrão, que nada mais é do que a raiz quadrada da função variância. Ao plotar os pontos dessa função nos gráficos (veja na Figura 7), à primeira vista parece relativamente semelhantes em ambos os sexos. Mas quando paramos para analisar mais detalhadamente, nota-se que a variabilidade dos dados funcionais da altura dos meninos possui um pico próximo a oito centímetros, enquanto que esse pico no gráfico de desvios das meninas é menor. Ao fazer uma média aritmética dos valores da função desvio padrão, encontramos uma variabilidade "média" de 5.507 para os dados funcionais das alturas dos meninos e de 5.469 para os dados das meninas.

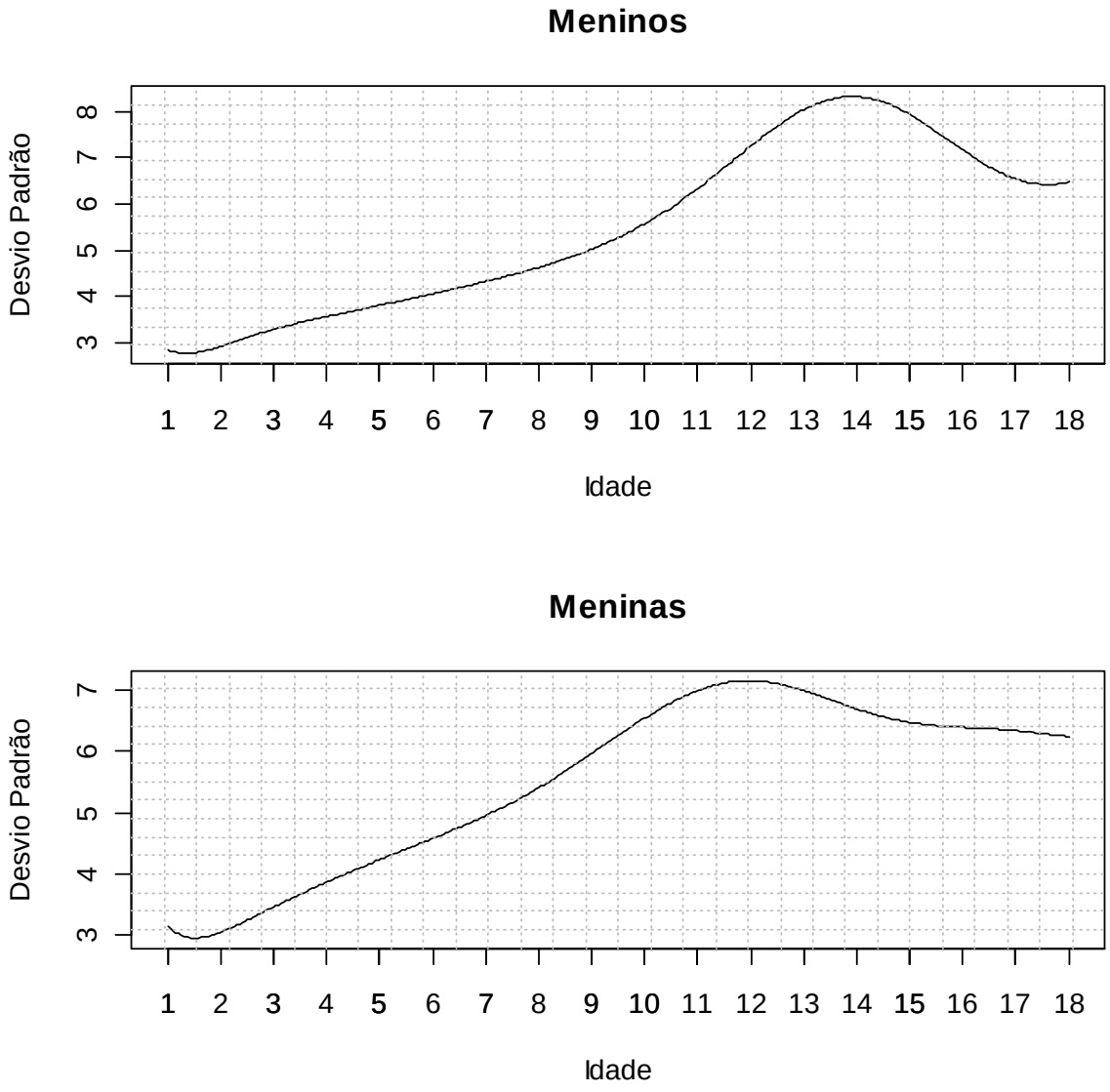

Figura 7: Função desvio padrão dos dados funcionais do Estudo de Crescimento de Berkeley

Na realidade, para termos uma idéia de variabilidade de nossos dados devemos avaliar o gráfico das funções médias desses dados funcionais, com suas respectivas funções desvios

padrões, como foi realizado na Figura 8. Veja que as curvas sólidas coloridas são as funções médias dos dados funcionais referentes às alturas dos meninos e das meninas do Estudo de Crescimento de Berkeley. Enquanto que as curvas pontilhadas representados dois desvios para cima e para baixo relativos àquelas funções médias, ou seja, entre essas curvas pontilhadas estão 95,44% da função média populacional. Uma conclusão sobre esses gráficos parte, novamente da ideia de puberdade. Nas proximidades da idade que inicia o período de puberdade, tanto para as curvas das meninas como para os meninos, a linha dos dois desvios está mais longe da função média, ou seja, a variabilidade é maior nesse local.

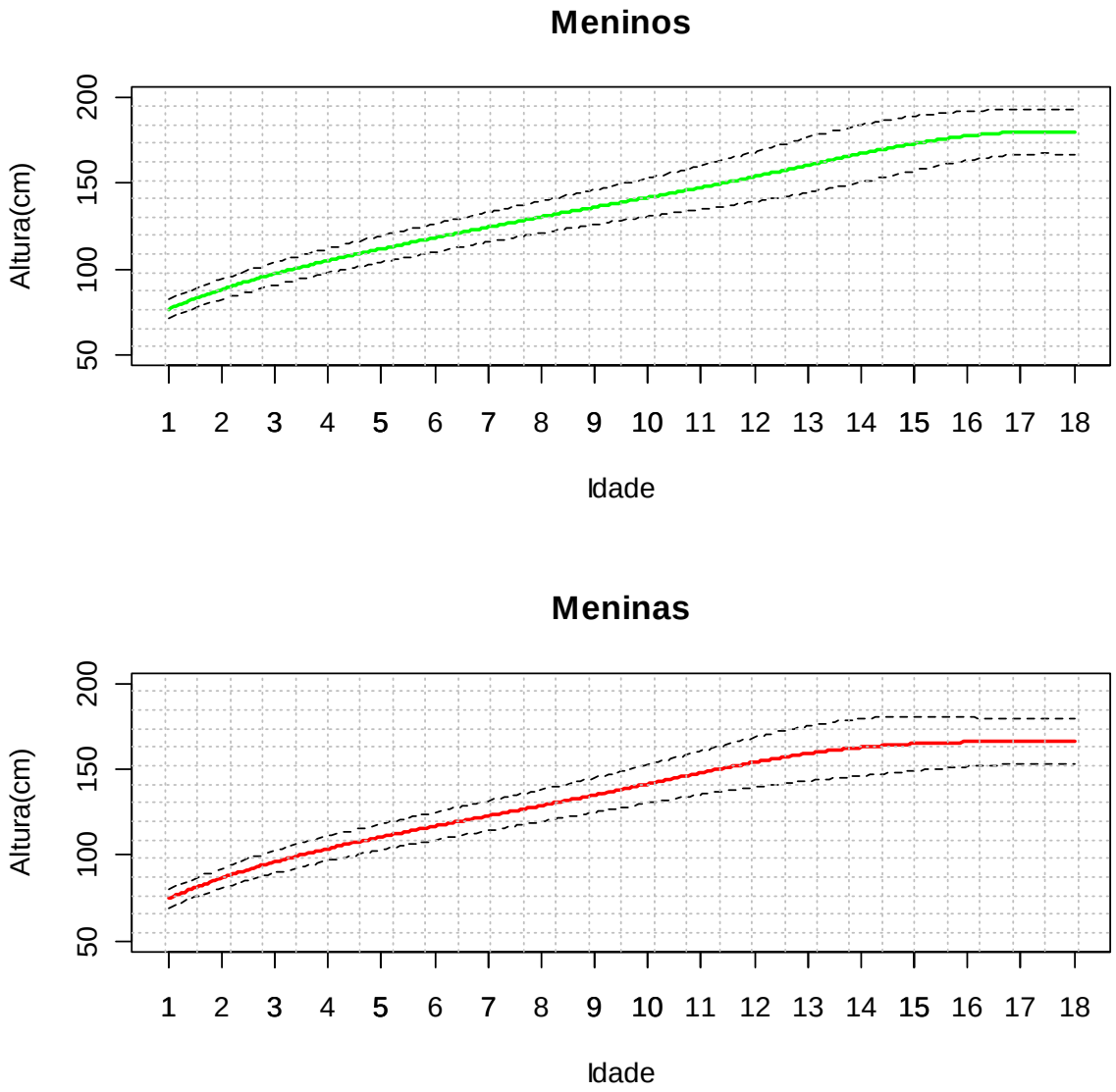

Figura 8: As curvas sólidas representam as funções médias dos dados funcionais. As curvas pontilhadas representam os dois desvios da função média.

Para mais detalhes de aplicações a dados, veja Ramsay and Silverman (2002) e Ramsay, Hooker and Spencer (2009). Este último possui comandos dos *softwares* R e Matlab.

5. Conclusão

Ao longo deste trabalho, apresentamos os fundamentos da Análise de Dados Funcionais (ADF), que consiste em representar um conjunto de observações através de funções. Visando contornar as limitações das técnicas quando os dados são representados de sua maneira mais usual, a ADF surgiu para: i) ser uma alternativa para problemas desse tipo, ii) buscar a informação em diversas fontes pouco exploradas pela análise convencional como, por exemplo, as derivadas das funções e iii) realizar análises que seriam incompatíveis ou que não trariam, aparentemente, nenhum resultado se os dados fossem tratados de uma forma não-funcional.

Por ser ainda um tema recente, a Análise de Dados Funcionais ainda é desconhecida por muitos. Contudo, no meu ponto de vista, a ADF aos poucos vai ganhar seu espaço entre a comunidade estatística mundial. Isso se deve ao aspecto em que ela analisa os dados, retirando informações ocultas através das características possuídas pelas funções.

As principais limitações deste trabalho foram encontradas na literatura. Basicamente, quando se trata de publicações de livros, a literatura da ADF ainda está bastante escassa. Mas, em breve, à medida que for aumentada a divulgação da ADF, as publicações de livros e artigos também deverão crescer. Outra limitação foi a abordagem que foi realizada, uma abordagem basicamente que tratou da Análise Descritiva de Dados Funcionais.

Como trabalhos futuros, o interessante seria divulgar a Análise de Componentes Principais sob o enfoque funcional, técnica esta que é a mais difundida das técnicas da ADF. Outra técnica que poderia ser pesquisada trata-se dos Modelos Lineares Funcionais.

6. Referências Bibliográficas

Bathia, N.; Yao, Q. and Ziegelmann, F. *Identifying the finite dimensionality of curve time series*. To appear. Annal of Statistics.

Craven, P. and Wahba, G. (1979). *Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation.* Numerische Mathematik, 31, p. 377-403.

De Boor, C. (1978). *A Practical Guide to Splines*, New York: Springer-Verlag

De Boor, C. (2001). *A Practical Guide to Splines*. Edição revisada. New York: Springer.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker, 360 p.

Ferraty, F.; Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. New York, N.Y.: Springer, 258 p.

Green, P. J. and Silvermann, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. London: Chapman & Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Hall, P. and C. Vial (2006). Assessing the finite dimensionality of functional data. Journal of the Royal Statistical Society, Series B 68, 689–705.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. New York, N.Y.: Springer, 190 p.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. 1st ed. New York, N.Y.: Springer.

Ramsay, J. O and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. New York, N.Y.: Springer, 426 p.

Ramsay, J. O.; Hooker, G. and Spencer, G. (2009). *Functional Data Analysis with R and MATLAB*. 1nd ed. New York, N.Y.: Springer, 213 p.

Ramsay, J. O. and Dalzell, C. J. (1991). *Some tools for functional data analysis (with Discussion)*. Journal of the Royal Statistical Society, Series B, 53, p. 539-572.

Ramsay, J.O. *et al.* (2010). Pacote *fda* (versão 2.2.1) do software *R*.

Schoenberg, I. (1964a). *Spline functions and the problem of graduation*. Proceedings of the National Academy of Sciences. U.S.A., 52, p. 947-950.

Schoenberg, I. (1964b). *On interpolation by spline functions and its minimum properties*. Int. Ser. Numer. Anal., 5, p. 109-129

Schumaker, L. (1981). *Spline Functions: Basic Theory*. New York: Wiley.

Silverman, B. W. (1984). *Spline smoothing: the equivalent variable kernel method*. The annal of Statistics, 12 (3), p. 898-916.

Tuddenham, R. D. and Snyder, M. M. (1954). *Physical growth of California boys and girls from birth to eighteen years*. University of California Publications in Child Development 1, 183-364.

ANEXOS

ANEXO A – Notações e Espaço de Funções

Este anexo expõe uma breve revisão da notação frequentemente utilizada na literatura. Centrada em Ramsay and Silverman (2005), o livro mais difundido na área de Análise de Dados Funcionais, a notação adotada é bem conhecida na área.

A.1. Notações Elementares

A notação usada apresenta a estrutura do que está sendo discutido no momento. Por exemplo, x pode se referir a um escalar ou a uma função. O contexto, na maioria das vezes, deverá esclarecer situações desse tipo de modo a evitar ambiguidades.

No caso de vetores e matrizes, a notação convencional será mantida. Os vetores denotados por letras minúsculas em negrito, como em $\mathbf{x}$, e as matrizes por letras maiúsculas em negrito, como em $\mathbf{X}$. A notação $\mathbf{x}'$ é usada para a transposta de um vetor $\mathbf{x}$.

A notação para a derivada de ordem m de uma função x é $D^m x$, uma fórmula compacta para enfatizar que a diferenciação é um operador que atua em uma função x para produzir outra função Dx. Naturalmente, $D^0 x$ se refere à própria x e $D^{-1}x$ a integral indefinida de x. No caso das integrais, a notação para a integral definida $\int_a^b x(t)\,dt$ será muitas vezes abreviada para $\int x$ quando o contexto informar os limites de integração (a e b) e a variável t sobre a qual a integração ocorre.

A.2. Espaço de Funções

O espaço onde estão contidas todas as funções de interesse que são consideradas respostas plausíveis dos diversos problemas existentes é chamado de espaço de funções. Esse espaço é o conjunto de todas as funções com quadrado integrável num certo intervalo $[a,b]$ e é denotado por $\mathfrak{L}_2[a,b]$. Em algumas situações não é relevante discriminar o intervalo de integração, por isso usaremos apenas $\mathfrak{L}_2$.

O espaço $\mathfrak{L}_2[a,b]$ é uma coleção muito rica de funções, contudo possui alguns inconvenientes. Por exemplo, para *x* e *y* serem consideradas funções idênticas em $\mathfrak{L}_2[a,b]$, basta que $\|x-y\|=0$, onde $\|x-y\|=\sqrt{\langle x-y,x-y\rangle}$. Consequentemente, avaliar um elemento de $\mathfrak{L}_2[a,b]$ em algum ponto do intervalo $[a,b]$ não é uma operação bem definida como cita Eubank (1988).

ANEXO B – Produto Interno

O produto interno, no nosso caso, entre funções é muito útil para calcular algumas estatísticas descritivas funcionais, como também ajudar na suavização por *splines*.

Inicialmente, será apresentada uma notação genérica para produto interno entre dois vetores **x** e **y**. Considere o produto interno euclidiano $\mathbf{x}'\mathbf{y}$, onde **x** e **y** são vetores de mesma dimensão, o qual possui as seguintes propriedades:

Simetria: $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$ para todo **x** e **y**,

Positividade: $\mathbf{x}'\mathbf{x} > 0$ para todo **x**, com $\mathbf{x}'\mathbf{x} = 0$ se e somente se $\mathbf{x} = 0$, e

Bilinearidade: para todos os números reais *a* e *b*, $(a\mathbf{x} + b\mathbf{y})'\mathbf{z} = a\mathbf{x}'\mathbf{z} + b\mathbf{y}'\mathbf{z}$ para todos os vetores **x**, **y** e **z**.

Essas propriedades seguem da definição

$$\mathbf{x}'\mathbf{y} = \sum_i x_i y_i \,. \tag{7.34}$$

Considere agora que *x* e *y* são funções com valores $x(t)$ e $y(t)$, respectivamente. O funcional equivalente para $x'y$ é $\int x(t)\,y(t)\,dt$, substituindo o somatório em pela integral. Essa substituição conserva as propriedades originais: simétrico em *x* e *y*, linear em cada função e satisfaz a exigência da positividade.

É realizada uma generalização usando uma notação comum para essas operações que implicam em simetria, positividade e bilinearidade. Denominamos esta operação de produto interno, onde $\langle x, y \rangle$ é a notação do produto interno entre *x* e *y*, cujas propriedades são as seguintes:

Simetria: $\langle x, y \rangle = \langle y, x \rangle$ para todo *x* e *y*;

Positividade: $\langle x, x \rangle \geq 0$ para todo *x*, com $\langle x, x \rangle = 0$ se e somente se $x = 0$;

Bilineariidade: para todos os números reais *a* e *b*, $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$ para todos vetores *x*, *y* e *z*.

B.1. Propriedades Gerais

O produto interno pode ser pensado como sendo uma medida escalar de associação entre pares de quantidades *x* e *y*. A simetria dessa medida significa que, como usualmente desejamos, é invariante em relação à ordem das quantidades. A bilinearidade significa que ao mudar a escala de qualquer argumento obtemos o mesmo efeito de escala na medida de associação e ao usar uma soma como um dos argumentos (a soma de duas medidas de associação de uma quantidade é a soma das medidas individuais de associação) deixa a medida de associação inalterada em suas propriedades fundamentais. A positividade significa que o produto interno de qualquer *x* com ele mesmo é uma medida de seu tamanho. A raiz quadrada positiva dessa medida de tamanho é chamada de norma de *x*, $\|x\|$, de modo que

$$\|x\|^2 = \langle x, x \rangle , \tag{7.35}$$

com $\|x\| \geq 0$. No caso especial em que *x* é um vetor $n \times 1$ e o produto interno é o produto interno euclidiano , a norma de *x* é o tamanho do vetor medido no espaço *n*-dimensional. Por exemplo, para a função *f*, um tipo de norma é $\|f\| = \sqrt{\int f^2}$, a qual é denominada de norma L^2.

As propriedades padrão de qualquer produto interno retornam as seguintes propriedades da norma:

1. $\|x\| \geq 0$ e $\|x\| = 0$ se e somente se $x = 0$
2. $\|ax\| = |a| \|x\|$ para todo número real *a*
3. $\|x + y\| \leq \|x\| + \|y\|$.

A relação particular $\langle x, y \rangle = 0$, chamada ortogonalidade, implica que *x* e *y* são perpendiculares entre si. A ortogonalidade desempenha um papel fundamental na operação de projeção.

Do produto interno, deriva-se também uma medida de distância entre *x* e *y*,

$$d_{xy} = \|x - y\| = \sqrt{\langle x - y, x - y \rangle} , \tag{7.36}$$

que possui uma aplicação extremamente ampla.

A estrutura do produto interno depende de algo mais fundamental sobre *x* e *y*. Eles são elementos de um espaço vetorial em que elementos podem ser adicionados, multiplicados, (entre outras operações) por números reais para gerar novos vetores. O conjunto de um espaço vetorial e um produto interno associado é chamado de um espaço de produto interno.

B.2. Outros aspectos dos espaços de produto interno

Sejam $u_1,\ldots,u_n$ quaisquer *n* elementos de um espaço vetorial e seja U o subespaço constituído de todas possíveis combinações lineares dos u_i. Seja *u* o vetor de dimensão *n* cujos elementos são $u_1,\ldots,u_n$. Então todos os membros de *U* são da forma $u'c$ para qualquer vetor *c* de dimensão *n*.

A projeção ortogonal sobre *U* está associada a um subespaço de *U* e é definida como um operador linear *P* que possui as seguintes propriedades:

1. Para todo *z*, o elemento $Pz \in U$ e é uma combinação linear das funções $u_1,\ldots,u_n$.
2. Se *y* já está contido em *U*, então $Py = y$.
3. Para todo *z*, o resíduo $z - Pz$ é ortogonal para todos os elementos υ de *U*.

Das duas primeiras propriedades segue que $PP = P^2 = P$. E da terceira propriedade, é fácil mostrar que o operador *P* mapeia cada elemento de *z* para seu ponto mais próximo em *U*.

ANEXO C – Funções utilizadas do pacote fda

Agora, são apresentadas as funções do *software* R (versão 2.11.1), contidas no pacote *fda*, utilizadas neste trabalho.

as.fd

Transforma um objeto do tipo *spline* (ou seja, um objeto proveniente do processo de suavização via *splines*) para dados funcionais.

create.bspline.basis

Como vimos os dados funcionais são construídos, indicando um conjunto de funções de base e um conjunto de coeficientes que definem uma combinação linear destas funções de base. Assim, a função *create.bspline.basis* tem por característica construir funções de bases B-*splines* para auxiliar na transformação de observações comuns em dados funcionais.

Data2fd

Converte uma matriz *y* de valores da amostra de curvas, acrescido de uma matriz de valores do argumento, para um dado funcional.

fdPar

Transforma um dado funcional, uma função base ou uma matriz em um parâmetro funcional. Parâmetros funcionais são usados como argumentos para funções a fim de realizar estimativas como, por exemplo, em funções de suavização.

smooth.basis

Esta é a principal função usada para a suavização de dados, cujo método é baseado na penalização da rugosidade (das curvas). Esta função controla a natureza e o grau de suavização através dessa penalização.

smooth.basisPar

É uma função de suavização de dados usando o método de penalização da rugosidade previamente especificada através dos argumentos restantes: *fdobj* (basicamente, dados funcionais) e *Lfobj* (um número inteiro positivo).

ANEXO D – Programação da Análise de Dados Reais

A seguir são apresentados os comandos do *software* R (versão 2.11.1) utilizados no Capítulo 5. No *site www.functionaldata.org* (ou instalado a partir do próprio programa) está disponível o pacote de funções *fda* essencial para a Análise de Dados Funcionais, utilizado amplamente no desenvolvimento deste trabalho. Esse pacote contém também alguns bancos de dados, como por exemplo, os dados do Estudo de Crescimento de Berkeley.

Primeiramente, devemos instalar o pacote *fda* e, assim, poderemos prosseguir com a programação.

```
require(fda)  #Carrega o pacote fda

#Figura 1: Dados brutos para dois determinados indivíduos em um estudo de crescimento.
with(growth,matplot(age,hgtf[,1],type="n",xlab="Idade",ylab="Altura(cm)",ylim=c(50,200)))
lines(growth$age,growth$hgtm[,5], type="p",col="blue",lwd=2)
lines(growth$age,growth$hgtf[,1], type="p",col="green",lwd=2)
legend("topright",c("Menino","Menina"), col = c("blue", "green"),lty=c(2, 5))
axis(1,seq(1,21,1))
axis(2,seq(50,200,25))
grid(nx=30,ny=30,col="Gray")


#Figura 2: Dados brutos do Estudo de Crescimento de Berkeley. (A) e (B) são as curvas que
correspondem as alturas (em cm) das meninas; (C) e (D) as alturas dos meninos.
par(mfrow=c(2,2))
with(growth,matplot(age,hgtf,type="l",xlab= "Idade", ylab= "Altura (cm)"))            #(A)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
with(growth,matplot(age,hgtf[,1:10],type="l",xlab= "Idade", ylab= "Altura (cm)"))    #(B)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
with(growth,matplot(age,hgtm,type="l",xlab= "Idade", ylab= "Altura (cm)"))           #(C)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
with(growth,matplot(age,hgtm[,1:10],type="l",xlab= "Idade", ylab= "Altura (cm)"))  #(D)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")

#Figura 3: Suavização spline dos dados brutos do Estudo de Crescimento de Berkeley.
girlGrowthSm<-with(growth,smooth.basisPar(argvals=age,y=hgtf)) #lambda=0
girlGrowth.fd<-as.fd(girlGrowthSm)  #criando dados funcionais
boyGrowthSm<-with(growth,smooth.basisPar(argvals=age,y=hgtm))
boyGrowth.fd<-as.fd(boyGrowthSm)
girlGrowthSm1<-with(growth,smooth.basisPar(argvals=age,y=hgtf,lambda=1)) #lambda=1
girlGrowth.fd1<-as.fd(girlGrowthSm1)
```

```
boyGrowthSm1<-with(growth,smooth.basisPar(argvals=age,y=hgtm,lambda=1))
boyGrowth.fd1<-as.fd(boyGrowthSm1)
par(mfrow=c(2,2))
plot(girlGrowthSm$fd,xlab="Idade",ylab="Altura(cm)")
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(boyGrowthSm$fd,xlab="Idade",ylab="Altura(cm)")
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(girlGrowthSm1$fd,xlab="Idade",ylab="Altura(cm)")
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(boyGrowthSm1$fd,xlab="Idade",ylab="Altura(cm)")
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
```

#**Figura 4:** Dados funcionais referentes as alturas das 10 primeiras meninas do Estudo de Crescimento de Berkeley.

```
girlGrowthSm<-with(growth,smooth.basisPar(argvals=age,y=hgtf, lambda=1))
girlGrowth.fd<-as.fd(girlGrowthSm)
plot(girlGrowthSm$fd[,1:10],xlab="Idade",ylab="Altura(cm)")
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
```

#**Figura 5:** Curvas estimadas de velocidade (A) e de aceleração (B) do crescimento para as 10 primeiras garotas.

```
#Estimando a aceleração e velocidade para as curvas de crescimento
age<-c(seq(1,2,0.25),seq(3,8,1),seq(8.5,18,0.5))
rng<-c(1,18)
knots<-age
norder<-6
nbasis<-length(knots)+norder-2 #escolha de lambda automática
hgtbasis<-create.bspline.basis(rng,nbasis,norder,knots) #criando um base b-spline
Lfdobj<-4
lambda<-df2lambda(age,hgtbasis,df=12)
growfdPar<-fdPar(hgtbasis,Lfdobj,lambda)
```

#Suavizando os dados.Os dados de garotas e garotos são as matrizes hgtm e hgtf, respectivamente.

```
hgtmfd<-smooth.basis(age,growth$hgtm,growfdPar)$fd
hgtffd<-smooth.basis(age,growth$hgtf,growfdPar)$fd
```

```
#Computando as funções de velocidade
velmfd<-deriv.fd(hgtmfd)
velffd<-deriv.fd(hgtffd)
plot(velmfd,xlab="Idade", ylab="Velocidade do Crescimento (cm/ano)") #meninos
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
```

```
plot(velffd,xlab="Idade", ylab=" Velocidade do Crescimento (cm/ano)") #meninas
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(velffd[1:10,1:10],xlab="Idade", ylab=" Velocidade do Crescimento (cm/ano)") #10 meninas
lines(mean(velffd[1:10,1:10]),lwd=2)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")

#Computando as funções de aceleração
accmfd<-deriv.fd(hgtmfd,2)
accffd<-deriv.fd(hgtffd,2)
par(mfrow=c(2,1))
plot(accmfd,xlab="Idade", ylab="Aceleração do Crescimento (cm/ano^2)") #meninos
lines(mean(accmfd),lwd=2)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(accffd,xlab="Idade", ylab="Aceleração do Crescimento (cm/ano^2)") #meninas
lines(mean(accffd),lwd=2)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(accffd[1:10,1:10], xlab="Idade", ylab="Aceleração do Crescimento (cm/ano^2)") #10 meninas
lines(mean(accffd[1:10,1:10]),lwd=2)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")


#Figura 6: Função média dos dados funcionais do Estudo de Crescimento de Berkeley
age<-c(seq(1,2,0.25),seq(3,8,1),seq(8.5,18,0.5))
rng<-c(1,18)
knots<-age
norder<-6
nbasis<-length(knots)+norder-2 #escolha de lambda automática
hgtbasis<-create.bspline.basis(rng,nbasis,norder,knots) #criando um base b-spline
Lfdobj<-4
lambda<-df2lambda(age,hgtbasis,df=12)
growfdPar<-fdPar(hgtbasis,Lfdobj,lambda)
hgtmfd<-smooth.basis(age,growth$hgtm,growfdPar)$fd
hgtffd<-smooth.basis(age,growth$hgtf,growfdPar)$fd
girlsmean<-mean.fd(hgtffd)
boysmean<-mean.fd(hgtmfd)
plot(boysmean,col="green",lwd=2, xlab="Idade",ylab="Altura(cm)", ylim=c(50,200)) #Função média dos meninos
lines(girlsmean,col="red",lwd=2) #Função média dos meninas
legend("topright",c("Menino","Menina"), col = c("green", "red"),lty=c(2, 5))
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
```

#**Figura 7:** Função desvio padrão dos dados funcionais do Estudo de Crescimento de Berkeley

```
age<-c(seq(1,2,0.25),seq(3,8,1),seq(8.5,18,0.5))
rng<-c(1,18)
knots<-age
norder<-6
nbasis<-length(knots)+norder-2 #escolha de lambda automática
hgtbasis<-create.bspline.basis(rng,nbasis,norder,knots) #criando um base b-spline
Lfdobj<-4
lambda<-df2lambda(age,hgtbasis,df=12)
growfdPar<-fdPar(hgtbasis,Lfdobj,lambda)
hgtmfd<-smooth.basis(age,growth$hgtm,growfdPar)$fd
hgtffd<-smooth.basis(age,growth$hgtf,growfdPar)$fd
boys.sd<-sd.fd(hgtmfd)
boys.sd $coefs  #valores da função desvio padrão
girls.sd <-sd.fd(hgtffd)
girls.sd $coefs
par(mfrow=c(2,1))
plot(boys.sd, xlab="Idade",ylab="Desvio Padrão",main="Meninos")
axis(1,seq(1,21,1))
grid(nx=30,ny=15,col="Gray")
plot(girls.sd, xlab="Idade",ylab="Desvio Padrão",main="Meninas")
axis(1,seq(1,21,1))
grid(nx=30,ny=15,col="Gray")
```

#**Figura 8:** As curvas sólidas representam as funções médias dos dados funcionais. As curvas pontilhadas representam os dois desvios da função média.

```
age<-c(seq(1,2,0.25),seq(3,8,1),seq(8.5,18,0.5))
rng<-c(1,18)
knots<-age
norder<-6
nbasis<-length(knots)+norder-2 #escolha de lambda automática
hgtbasis<-create.bspline.basis(rng,nbasis,norder,knots) #criando um base b-spline
Lfdobj<-4
lambda<-df2lambda(age,hgtbasis,df=12)
growfdPar<-fdPar(hgtbasis,Lfdobj,lambda)
hgtmfd<-smooth.basis(age,growth$hgtm,growfdPar)$fd
hgtffd<-smooth.basis(age,growth$hgtf,growfdPar)$fd
girlsmean<-mean.fd(hgtffd)
boysmean<-mean.fd(hgtmfd)
boys.sd<-sd.fd(hgtmfd)
girls.sd <-sd.fd(hgtffd)
x= boysmean+ 2*boys.sd
y= boysmean- 2*boys.sd
z= girlsmean+ 2*boys.sd
w= girlsmean- 2*boys.sd
par(mfrow=c(2,1))
plot(boysmean,col="green",lwd=2,xlab="Idade",ylab="Altura(cm)",ylim=c(50,200),
main="Meninos")
```

```
lines(x, lty=2)
lines(y, lty=2)
axis(1,seq(1,21,1))
grid(nx=30,ny=30,col="Gray")
plot(girlsmean,col="red",lwd=2,xlab="Idade",ylab="Altura(cm)",ylim=c(50,200),
main="Meninas")
lines(z, lty=2)
lines(w, lty=2)
```