Introdução à Análise de Dados Funcionais

Ronaldo Dias
IMECC - UNICAMP

Camila Pedroso Estevam de Souza
Dept. de Estatística - University of Bristish Columbia

Minicurso 19º SINAPE
26 a 30 de julho de 2010
São Pedro - SP - Brasil

Prefácio

Com o desenvolvimento de tecnologias mais modernas, dados funcionais tem sido observados com frequência cada vez maior em diversos campos. Em muitos casos, o interesse esta na estimação não somente das curvas mas também de outros funcionais como por exemplo, derivadas e integrais destas curvas. Considere o seguinte problema: no estudo de crescimento de criança pode-se estar interessado em, além de estimar a curva de crescimento, simultaneamente estimar a velocidade de crescimento ou aceleração como função do tempo para cada indivíduo. Desenvolveu-se, portanto, uma nova metodologia, chamada Análise de Dados Funcionais (*Functional Data Analysis*) para contemplar este tipo de problema. Aqui, o termo funcional se refere à estrutura dos dados e não a sua forma explícita, pois na prática os dados são observados de maneira discreta. Parte substancial das técnicas desenvolvidas para a análise de dados funcionais foi iniciada por Ramsay e Dalzell (1991) e Ramsay e Silverman (1997). As técnicas não paramétricas são particularmente apropriadas para a modelagem de dados funcionais. No caso de estimação "pontual" de curvas médias pode-se utilizar diversas metodologias, entre elas métodos baseados em *kernel* tais como Ferraty e Vieu (2004), Fan e Zhang (2000) e *splines*; Nielson, (1974), Guo (2004) e Ramsay e Dalzell (1991). No caso de estimação pontual de curvas médias e resumo de informação, uma técnica adaptativa é proposta em Anselmo, Dias e Garcia (2005). Assim como metodologias mais avançadas em estimação para dados funcionais agregados podem ser encontradas em Dias, Garcia e Martarelli (2009).

Embora esta área seja cativante, de crescente desenvolvimento e de grande interesse pela comunidade estatística internacional, ainda está incipiente no Brasil, com poucos autores nacionais. O objetivo deste trabalho é expor a

comunidade estatística brasileira a uma introdução de algumas técnicas de Análise de Dados Funcionais, que por nós são consideradas como fundamentais para o entendimento de metodologias mais avançadas nessa área.

Gostaríamos de agradecer nossas famílias pelo apoio e compreensão.

Ronaldo Dias e Camila P. E. de Souza
Campinas, 2010.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 O que são dados funcionais?

Em geral, a análise de dados funcionais (ADF) lida com dados nos quais a i-ésima observação é uma função real, $x_i(t)$, $i = 1, \ldots, n$, $t \in \mathcal{T}$, onde $\mathcal{T}$ é um intervalo real; portanto, cada x_i é um ponto em algum espaço de funções $\mathcal{H}$.

As razões práticas para analisar dados de uma perspectiva funcional estão citadas nos quatro itens abaixo.

- As observações funcionais ocorrem cada vez mais frequentemente em contextos aplicados à medida que as facilidades de coleta automatizada de dados se tornam acessíveis para mais pesquisadores. 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, como seria o caso se um número finito de dados fosse usado para estimar toda uma função, suas derivadas ou valores de outros funcionais.

- Levar em consideração certos aspectos, tais como a suavidade, para da-

dos multivariados ocasionados por processos funcionais pode ter implicações importantes na análise desses dados.

Quais tipos de dados podem ser chamados de dados funcionais? Nos casos mais comuns, as observações originais são interpoladas de dados longitudinais, que são quantidades observadas conforme elas evoluem através do tempo. Contudo, existem muitas outras formas de origem para dados funcionais. Por exemplo, pode-se obter um grande número de observações numéricas independentes para cada indivíduo em estudo. O dado funcional de um indivíduo poderia ser a densidade estimada dessas observações. Às vezes, os dados são curvas traçadas numa superfície ou espaço. Num exemplo de arqueologia, a forma de uma imagem bidimensional de cada osso em estudo é o dado funcional em questão. É importante dizer que imagens, bem como curvas, podem aparecer como dados funcionais ou como parâmetros funcionais em modelos.

1.2 Quão característica é a Análise de Dados Funcionais?

O termo atual *análise de dados funcionais* (ADF) foi criado por Ramsay e Dalzell (1991), embora algumas ideias de certa forma já existissem muito antes. O que tem sido mais característico nas recentes pesquisas é a noção de ADF como uma maneira de pensar, ao invés de um conjunto de métodos e técnicas distintas.

A tentativa de se encontrar uma definição exaustiva de ADF tem sido intencionalmente contida, porque não se deseja colocar grandes limites ao redor desse campo. No entanto, vale a pena notar alguns aspectos comuns dos dados funcionais que aparecem frequentemente na literatura. Nesse caso, é possível citar algumas referências, tais como: Ramsay e Silverman (2002), Hall et al. (2006) e Rice (2004).

- *Conceitualmente, dados funcionais são continuamente definidos.* Claro que na prática eles geralmente são observados em pontos discretos, e também têm sido armazenados de maneira finito-dimensional dentro do computador, mas isso não altera a maneira de pensar destacada acima.

- *O dado individual é toda a função*, ao invés do seu valor em algum ponto em particular. Os diversos dados funcionais irão frequentemente ser independentes uns dos outros, mas não há suposições sobre a independência de diferentes valores dentro do mesmo dado funcional.

- Em alguns casos os dados são funções do tempo, mas *não há nada de especial com o tempo como uma variável*. Em certos estudos os dados são funções de uma variável unidimensional, mas a maioria das ideias leva diretamente à funções de variáveis com dimensões maiores.

- Não há exigência para que os dados sejam suaves, mas *frequentemente a suavidade ou outra regularidade será um aspecto chave da análise.* Em alguns casos, derivadas das funções observadas serão importantes. Em outras ocasiões, embora os dados em si não tenham que ser suaves, suposições de suavidade serão apropriadas para funções médias ou outras funções envolvidas em modelar os dados observados.

1.3 Metas da Análise de Dados Funcionais

As metas da análise de dados funcionais são essencialmente as mesmas que de outros ramos da estatística. Elas incluem os seguintes objetivos:

- 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 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.

Cada uma dessas atividades pode ser conduzida com técnicas apropriadas para cada tipo de problema.

Uma estratégia para analisar dados funcionais é usar três etapas: exploratória, confirmatória e preditiva. Na parte exploratória, as perguntas lançadas aos dados tendem a ser bastante ilimitadas, no sentido de que a técnica correta é esperada tanto para revelar novos e interessantes aspectos dos dados, quanto para mostrar características óbvias. Investigações exploratórias tendem a considerar somente os dados em mãos, com menos preocupação para afirmações sobre questões mais amplas, tais como características de populações ou eventos não observados nos dados. Análises confirmatórias, por outro lado, tendem a ser inferenciais e determinadas por perguntas específicas sobre os dados. Assume-se que algum tipo de estrutura está presente nos dados e deseja-se saber se certas afirmações específicas ou hipóteses podem ser confirmadas pelos dados. A linha que separa as análises exploratórias das confirmatórias tende a ser o grau em que a teoria de probabilidade é usada, no sentido de que a maioria das análises confirmatórias são resumidas por uma ou mais afirmações de probabilidade. Estudos preditivos são menos comum e o foco é usar os dados em mãos para fazer afirmações sobre estados não observados, tal como o futuro.

1.4 ADF versus ADL

Avanços na tecnologia moderna, incluindo ambientes computacionais, têm facilitado a coleta e análise de dados de alta dimensão, ou dados que são formados por medidas repetidas de um mesmo objeto. Se as medidas são tomadas durante um intervalo de tempo $\mathcal{T}$, há geralmente duas abordagens distintas para tratá-las, dependendo se as medidas estão disponíveis em um intervalo denso de pontos, ou se elas estão registradas relativamente de forma esparsa.

Quando os dados são registrados densamente ao longo do tempo, frequentemente por máquinas, eles são tipicamente chamados de dados funcionais, com uma curva (ou função) observada por objeto em estudo. Esse é frequentemente o caso mesmo quando os dados são observados com erro experimental, já que a operação de suavizar dados registrados em pontos próximos do tempo pode reduzir grandemente os efeitos de ruído. Em tais casos, pode-se considerar toda a curva para o i-ésimo objeto sendo observada no contínuo, mesmo que na realidade os tempos de registro sejam discretos. Tal curva é representada pelo gráfico da função $X_i(t)$ (ou $x_i(t)$ para facilitar a notação). A análise

estatística de uma amostra de n curvas como essas é comumente chamada de *análise de dados funcionais*.

Estudos biomédicos longitudinais são semelhantes à ADF em importantes aspectos, exceto que é raro observar toda a curva. As medidas são frequentemente registradas somente em alguns pontos espalhados do tempo, que variam entre os sujeitos em estudo. Pode-se representar os tempos das observações para o sujeito i por variáveis aleatórias T_{ij}, para $j = 1, \ldots, m_i$, sendo m_i o número de observações obtidas para o i-ésimo sujeito em estudo. Dessa forma, os dados resultantes são $(X_i(T_{i1}), \ldots, X_i(T_{im_i}))$, geralmente observados com ruído. O estudo de informações dessa forma é frequentemente referido como *análise de dados longitudinais* (ADL).

As abordagens estatísticas para analisar dados longitudinais e dados funcionais são geralmente distintas. Técnicas paramétricas, tais como equações de estimação generalizadas ou modelos de efeitos mistos lineares generalizados, têm sido os métodos dominantes para dados longitudinais, enquanto que técnicas não paramétricas são tipicamente empregadas para dados funcionais.

Uma diferença significante, intrínseca entre os dois tipos de dados, está na percepção de que dados funcionais são observados no contínuo, sem ruído, enquanto que dados longitudinais são observados em pontos do tempo esparsamente distribuídos e estão frequentemente sujeitos a erro experimental. Contudo, dados funcionais são às vezes analisados depois de suavizar observações que foram feitas em um número relativamente pequeno de pontos, talvez somente uma dúzia de pontos se, por exemplo, curvas de todo o ano são calculadas a partir de medidas mensais. Tais casos indicam que as diferenças entre os dois tipos de dados estão relacionadas com a maneira em que o problema é considerado e elas são possivelmente mais conceituais do que reais.

As metas da ADF e da ADL são também de alguma forma diferentes, o que se deve em parte ao fato de que os assuntos científicos são diferentes. As metas da ADF tendem a ser exploratórias, enquanto que na ADL inferências são realizadas.

Apesar das diferenças citadas acima existem muitas metas em comum entre a ADF e a ADL, sendo possível citar as seguintes:

1. Caracterização da média ou curso de tempo típico.

2. Estimação de curvas individuais de dados com ruído e, geralmente em

estudos de ADL, de dados esparsos. Funcionais dessas curvas, tais como derivadas, posições e valores extremos são às vezes de interesse também.

3. Caracterizar homogeneidade e padrões de variabilidade entre as curvas e identificar curvas que não são úteis.

4. Estimar as relações das formas de curvas com covariáveis.

1.5 Análise de dados funcionais utilizando R

Através do pacote `fda` é possível desenvolver em R inúmeras técnicas e exemplos envolvendo dados funcionais. Ramsay et al. (2009) apresenta uma completa descrição do pacote `fda` não só para R mas também para o programa Matlab.

Ao longo desse livro porções dos códigos em R usados na elaboração dos exemplos estarão descritos utilizando a fonte *typewriter*, como por exemplo, `plot(x,y)`.

1.6 Motivação: Crescimento Humano

1.6.1 Introdução

O estudo sobre o crescimento humano é essencial para definir o que é um crescimento normal. Dessa forma é possível detectar, o mais cedo possível, se algo está errado com o processo de crescimento. Os cientistas precisam de dados de alta qualidade para melhorar o entendimento sobre como o corpo regula seu próprio crescimento.

A coleta de dados sobre o crescimento exige um alto custo e persistência, uma vez que as crianças devem ser levadas ao laboratório em idades pré-definidas ao longo de 20 anos. É difícil também obter a medida exata da altura; isso requer um treinamento considerável. Os procedimentos mais cuidadosos ainda exibem desvios padrão sobre medidas repetidas de aproximadamente 3 milímetros.

Registros da altura de uma criança durante 20 anos mostram certas características que são difíceis para um analista modelar. A abordagem clássica tem sido usar funções matemáticas que dependem de um número limitado de

constantes desconhecidas. Os melhores modelos paramétricos apresentam oito ou mais parâmetros e, ainda assim, alguns aspectos do crescimento não são detectados.

Técnicas não paramétricas, como suavização por *kernel* e *splines*, têm tido sucesso em identificar características não detectadas pelos modelos paramétricos, mas elas não garantem a produção de curvas que são monótonas, estritamente crescentes. Mesmo uma pequena falha na monotonicidade pode produzir sérias consequências na velocidade de crescimento, bem como nas curvas de aceleração, que são especialmente importantes na identificação de processos que regulam o crescimento.

Nessa seção, alguns desenvolvimentos na análise de dados de crescimento podem ser observados. Um método desenvolvido para suavização monótona é aplicado em alguns dados (Figura 1.2). Tal método está bem descrito no Capítulo 5.

Mais detalhes e outros exemplos sobre o crescimento humano podem ser encontrados em Ramsay e Silverman (2002) e no site `www.functionaldata.org`.

1.6.2 Medidas de altura em duas escalas

A Figura 1.1 mostra, para cada uma de 10 garotas, a função altura $H(t)$ estimada a partir de 31 observações tomadas entre 1 e 18 anos. Esses dados correspondem a uma parte dos dados do Estudo de Crescimento de Berkeley. Eles estão publicados e, portanto, disponíveis gratuitamente, inclusive no pacote `fda`. 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 puberal, que ocorre em idades variando de aproximadamente 9 a 15 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. O código a seguir foi utilizado na elaboração desse exemplo.

```
library(fda)
altGarotas <- growth$hgtf[,1:10]
idade <- growth$age
amplitude <- range(idade)
idade2 <- seq(1,18,length=101)
nos <- idade   # em inglês knots
```

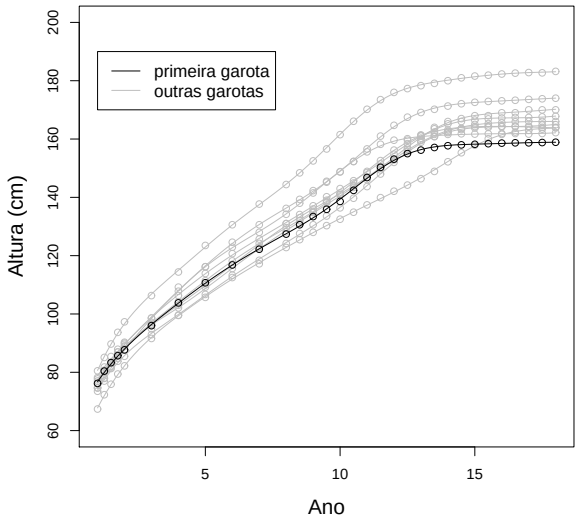


Figura 1.1: Alturas das primeiras 10 garotas do Estudo de Crescimento de Berkeley. Os círculos indicam os dados observados. Cada curva sólida é o ajuste suave aos dados obtido através da penalização da não suavidade usando a quarta derivada.

```
nOrdem <- 6
nBases <- length(nos) + nOrdem - 2
altBases <- create.bspline.basis(amplitude, nBases, nOrdem, nos)
garotasfdPar <- fdPar(altBases, Lfdobj=4, lambda=1e-1)
objFuncGarotas <- smooth.basis(idade, altGarotas, garotasfdPar)$fd
altGarotasSuave <- eval.fd(idade2, objFuncGarotas)
```

A Figura 1.2 apresenta registros da altura de um garoto em 83 dias durante um ano escolar, as falhas correspondem as férias. Esse conjunto de dados também está disponível no pacote `fda`. A curva sólida representa o ajuste suave dos dados obtido através do método de suavização monótona. O ruído de medida nos dados, com desvio padrão de aproximadamente 3 mm, é evidente. O ajuste apresenta mais protuberâncias, com a altura aumentando mais rapidamente durante algumas semanas do que em outras.

```
library(fda)
```

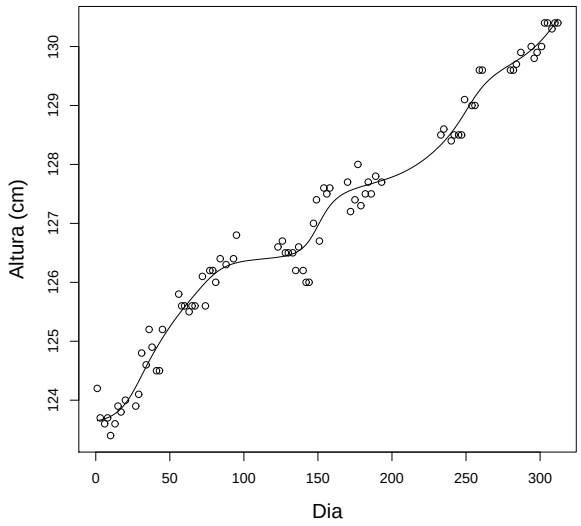


Figura 1.2: Os círculos são 83 medidas da altura de um garoto de 10 anos durante o ano escolar. A curva sólida é o ajuste suave obtido por suavização monótona dos dados.

```
altGaroto <- onechild$height ; dia <- onechild$day
dia2 <- seq(1,max(dia),length=151)
nBases <- 43
altBases <- create.bspline.basis(range=range(dia), nbasis=nBases,
norder=4)
cvec0 <- matrix(0,nBases,1) ; Wfd0  <- fd(cvec0, altBases)
garotofdPar <- fdPar(Wfd0, Lfdobj=2, lambda=1)
resultado <- smooth.monotone(dia, altGaroto,garotofdPar,
 active=c(TRUE,rep(TRUE,nBases-1)))
objFuncGaroto  <- resultado$Wfdobj
beta <- resultado$beta
altGarotoSuave <- beta[1] + beta[2]*eval.monfd(dia2, objFuncGaroto)
```

1.6.3 Das observações para dados funcionais

Considere as medidas de altura das garotas presentes na Figura 1.1. Cada garota foi medida em 31 idades que variam entre 1 e 18 anos.

Uma função pode ser expressa como uma soma ponderada ou combinação linear de blocos de construção funcional elementares chamados funções base. Portanto, a conversão de dados do crescimento para a forma funcional requer dois passos:

1. Escolher e definir um conjunto de funções para formar uma base. Por simplicidade, tais funções são chamadas de funções base.

2. Calcular a melhor combinação linear para cada conjunto de medidas de altura discretas.

Nesse caso, foram escolhidas as funções base B-*spline* como o sistema de funções base a ser utilizado.

As funções B-*spline* são blocos de construção extremamente flexíveis para ajustar curvas. Deseja-se também observar a velocidade e a aceleração do crescimento e isso requer o controle da suavidade das funções base. Isso pode ser feito facilmente com *splines*.

Por outro lado, as séries de Fourier são frequentemente úteis, mas somente para dados que são fortemente periódicos, e que, portanto, mostram claramente ciclos repetitivos. Os dados de crescimento não são assim.

É importante dizer para considerações futuras que os métodos de ajuste descritos aqui não reconhecem que a altura cresce, e que as curvas devem em princípio ser sempre crescentes.

Como os B-*splines* são segmentos polinomiais unidos suavemente duas características devem ser especificadas. Primeiro, é necessário escolher o grau dos segmentos polinomiais, ou, equivalentemente, a ordem (ordem=grau+1). A segunda característica é um conjunto de pontos chamados nós (em inglês, *knots*) nos quais esses polinômios se juntam. A sequência de nós deve ser crescente, ou não decrescente no caso de múltiplos nós. O primeiro nó deve estar na menor idade ou abaixo dela e o último deve estar na maior idade ou acima dela. O número de nós entre o primeiro e o último mais a ordem dos segmentos polinomiais determinam o número total de funções base.

Mais detalhes sobre as funções B-*spline* estão descritos na Seção 3.2.7.

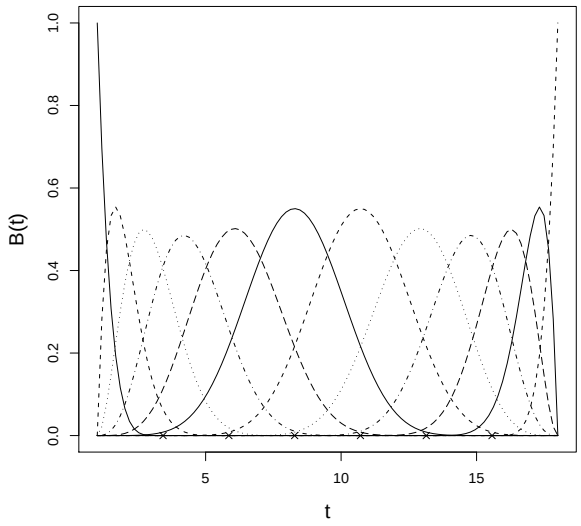
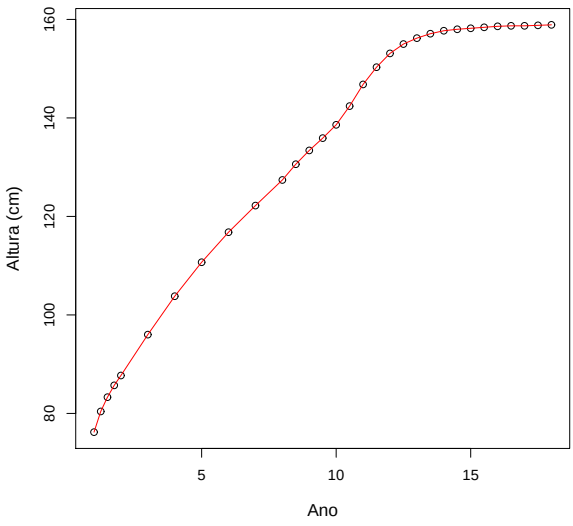


Figura 1.3: B-*splines* de sexta ordem, com seis nós interiores igualmente espaçados.

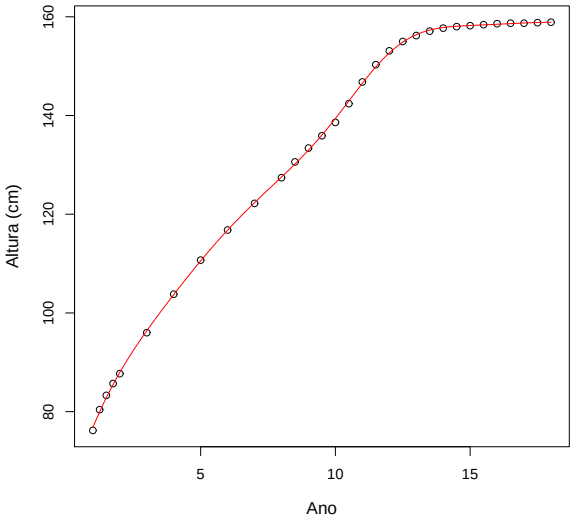
Na Figura 1.3 pode-se observar as funções B-*splines*; a ordem dos polinômios é 6 e o número de nós interiores também é 6. Portanto, o número de funções base é 12 ($6 + 6 = 12$).

```
t <- seq(1,18,length=101)
bases <- create.bspline.basis(rangeval=c(1,18), nbasis=12, norder=6)
nos <- knots.basisfd(bases)
matrizBases <- getbasismatrix(t, bases, nderiv=0)
matplot(t,matrizBases,ylab='B(t)',xlab='t',type='l',cex.lab=1.4)
points(x=nos,y=matrix(0,length(nos), 1),pch=4)
```

Nesse exemplo sobre o crescimento das garotas a ordem também é 6, ou seja, os B-*splines* são segmentos polinomiais de quinto grau. Em um primeiro caso, o número total de funções base é igual ao número de idades em que as observações foram feitas, ou seja, 31. A curva estimada é obtida calculando a combinação linear dessas funções base que melhor ajusta os dados para cada garota. O resultado pode ser visto na Figura 1.4-(a), onde é possível observar



(a)



(b)

Figura 1.4: (a) Curva ajustada para a primeira garota encontrando a melhor combinação linear de 31 B-*splines*. (b) Curva ajustada para a primeira garota penalizando a não suavidade usando a quarta derivada.

a curva obtida para a primeira garota do conjunto de dados. Nota-se que a curva se ajustou exatamente nas observações. Sabe-se que há um erro de medida das alturas de aproximadamente 3 mm, portanto não seria melhor tentar encontrar uma curva que se aproximasse da verdade do que interpolar os dados com ruído? Uma maneira simples de controlar a suavidade do ajuste é limitar o número de funções base.

Por outro lado, assumindo que o processo que gerou os dados é suave, pode-se fazer algo muito melhor suavizando a curva ajustada. Nesse segundo caso, as próprias idades de observação são usadas como nós. Há 31 idades, então o número total de funções base B-*spline* é 29+6=35. Um método comum para forçar a curva a ser suave é penalizar a sua curvatura. É comum definir a curvatura total da curva como sendo a integral da sua segunda derivada ao quadrado. Isso é uma maneira de medir a não suavidade da curva. Aqui há também um interesse pelas curvas de aceleração, que são obtidas através da segunda derivada das respectivas funções altura. Portanto, é necessário agora controlar não somente a suavidade da função em si, mas também a suavidade de sua segunda derivada. Para isso é preciso penalizar a curvatura da segunda derivada, ou seja, a integral da quarta derivada ao quadrado. Os resultados obtidos podem ser vistos nas Figuras 1.1 e 1.4-(b). Mais detalhes sobre o método de penalização da não suavidade podem ser encontrados no Capítulo 4.

1.6.4 Velocidade e aceleração

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"deveria ser usado para $V(t)$. Devido a altura não decrescer (pelo menos durante os anos de crescimento), a velocidade ou crescimento é necessariamente positivo. 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,

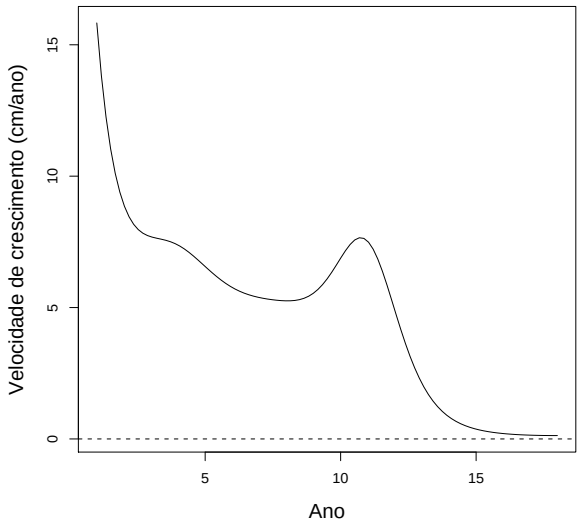
$$V(t_i) = [H(t_{i+1}) - H(t_i)]/(t_{i+1} - t_i),$$

mas 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 muito melhor ajustar os dados de altura com uma curva suave apropriada e, então, estimar a velocidade encontrando a inclinação dessa curva suave.

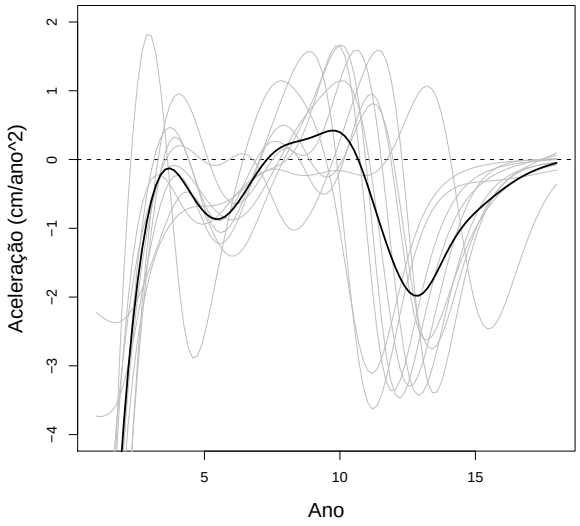
A Figura 1.5-(a) mostra a curva estimada de velocidade para a primeira garota no Estudo de Crescimento de Berkeley. Agora é possível ver mais claramente o que está acontecendo. O estirão de crescimento na Figura 1.5-(a) é certamente mais óbvio. Nota-se que por volta dos dez anos a velocidade de crescimento aumenta consideravelmente. Agora também sabe-se que é necessário trabalhar muito para encontrar bons métodos para estimar a velocidade.

```
altGarotas <- growth$hgtf[,1:10]
idade <- growth$age
amplitude <- range(idade)
idade2    <- seq(1,18,length=101)
nos  <- idade   # em inglês knots
nOrdem <- 6
nBases <- length(nos)+nOrdem-2
altBases <- create.bspline.basis(amplitude, nBases, nOrdem, nos)
garotasfdPar <- fdPar(altBases, Lfdobj=3, lambda=10^(-1.5))
altMonotona <- matrix(0, length(idade2), 10)
velocidade <- matrix(0, length(idade2), 10)
aceleracao <- matrix(0, length(idade2), 10)
for (i in 1:10) {
   resultado <-smooth.monotone(idade, altGarotas[,i],garotasfdPar,
   active=c(TRUE,rep(TRUE,nBases-1)))
   Wfd     <- resultado$Wfdobj
   beta    <- resultado$beta
   altMonotona[,i] <- beta[1] + beta[2]*eval.monfd(idade2, Wfd)
   velocidade[,i] <- beta[2]*eval.monfd(idade2, Wfd, 1)
   aceleracao[,i] <- beta[2]*eval.monfd(idade2, Wfd, 2)}
```

Pode-se obter mais informação do processo de crescimento estudando a taxa de mudança na velocidade, ou seja, a aceleração, denotada por $A(t)$. As curvas de aceleração estimadas para as dez garotas do Estudo de Crescimento de Berkeley podem ser observadas na Figura 1.5-(b) juntamente com a curva média. Agora, pode-se ver ainda mais claramente o que acontece no estirão de crescimento puberal. Observa-se um grande aumento repentino na aceleração no início do estirão de crescimento puberal, 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. É possível observar que o momento do estirão de crescimento puberal varia bastante de garota para garota. Pode-se também notar que existe uma ou mais oscilações na aceleração antes do estirão de crescimento puberal. A capacidade de detectar essas oscilações foi um dos importantes avanços recentes da tecnologia não paramétrica de estimação nessa área.



(a)



(b)

Figura 1.5: (a) Velocidade estimada de crescimento, ou taxa de crescimento da primeira garota. (b) Curvas de aceleração do crescimento estimadas para as 10 garotas, cujos os dados estão na Figura 1. Em destaque tem-se a curva média de aceleração.

Capítulo 2

Notação e técnicas algébricas

2.1 Notação geral

A notação utilizada nesse trabalho está baseada em Ramsay e Silverman (2006). Assim, o símbolo x pode se referir a um escalar ou a uma função com valores $x(t)$. Através do contexto será sempre possível entender quando um símbolo se refere a um escalar ou a uma função.

Por outro lado, a notação convencional para vetores e matrizes será adotada: vetores serão denotados por letras minúsculas em negrito, como em $\mathbf{x}$, e matrizes por letras maiúsculas em negrito, como em $\mathbf{X}$. A notação $\mathbf{x}'$ será usada sempre para a transposta de um vetor $\mathbf{x}$.

Os elementos de um vetor $\mathbf{x}$, x_i, ou os valores de uma função x, $x(t)$, são geralmente escalares, mas, às vezes, pode ser apropriado que x_i ou $x(t)$ seja um vetor e, então, a notação em negrito será utilizada. É útil também utilizar a notação $x(\mathbf{t})$ para denotar o vetor contendo os valores da função x para cada elemento do vetor $\mathbf{t}$.

A notação para a derivada de ordem m de uma função x é $D^m x$. Essa notação produz fórmulas mais simples do que $d^m x/dt^m$ e também enfatiza que a diferenciação é um operador que atua em uma função x para produzir uma outra função Dx.

O objetivo principal da notação aqui apresentada é a facilitação da leitura e do entendimento desse texto.

2.2 Notação estocástica

2.2.1 O que é uma variável funcional?

Considere a seguinte situação usual onde alguma variável aleatória pode ser observada em vários pontos diferentes no intervalo $\mathcal{T} = [a, b]$. Uma observação pode ser expressa pela família aleatória $\{X(t_j)\}_{j=1,\ldots,J}$. Na estatística moderna, os instantes consecutivos se tornaram mais e mais próximos. Uma maneira de levar isso em consideração é assumir agora que os dados são uma observação da família contínua $\mathcal{X} = \{X(t); t \in \mathcal{T}\}$. Esse é, por exemplo, o caso dos dados do crescimento humano apresentados na Seção 1.6. De forma geral, uma variável aleatória $\mathcal{X}$ é chamada de variável funcional se ela assume valores em um espaço infinito dimensional (ou espaço funcional). Uma observação χ de $\mathcal{X}$ é chamada de dado funcional. Por questão de simplicidade, χ poderia também ser descrito simplesmente como x e, da mesma forma, $\mathcal{X}$ como X.

Note que, quando $\mathcal{X}$ denota uma curva aleatória isso implica que $\mathcal{T} \subset \mathbb{R}$. Aqui, é importante dizer que a noção de variável funcional abrange uma grande área de pesquisa, ou seja, não somente análise de curvas. Em particular, uma variável funcional pode ser uma superfície aleatória (e, nesses casos, $\mathcal{T} \subset \mathbb{R}^2$) ou qualquer outro objeto matemático infinito dimensional mais complicado.

2.2.2 O que são conjuntos de dados funcionais?

Um conjunto de dados funcionais $\chi_1, \ldots, \chi_n$ é formado pelas observações de n variáveis funcionais $\mathcal{X}_1, \ldots, \mathcal{X}_n$ identicamente distribuídas como $\mathcal{X}$.

Essa definição abrange muitas situações e a mais popular delas ocorre quando o conjunto de dados é formado por curvas. Sabe-se que os dados são coletados de forma discreta, portanto um estágio preliminar consiste em apresentar os dados de forma a facilitar o processamento funcional. Na maioria das situações, métodos de suavização podem ser utilizados, tais como os que estão descritos nos Capítulos 3 e 4. Existem, porém, situações onde são necessárias técnicas de suavização mais sofisticadas, por exemplo, quando se tem poucas observações por objeto em estudo (dados esparsos). Mais detalhes podem ser encontrados em Ferraty e Vieu (2006).

2.3 Produtos internos

2.3.1 Notação genérica e exemplos

Deseja-se usar uma notação comum para o produto interno de vetores ou funções sem se preocupar com os detalhes do cálculo. Portanto, a notação genérica $\langle x, y\rangle$ é usada para denotar o produto interno entre x e y. As propriedades fundamentais de um produto interno são:

- **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 $x = 0$;
- **Bilinearidade**: para todos números reais a e b, $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$ para todo x, y e z.

Como exemplo, considere o produto interno euclidiano

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y}\rangle &= \mathbf{x}'\mathbf{y} \\ &= \sum_i x_i y_i \end{aligned} \tag{2.1}$$

onde $\mathbf{x}$ e $\mathbf{y}$ são vetores de mesmo tamanho. É fácil verificar que o produto interno euclidiano satisfaz as propriedades de simetria, positividade e bilinearidade.

Note que $\mathbf{x}'\mathbf{W}\mathbf{y}$, onde $\mathbf{W}$ é uma matriz definida positiva de ordem apropriada, também possui as propriedades de simetria, positividade e bilinearidade e pode ser usado em quase todos os lugares onde o produto interno euclidiano é usado. Por exemplo, $\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{y}$, onde $\mathbf{\Sigma}$ é uma matriz de covariância, é usado para definir a distribuição normal multivariada, encontrar as estimativas de mínimos quadrados generalizados, dentre outras coisas também úteis.

Agora suponha que x e y não são vetores, mas sim funções com valores $x(t)$ e $y(t)$, respectivamente. O funcional natural para $x'y$ é $\int x(t)y(t)dt$, substituindo a soma em (2.1) por uma integral. Isso também é simétrico em x e y, linear em cada função e satisfaz o requerimento de positividade. As mesmas conclusões podem ser tiradas para a operação $\int w(t)x(t)y(t)dt$, onde w é uma função peso estritamente positiva, e também para a operação mais

geral $\int\int w(s,t)x(s)y(t)dsdt$ se w é estritamente positiva definida, o que faz com que o requerimento de positividade para o produto interno seja satisfeito.

2.3.2 Propriedades gerais

É possível considerar um produto interno como sendo uma medida escalar de associação entre pares de quantidades x e y. A natureza simétrica da medida significa que ela é invariante com respeito à ordem das quantidades. A bilinearidade significa que ao mudar a escala de um dos argumentos, tem-se o mesmo efeito de escala na medida de associação, e que também a medida de associação entre uma quantidade e a soma de outras duas é a soma das medidas individuais de associação. Essas são duas propriedades naturais que uma medida de associação deve ter.

A positividade significa que o produto interno de qualquer x com ele mesmo é essencialmente uma medida de seu tamanho. A raiz quadrada positiva dessa medida de tamanho é chamada de norma de x, $\|x\|$, de forma que

$$\|x\|^2 = \langle x, x \rangle ,$$

com $\|x\| \geq 0$. Se x é um vetor $n \times 1$ e o produto interno é o produto interno euclidiano (2.1), a norma de x é simplesmente o tamanho do vetor medido no espaço n-dimensional. No caso de uma função f, um tipo básico de norma é

$$\|f\| = \sqrt{\int f^2}$$

chamada de norma $\mathcal{L}^2$.

As propriedades de produtos internos levam às seguintes propriedades de norma:

- $\|x\| \geq 0$ e $\|x\| = 0$ se e somente se $x = 0$;
- $\|ax\| = |a| \, \|x\|$ para todo número real a;
- $\|x + y\| \leq \|x\| + \|y\|$.

Das propriedades de produto interno também segue a bem conhecida desigualdade de Cauchy-Schwarz,

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| ,$$

que por sua vez leva à desigualdade do cosseno:

$$-1 \leq \frac{\langle x, y\rangle}{\|x\| \, \|y\|} \leq 1.$$

A desigualdade do cosseno relaciona o produto interno com o conceito geométrico de ângulo. O ângulo entre x e y pode ser definido como sendo o ângulo θ tal que

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\| \, \|y\|}. \tag{2.2}$$

Quando x e y são vetores de mesmo tamanho n e o produto interno é o euclidiano, θ é o ângulo entre x e y no sentido geométrico usual. Semelhantemente, quando x e y são funções, o cosseno do ângulo entre elas também é dado por (2.2), ou seja,

$$\cos\theta = \frac{\int xy}{\sqrt{(\int x^2)(\int y^2)}}.$$

A particular relação $\langle x, y\rangle = 0$, chamada ortogonalidade, implica que x e y são ortogonais, ou seja, formam um ângulo reto. A ortogonalidade desempenha um papel importante na operação de projeção que será discutida na Seção 2.4.

A partir do produto interno, deriva-se também uma medida de distância entre x e y definida por:

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

Novamente, no caso euclidiano, a distância d_{xy} corresponde à definição usual de geometria.

Pode-se concluir que as simples propriedades algébricas de produto interno (simetria, positividade e bilinearidade) levam a construção de medidas muito úteis, tais como: o tamanho de x, o ângulo e a distância entre x e y. Além disso, não importa a maneira como $\langle x, y\rangle$ está definido em uma certa aplicação, as características essenciais dessas medidas permanecem as mesmas.

A natureza do produto interno depende de algo mais fundamental sobre x e y: eles são elementos de um espaço vetorial, no qual os elementos podem ser adicionados ou multiplicados por números reais para formar novos elementos, e também no qual a adição distribui com respeito a multiplicação por um escalar. A união entre um espaço vetorial e um produto interno associado é chamada de espaço de produto interno.

2.3.3 Estatísticas descritivas em notação de produto interno

Sejam $\mathbf{x} = (x_1, \ldots, x_n)$ e $\mathbf{y} = (y_1, \ldots, y_n)$ amostras univariadas de tamanho n e seja $\mathbf{u}$ um vetor de uns de tamanho n ($\mathbf{u} = (1, 1, \ldots, 1)'$). Utilizando o produto interno euclidiano definido em 2.1, tem-se as seguintes estatísticas descritivas já familiares:

- Média: $\bar{x} = n^{-1} \langle \mathbf{x}, \mathbf{u} \rangle$. Note que $\bar{x}$ é um escalar e não um vetor. O vetor de tamanho n em que todos os elementos são $\bar{x}$ é dado por $\bar{x}\mathbf{u}$.
- Variância: $s_x^2 = (n-1)^{-1} \langle \mathbf{x} - \bar{x}\mathbf{u}, \mathbf{x} - \bar{x}\mathbf{u} \rangle = n^{-1} \|\mathbf{x} - \bar{x}\mathbf{u}\|^2$.
- Covariância: $s_{xy} = (n-1)^{-1} \langle \mathbf{x} - \bar{x}\mathbf{u}, \mathbf{y} - \bar{y}\mathbf{u} \rangle$.
- Correlação: $r_{xy} = s_{xy}/(s_x s_y)$.

Suponha que agora, ao invés do produto interno euclidiano, seja usado o seguinte produto interno:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i w_i x_i y_i,$$

onde w_i é um peso não negativo aplicado à i-ésima observação. Isso não faz nenhuma diferença, exceto que agora é necessário dividir $\bar{x}$, s_x^2 e s_{xy} por uma constante $\sum_i w_i$, ao invés de dividir por n ou $(n-1)$. É claro que os pesos afetam os valores das estatísticas, mas os seus significados permanecem basicamente os mesmos.

Essa ideia pode ser generalizada. Suponha uma sequência de observações correlacionadas, com matriz de covariância $\mathbf{\Sigma}$. Assim, pode-se usar $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{y}$ como uma base para uma estatística descritiva que leva em conta a estrutura de covariância das observações.

Agora, considere essas mesmas estatísticas no contexto em que a amostra é uma função x com valores $x(t)$, onde o argumento t assume valores dentro de algum intervalo na reta, tal como $[0, T]$. Então, o índice i, assumindo n possíveis valores, é agora substituído pelo índice t, que assume infinitos valores. Defina o produto interno entre x e y como

$$\langle x, y \rangle = \int_0^T x(t)y(t)dt,$$

onde as funções são suficientemente bem comportadas de tal forma que a integral seja sempre definida e finita. Seja u a função que vale 1 para qualquer valor de t, ou seja, $u(t) = 1 \, \forall \, t$. Assim, de forma semelhante ao caso vetorial, as estatísticas descritivas agora são dadas por:

- $\bar{x} = \langle x, u \rangle \, / \int_0^T 1 dt = \langle x, u \rangle \, / \, \|u\|^2$,
- $s_x^2 = \|x - \bar{x}u\|^2 \, / \, \|u\|^2$ e
- $s_{xy} = \langle x - \bar{x}u, y - \bar{y}u \rangle \, / \, \|u\|^2$.

Nesse caso, $\bar{x}$ se torna o nível médio da função x; s_x^2 se torna uma medida da sua variação em torno do seu nível médio, e s_{xy} e r_{xy} medem o grau de relacionamento entre x e y. Considerando

$$\langle x, y \rangle = \int_0^T w(t)x(t)y(t)dt$$

para alguma função peso positiva w e dividindo por $\int w(t)dt$ ao invés de $\|u\|^2$, essas interpretações não mudariam de forma essencial, exceto que diferentes partes da amplitude de t teriam importâncias diferentes.

2.3.4 Outros usos da notação de produto interno

Até aqui o resultado de um produto interno tem sido sempre um simples número real. Uma forma de tornar a notação mais abrangente é a seguinte: seja $x = (x_1, \ldots, x_m)'$ um vetor de tamanho m, onde cada x_i é um elemento de algum espaço vetorial finito-dimensional ou funcional. Então, a notação $\langle x, y \rangle$, onde y é um simples elemento do mesmo espaço vetorial, indica o vetor m-dimensional, cujos elementos são $\langle x_1, y \rangle, \ldots, \langle x_m, y \rangle$. Além disso, se y é da mesma forma que x, porém de tamanho n, então a notação $\langle x, y \rangle$ define uma matriz com m linhas e n colunas contendo os valores $\langle x_i, y_j \rangle$ para $i = 1, \ldots, m$ e $j = 1, \ldots, n$. Essa convenção só é usada em situações onde o contexto deixa claro se x e/ou y são vetores de elementos do espaço em questão.

No contexto funcional, às vezes se escreve

$$\langle z, \beta \rangle = \int z(s)\beta(s)ds,$$

mesmo quando as funções z e β não estão no mesmo espaço. Espera-se que o contexto deixe claro que um verdadeiro produto interno não está envolvido nesse caso. Uma alternativa seria usar uma notação diferente, como por exemplo (z, β).

Uma propriedade importante é que $\langle z, \beta \rangle$ é sempre um operador linear, quando considerado como uma função de um dos seus argumentos. Um operador linear num espaço de funções é um mapeamento A, tal que, para todo f_1 e f_2 no espaço e para todos os escalares a_1 e a_2, $A(a_1 f_1 + a_1 f_2) = a_1 A f_1 + a_2 A f_2$.

2.4 Projeção ortogonal

Sejam $v_1, \ldots, v_n$ elementos de um espaço vetorial e seja $\mathcal{V}$ o subespaço com todas as possíveis combinações lineares desses elementos. Pode-se caracterizar o subespaço $\mathcal{V}$ usando uma notação vetorial adequada. Seja v o vetor n-dimensional cujos elementos são os $v_1, \ldots, v_n$. Então, todo membro de $\mathcal{V}$ é da forma $v'c$, para algum vetor real c de dimensão n.

Associado com o subespaço $\mathcal{V}$ está a projeção ortogonal sobre $\mathcal{V}$, definida como sendo um operador linear P com as seguintes propriedades:

1. Para todo z, o elemento $Pz \in \mathcal{V}$ e, assim, é uma combinação linear de $v_1, \ldots, v_n$.
2. Se y já está em $\mathcal{V}$, então $Py = y$.
3. Para todo z, o resíduo $z - Pz$ é ortogonal a todos os elementos de $\mathcal{V}$.

Das primeiras duas propriedades segue que $PP = P^2 = P$. Através da terceira propriedade pode-se mostrar que o operador P mapeia cada elemento z para o seu ponto mais próximo em $\mathcal{V}$. Isso faz com que as projeções sejam muito importantes nos contextos estatísticos, tais como a estimação por mínimos quadrados.

2.5 Estatísticas descritivas para dados funcionais

2.5.1 Média e variância amostral

As estatísticas descritivas para dados funcionais são obtidas de forma semelhante ao caso de dados univariados. Assim, a função média para uma amostra

contendo n funções é dada por

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t), \tag{2.3}$$

para cada ponto $t \in \mathcal{T} = [a, b]$. Da mesma forma, a função variância é definida por

$$var_x(t) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(t) - \bar{x}(t)]^2, \tag{2.4}$$

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

2.5.2 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 seguinte fórmula:

$$cov_x(t_1, t_2) = \frac{1}{n-1}\sum_{i=1}^{n}\{x_i(t_1) - \bar{x}(t_1)\}\{x_i(t_2) - \bar{x}(t_2)\}. \tag{2.5}$$

A função de correlação por sua vez é dada por

$$corr_x(t_1, t_2) = \frac{cov_x(t_1, t_2)}{\sqrt{var_x(t_1)var_x(t_2)}}. \tag{2.6}$$

Dessa forma é possível obter de forma análoga à análise multivariada, as matrizes de covariância e de correlação.

Capítulo 3

Suavização de dados funcionais

3.1 Introdução

3.1.1 Dados funcionais observados

A ideia básica da análise de dados funcionais é considerar os dados observados como entidades únicas, e não meramente como uma sequência de observações individuais. O termo funcional se refere à estrutura intrínseca dos dados ao invés da sua forma explícita, já que na prática esses dados são geralmente observados e registrados discretamente. O registro de uma observação funcional x consiste de n pares (t_j, y_j), onde y_j é uma observação de $x(t_j)$, um retrato da função no argumento t_j. Na grande maioria das vezes os dados funcionais são registrados ao longo do tempo. Dessa forma o argumento t_j é chamado de tempo t_j. Porém, o argumento poderia ser uma outra medida no contínuo, tal como: posição espacial, frequência, peso, etc.

Nesse capítulo, algumas técnicas usadas para converter dados funcionais (como são coletados) em sua verdadeira forma funcional serão consideradas. Devido à presença de ruído na maioria dos conjuntos de dados, a representação funcional geralmente envolve suavização e, portanto, algumas técnicas de suavização serão revistas. Uma atenção especial será dada para a estimação de derivadas, uma vez que elas são importantes em muitas análises de

dados funcionais. De acordo com a notação já estabelecida, $D^m x$ indica a m-ésima derivada de uma função univariada x. O valor da m-ésima derivada no argumento t é denotado por $D^m x(t)$.

O que significa uma observação funcional x ser conhecida na sua verdadeira forma funcional? É claro que isso não significa que x é avaliada em todo valor de t, pois isso envolveria um número incontável de valores. Na verdade, isso significa que é suposta a existência de uma função x, baseada nos dados observados, implicando, a princípio, que é possível avaliar x em qualquer ponto t, e ainda avaliar também qualquer uma das suas derivadas $D^m x$ que existam em t. Mas como somente valores discretos estão disponíveis, avaliar $x(t)$ e $D^m x(t)$ em qualquer valor arbitrário s envolverá os métodos de suavização a serem apresentados.

Em geral, trabalha-se com uma coleção de dados funcionais, ao invés de uma única função x. Especificamente, a observação da função x_i é formada por n_i pares (t_{ij}, y_{ij}), $j = 1, \ldots, n_i$. Pode ser que os argumentos t_{ij} sejam os mesmos para cada observação, mas eles também podem ser diferentes. Na reconstrução das observações funcionais, as curvas são geralmente consideradas uma por uma. Por uma questão de simplicidade, nesse capítulo será considerado que uma única função x é observada.

Assume-se sempre que a amplitude dos valores de interesse para o argumento t é um intervalo limitado $\mathcal{T}$ e, implicitamente ou explicitamente, que x satisfaz condições razoáveis de continuidade e suavidade em $\mathcal{T}$. Sem algumas condições como essas, é impossível fazer qualquer inferência sobre os valores $x(t)$ em qualquer ponto t, com exceção dos atuais pontos de observação.

3.1.2 Erro observacional

A suavidade, no sentido de possuir um certo número de derivadas, é uma propriedade da função latente x, e pode não ser óbvia quando olhamos o vetor de dados observados $\mathbf{y} = (y_1, \ldots, y_n)$. Isso ocorre devido ao erro observacional ou ruído presente nos dados, causado por aspectos do processo de medida. Em termos de modelo, pode-se escrever

$$y_j = x(t_j) + \epsilon_j, \tag{3.1}$$

onde o ruído, erro, perturbação ou então termo exógeno ϵ_j contribui para a não suavidade dos dados. Uma das tarefas em representar os dados como funções

suaves é tentar filtrar esse ruído da forma mais eficiente possível. Em certos casos outra alternativa é adotada, deixa-se o ruído nos dados e somente se faz a suavização dos resultados da análise.

A equação (3.1) pode ser reescrita em notação vetorial como

$$\mathbf{y} = x(\mathbf{t}) + \mathbf{e}, \tag{3.2}$$

onde $\mathbf{y}$ e $\mathbf{e}$ são vetores coluna de tamanho n. Por sua vez, $x(\mathbf{t})$ é o vetor de tamanho n com os valores da função x avaliada nos pontos de observação $t_1, \ldots, t_n$.

A matriz de variância e covariância do vetor de observações $\mathbf{y}$ é igual à matriz de variância e covariância do correspondente vetor $\mathbf{e}$, uma vez que os valores $x(t_j)$ são considerados aqui efeitos fixos com variância zero. Seja $\mathbf{\Sigma}_e$ a matriz de variância e covariância dos erros. A suposição padrão para os ϵ_j's é que eles são independentes e identicamente distribuídos, com média zero e variância constante igual a σ^2. Consequentemente,

$$Var(\mathbf{y}) = \mathbf{\Sigma}_e = \sigma^2 \mathbf{I}, \tag{3.3}$$

onde $\mathbf{I}$ é uma matriz identidade de ordem n.

Em casos especiais, deve-se levar em conta a não homogeneidade da variância e a correlação entre os erros ϵ_j's para valores de argumento numa certa vizinhança.

Como palavra final sobre erro observacional, deseja-se dizer que o modelo clássico (3.1) poderia ser substituído pelo seguinte modelo

$$y_j = x(t_j) + \epsilon(t_j), \tag{3.4}$$

onde a função ruído ϵ é adicionada ao sinal suave x. Pode-se assumir que ϵ tem as características intuitivas de um ruído branco: média zero, variância constante e covariância zero para valores de argumento distintos.

Pode-se ter uma razão importante para preferir o modelo com ruído discreto (3.1) ao modelo com ruído funcional (3.4). Por exemplo, os dados do crescimento possuem claramente ruído discreto. Através de experimentos, sabe-se que medidas independentes de uma mesma criança em um tempo fixo apresentam um desvio padrão por volta de 1,5 mm.

Embora as técnicas nesse capítulo sejam projetadas para filtrar o erro observacional nas funções propriamente ditas, é importante dizer que o ruído, tanto discreto como funcional, pode afetar também uma ou mais derivadas.

3.1.3 Taxa de amostragem

A taxa de amostragem ou resolução dos dados observados é um fator chave para determinar o que é possível fazer na análise de dados funcionais. Isso é essencialmente uma propriedade local dos dados e pode ser descrita como a densidade dos valores de argumento t_j relativa à quantidade de curvatura nos dados, ao invés de ser simplesmente descrita como o número n de valores de argumento.

A curvatura de uma função x no argumento t é geralmente medida através do tamanho da segunda derivada, ou seja, por $\left|D^2x(t)\right|$ ou $[D^2x(t)]^2$. Quando a curvatura é grande precisa-se de um número suficiente de pontos para estimar a função efetivamente. Mas o que seria suficiente? Isso depende da quantidade de erro ϵ_j. Quando o nível de erro é pequeno e a curvatura é suave, pode-se trabalhar com baixa taxa de amostragem. Os dados de crescimento na Figura 1.1 possuem níveis de erro relativamente baixos, mas a curvatura nas funções de segunda derivada é bastante severa (Figura 1.5-(b)), de tal forma que a taxa de amostragem para esses dados não é muito adequada para fazer inferências sobre a aceleração.

Várias formas de representar as observações discretas y_j como uma função suave x devem ser investigadas, prestando atenção na estimação de suas derivadas. Mas, primeiramente, deve-se ter cuidado ao se tentar estimar as derivadas diretamente dos dados observados. Devido às funções observadas parecerem razoavelmente suaves, talvez alguém seja tentado a usar a diferença $(y_{j+1}-y_j)/(t_{j+1}-t_j)$ ou a diferença central $(y_{j+1}-y_{j-1})/[2(t_{j+1}-t_{j-1})]$ para estimar $Dx(t_j)$, porém, o resultado pode ser uma estimativa com bastante ruído. Isso ocorre especialmente quando se tem uma alta taxa de amostragem, pois quando se faz a diferença entre valores muito próximos a influência do erro aumenta muito mais. Embora seja uma prática comum, a estimação da derivada usando diferenças é raramente uma boa ideia.

3.1.4 Suavização linear

Um suavizador linear estima o valor da função $x(t)$ através de uma combinação linear de observações discretas, ou seja,

$$\hat{x}(t) = \sum_{j=1}^{n} S_j(t) y_j. \tag{3.5}$$

O comportamento do suavizador em t é determinado pelos pesos $S_j(t)$.

Suavizadores lineares podem ser representados de forma matricial. Suponha uma sequência $s_1 < s_2 < \cdots < s_m$ de valores de avaliação em $\mathcal{T}$ nos quais a função x deve ser estimada. Note que os valores de avaliação não precisam ser iguais aos valores observados t_j. Seja $\hat{x}(\mathbf{s})$ o vetor de tamanho m com valores $x(s_i)$ e $\mathbf{y}$ como sendo o vetor de dados observados y_j. Então, pode-se escrever

$$\hat{x}(\mathbf{s}) = \mathbf{S}\mathbf{y}, \tag{3.6}$$

onde $\mathbf{S}$ é uma matriz $m \times n$ com elementos $S_{ij} = S_j(s_i)$.

Muitos dos métodos mais comuns de suavização são lineares. A linearidade é uma característica desejada por diversas razões. Uma delas é a propriedade de linearidade

$$\mathbf{S}(a\mathbf{y} + b\mathbf{z}) = a\mathbf{S}\mathbf{y} + b\mathbf{S}\mathbf{z},$$

que é importante para entender outras propriedades da representação suave. Outra razão é a simplicidade desse suavizador que implica em uma computação relativamente rápida. Por outro lado, alguns suavizadores não lineares podem ser mais efetivos quando se tem comportamentos diferentes em partes distintas do intervalo de observação, e eles podem também ser mais robustos à observações aberrantes (*outliers*). A suavização através de uma mudança na transformada *wavelet*, apresentada na Seção 3.2.8, é um importante exemplo de suavização não-linear.

3.2 Métodos de funções base

3.2.1 O ajuste de mínimos quadrados

Um método comum de suavização consiste em representar a função como uma combinação linear de K funções base conhecidas ϕ_k, ou seja,

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t). \tag{3.7}$$

Seja $\mathbf{c} = (c_1, \ldots, c_K)'$ o vetor de coeficientes e $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_K(t))'$ o vetor cujos elementos são as funções base avaliadas em t. Assim, pode-se

escrever (3.7) como

$$x(t) = \mathbf{c}'\boldsymbol{\phi}(t). \tag{3.8}$$

O grau de suavidade que será aplicado nos dados y_j é determinado pelo número de funções base K. Seja $\mathbf{\Phi}$ a matriz $n \times K$ cujos elementos são $\Phi_{jk} = \phi_k(t_j)$, ou seja, o valor da k-ésima função base calculado no ponto de observação t_j para $k = 1, \ldots, K$ e $j = 1, \ldots, n$. Assumindo que $\mathbf{\Phi}$ tenha posto completo, é geralmente possível encontrar uma representação exata dos dados (interpolação) quando $K = n$, uma vez que se pode calcular os coeficientes c_k de tal forma a se ter $x(t_j) = y_j$ para cada j (ver Figura 3.1-(b) quando $K = n = 96$).

Por outro lado, quando $K < n$ o método de suavização mais simples, utilizando a representação por funções base, é obtido quando os coeficientes c_k são determinados minimizando o critério de mínimos quadrados, ou se preferir, a soma dos quadrados dos erros

$$SQE = \sum_{j=1}^{n} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2. \tag{3.9}$$

Em termos matriciais tem-se:

$$SQE = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c}) = \|\mathbf{y} - \mathbf{\Phi c}\|^2. \tag{3.10}$$

Sabe-se de análise de regressão que esse critério é minimizado quando $\mathbf{c} = (\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}$, chamado de $\hat{\mathbf{c}}$ (ver Seber (1977)).

Assumindo que os pontos de avaliação são exatamente os pontos de observação, a matriz de suavização $\mathbf{S}$ (ver equação 3.6), também chamada de matriz chapéu, é dada por

$$\mathbf{S} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'. \tag{3.11}$$

Para encontrar o resultado acima basta escrever $\hat{x}(t_j) = \sum_{k=1}^{K} \hat{c}_k \phi_k(t_j)$ para $j = 1, \ldots, n$. Assim, em termos matriciais tem-se que $\hat{x}(\mathbf{t}) = \mathbf{\Phi}\hat{\mathbf{c}}$. Substituindo $\hat{\mathbf{c}}$ por $(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}$ tem-se que $\hat{x}(\mathbf{t}) = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}$. Então, de acordo com a definição apresentada na Seção 3.1.4, $\mathbf{S} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'$.

A matriz $\mathbf{S}$ nesse caso é uma matriz de projeção ortogonal, uma vez que ela é simétrica e satisfaz a relação de idempotência ($\mathbf{S}^2 = \mathbf{S}$). Assim, um

suavizador baseado no método de mínimos quadrados usando a expansão por funções base é simplesmente uma projeção ortogonal de um vetor observado $\mathbf{y}$, no espaço gerado pelas colunas da matriz base $\mathbf{\Phi}$.

No caso mais geral, se os pontos de avaliação s_i não são necessariamente os pontos de observação, defina a matriz $m \times K$ $\tilde{\mathbf{\Phi}}$ cujos elementos são $\phi_j(s_i)$. Dessa forma,

$$\hat{x}(\mathbf{s}) = \tilde{\mathbf{\Phi}}\hat{\mathbf{c}} = \tilde{\mathbf{\Phi}}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}$$

e a matriz de suavização é dada por

$$\mathbf{S} = \tilde{\mathbf{\Phi}}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'.$$

A aproximação por mínimos quadrados simples é apropriada em situações onde se assume que os erros ϵ_j's, sobre a verdadeira curva, são independentes e identicamente distribuídos com média zero e variância constante σ^2, ou seja, esse método é preferido quando a suposição padrão para os erros discutida na Seção 3.1.2 é assumida.

Porém, a suposição padrão para os erros não é em geral compatível com a realidade. Para lidar com a não estacionariedade e/ou com erros autocorrelacionados, o critério de mínimos quadrados pode ser estendido para a forma

$$SQE = (\mathbf{y} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{y} - \mathbf{\Phi c}) = \|\mathbf{y} - \mathbf{\Phi c}\|^2_{\mathbf{W}}\,, \tag{3.12}$$

onde $\mathbf{W}$ é uma matriz conhecida simétrica definida positiva que permite colocar pesos diferentes aos quadrados e produtos dos erros. Tal extensão é conhecida como o método de mínimos quadrados ponderados ou generalizados.

Nesse caso, as estimativas dos coeficientes são $\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{y}$. Quando os pontos de avaliação são os mesmos que os de observação, a correspondente matriz de suavização é

$$\mathbf{S} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{W},$$

sendo ainda um operador de projeção ortogonal. Nesse caso, $\hat{\mathbf{y}} = \hat{x}(\mathbf{t}) = \mathbf{S}\mathbf{y}$ é chamada de projeção na métrica $\mathbf{W}$.

Uma questão importante é a escolha do número K de funções base. Quanto maior for o valor de K melhor será o ajuste dos dados, mas existe o risco de ajustar também ruído ou variação que deveriam ser ignorados. Por outro lado,

se K for muito pequeno, pode-se perder alguns aspectos importantes da função suave x que se deseja estimar. Além disso, o fato de K assumir somente valores inteiros faz com que o controle da suavidade possa ser relativamente grosseiro. A Seção 3.2.9 apresenta alguns tópicos com relação a essa importante escolha.

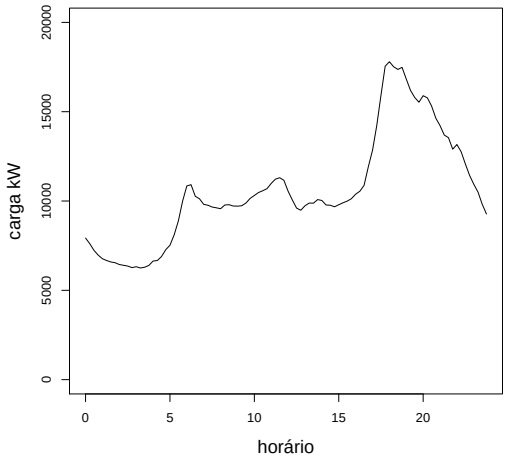
A Figura 3.1-(a) apresenta o consumo de energia elétrica ao longo de um dia em uma determinada região da cidade de Campinas. Na Figura 3.1-(b) é possível ver as representações para esse dado funcional utilizando diferentes valores de K. Nota-se que quando $K = n = 96$, os dados são interpolados, como já citado anteriormente, e não há suavidade. Por outro lado, quando K é pequeno há muita suavidade e a estimativa não descreve importantes características do dado funcional.

```
dados <- read.table('residencial.txt')
t <- dados[,1] ; consumo <- dados[,15]
library(fda)
dadosBases <- create.bspline.basis(range(t), nbasis=10, norder=4)
dadosfdPar <- fdPar(dadosBases, Lfdobj=int2Lfd(0), lambda=0)
objFunc <- smooth.basis(t,consumo, dadosfdPar)$fd
xhat1 <- eval.fd(t,objFunc)
plot(t,xhat1,ylim=c(0,20000),xlab="horário",ylab="carga kW",type="l")
```

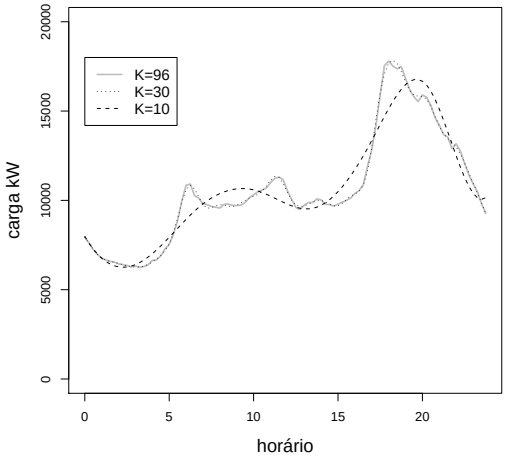
Quando o número de pontos de observação n é grande, uma computação eficiente é muito importante. Existem três tarefas essenciais para calcular as estimativas em pontos de avaliação em geral:

1. Calcular os produtos internos, que são K em $\mathbf{\Phi}'\mathbf{y}$ e $K(K+1)/2$ em $\mathbf{\Phi}'\mathbf{\Phi}$.
2. Resolver o sistema linear $\mathbf{\Phi}'\mathbf{\Phi}\mathbf{c} = \mathbf{\Phi}'\mathbf{y}$.
3. Calcular os m produtos internos $\tilde{\mathbf{\Phi}}\hat{\mathbf{c}}$, onde a matriz $\tilde{\mathbf{\Phi}}$ contém as funções base avaliadas nos argumentos de avaliação.

Algoritmos eficientes e estáveis de mínimos quadrados podem efetuar os cálculos em $O((n+m)K^2)$ operações. Isso é possível quando K é pequeno relativo a n e m. Para K grande é extremamente útil, tanto para economia computacional quanto para estabilidade numérica, que a matriz $\mathbf{\Phi}'\mathbf{\Phi}$ tenha uma estrutura de banda tal que valores diferentes de zero apareçam somente em um número fixo e limitado de posições em cada lado da diagonal. No pior cenário, sem a estrutura de banda e com K e m de $O(n)$, a computação é de $O(n^3)$, o que não é aceitável para n grande. Nesses casos uma computação eficiente é essencial.



(a)



(b)

Figura 3.1: (a) Aqui o dado funcional é o consumo de energia elétrica em kW ao longo de um bairro residencial em Campinas. (b) Curvas estimadas utilizando o método de expansão por funções base, nesse caso B-*splines*, para $K = n = 96$, $K = 30$ e $K = 10$.

3.2.2 Graus de liberdade

Na maioria das situações, o conceito de graus de liberdade de um ajuste significa simplesmente o número de parâmetros estimados a partir dos dados que são necessários para definir o modelo.

Esse mesmo conceito de graus de liberdade se aplica sem modificação na suavização de dados funcionais usando mínimos quadrados, onde o número de parâmetros é o tamanho do vetor de coeficientes $\mathbf{c}$, ou seja, K. Assim, o número de graus de liberdade para o erro é $n - K$.

Quando o método de penalização da não suavidade (Capítulo 4) for utilizado, as coisas não serão tão simples assim. Será necessária uma maneira mais geral para calcular os efetivos graus de liberdade de um ajuste suave aos dados e, consequentemente, os correspondentes graus de liberdade para o erro. Isso é feito usando a matriz chapéu $\mathbf{S}$. Seja $\mathbf{A}$ uma matriz quadrada, defina o traço de $\mathbf{A}$ como $\mathrm{tr}(\mathbf{A})$, lembrando que o traço de uma matriz quadrada é igual à soma dos elementos de sua diagonal. Assim, define-se o número de graus de liberdade para o ajuste suave como sendo

$$gl = \mathrm{tr}(\mathbf{S}). \tag{3.13}$$

Essa definição mais geral resulta em exatamente K quando se tem o ajuste por mínimos quadrados. Existem também algumas situações onde é a mais apropriada a seguinte definição:

$$gl = \mathrm{tr}(\mathbf{SS}'). \tag{3.14}$$

Contudo, na maioria das vezes a definição (3.13) é empregada. De qualquer forma, as duas definições produzem a mesma resposta para a estimação por mínimos quadrados.

3.2.3 Escolha da base

É importante que as funções base apresentem características semelhantes às das funções a serem estimadas. Teoricamente, uma base deveria ser escolhida por produzir um excelente ajuste e usar ao mesmo tempo um número pequeno de funções base. Isso não somente implica em menos esforço computacional, mas também os coeficientes em si podem ser usados para descrever os dados

de forma interessante. É importante dizer que existem bases que não são apropriadas para certas aplicações e que, infelizmente, não existe algo como uma base universal que seja boa para todos os casos.

A escolha da base é também muito importante para a estimativa da derivada

$$D\hat{x}(t) = \sum_{k=1}^{K} \hat{c}_k D\phi_k(t). \tag{3.15}$$

Certas bases que funcionam bem para a estimação da função podem produzir estimativas de derivadas ruins. Isso ocorre porque uma representação exata das observações tende a forçar $\hat{x}$ a ter pequenas oscilações com muita frequência, o que traz sérias consequências para as suas derivadas. Assim, um dos critérios para a escolha da base pode ser o bom comportamento da estimativa de uma ou mais derivadas.

3.2.4 Séries de Fourier

Talvez a expansão por base mais conhecida seja a obtida pelas séries de Fourier:

$$x(t) = c_0 + c_1 \text{sen}\omega t + c_2 \cos \omega t + c_3 \text{sen} 2\omega t + c_4 \cos 2\omega t + \cdots = \sum_{j=0}^{\infty} c_j \phi_j(t), \tag{3.16}$$

definida pela base $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \text{sen} r\omega t$ e $\phi_{2r}(t) = \cos r\omega t$ para $r \geq 1$. Essa base é periódica e o parâmetro ω determina o período $2\pi/\omega$, que é igual ao tamanho do intervalo $\mathcal{T}$ em que se está trabalhando. Suponha que t pertença ao intervalo $\mathcal{T} = [-\pi, \pi]$, pode-se mostrar que o sistema de funções $\{\phi_0(t), \phi_1(t), \phi_2(t), \phi_3(t), \ldots\}$ é um sistema ortogonal nesse intervalo, ou seja,

$$\langle \phi_i, \phi_j \rangle = \int_{-\pi}^{\pi} \phi_i(t)\phi_j(t)dt = \begin{cases} 0 & \text{se} \quad i \neq j \\ c_i & \text{se} \quad i = j. \end{cases}$$

Nesse caso, tem-se o que é chamado de estimador por séries ortogonais. Dividindo o sistema de funções por constantes apropriadas é possível obter um sistema ortonormal no intervalo $[-\pi, \pi]$, ou seja,

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 0 & \text{se} \quad i \neq j \\ 1 & \text{se} \quad i = j. \end{cases}$$

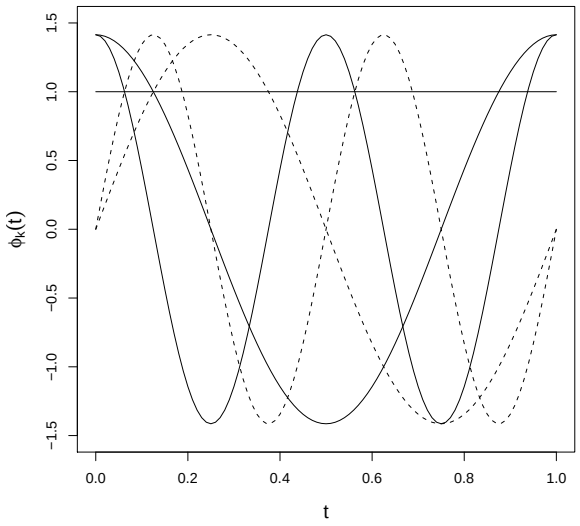


Figura 3.2: Cinco primeiras funções base de Fourier no intervalo $\mathcal{T} = [0, 1]$.

Uma vez que um número finito de observações está disponível, nem todos os coeficientes de Fourier poderão ser estimados. Portanto, pode-se considerar as primeiras K funções base da representação 3.16 e, dessa forma, a estimativa suave para $x(t)$ é dada por

$$\hat{x}(t) = \sum_{j=0}^{K-1} \hat{c}_j \phi_j(k). \tag{3.17}$$

A Figura 3.2 apresenta as cinco primeiras base de Fourier no intervalo $\mathcal{T} = [0, 1]$.

```
library(fda)
t <- seq(from=0,to=1,by=0.01)
basesFourier <- create.fourier.basis(rangeval=c(min(t), max(t)),
nbasis = 5)
MatrizBases <- getbasismatrix(t, basesFourier, nderiv=0)
```

A estimação das derivadas é simples, uma vez que $D\,\text{sen}r\omega t = r\omega\cos r\omega t$ e $D\cos r\omega t = -r\omega\text{sen}r\omega t$. Isso implica que expansão de Fourier de Dx é a expansão de x com os seguintes coeficientes:

$$(0, -\omega c_2, \omega c_1, -2\omega c_4, 2\omega c_3, \ldots).$$

Da mesma forma, pode-se obter as expansões de Fourier para derivadas de maior ordem.

As séries de Fourier são especialmente utilizadas para funções extremamente estáveis e que apresentam alguma periodicidade. Elas geralmente produzem estimativas que são uniformemente suaves, mas são inapropriadas até certo ponto para dados conhecidos ou suspeitos por refletir descontinuidades na função propriamente ou em suas primeiras derivadas. Pode-se notar na Figura 3.3 que quando a função não apresenta periodicidade é necessário um número grande de funções base para obter uma estimativa razoável. Já quando a função é periódica, no caso da Figura 3.3, com um número pequeno funções base já é possível construir uma ótima estimativa da verdadeira função.

```
library(fda)
b0 <- 2; b1 <- 1.5
t <- seq(from=0,to=1,by=0.01); x <- b0*exp(b1*2*t)
basesFourier <- create.fourier.basis(rangeval=c(min(t), max(t)),
nbasis = 5)
plot(data2fd(x, t, basesFourier),lty=2,ylim=c(0,40))
basesFourier <- create.fourier.basis(rangeval=c(min(t), max(t)),
nbasis = 15)
lines(data2fd(x, t, basesFourier),col='gray',lw=2)
lines(t,x,lty=1)
```

A Figura 3.5 apresenta a temperatura média diária em Montreal durante 101 dias de verão e 101 dias nde inverno. As médias foram obtidas a partir de observações ao longo de 34 anos, 1960-1994. As curvas estimadas foram obtidas através do método de mínimos quadrados simples utilizando 95 funções base de Fourier. Elas parecem ajustar bem as variações na temperatura que ocorrem em intervalos de tempo menores. Por exemplo, durante o inverno existe um notável período de aquecimento entre os dias 16 e 31 de janeiro.

```
library(fda)
CanadianWeather$dailyAv[1,,1]
montreal <- CanadianWeather$dailyAv[,'Montreal',1]
t <- seq(from=1,to=365,by=1)
```

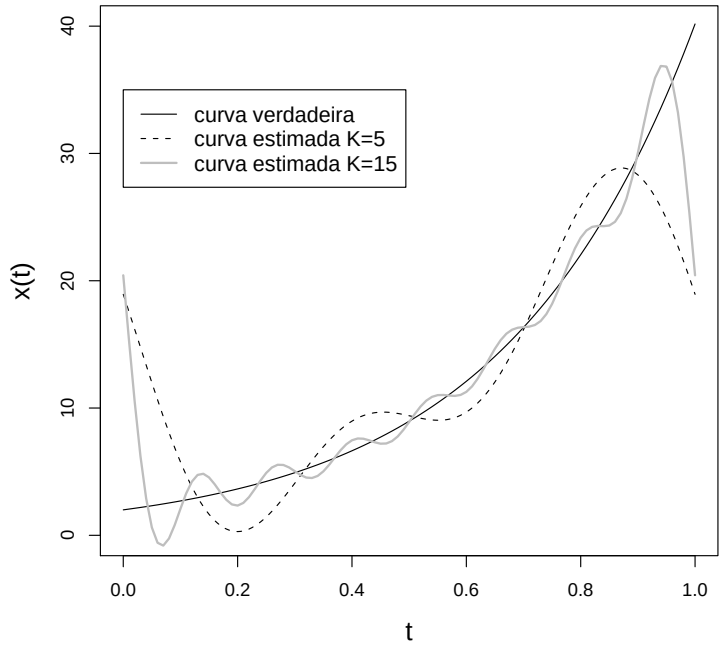


Figura 3.3: Estimação suave por séries de Fourier utilizando $K = 5$ e 15. Curva verdadeira $x(t) = 2e^{3t}$.

```
basesFourier<-create.fourier.basis(rangeval=c(min(t), max(t)),
nbasis=95)
MontrealfdPar <- fdPar(basesFourier, Lfdobj=int2Lfd(2), lambda=0)
montrealfd <- smooth.basis(t, montreal, MontrealfdPar)$fd
montrealfit <- eval.fd(t, montrealfd)
```

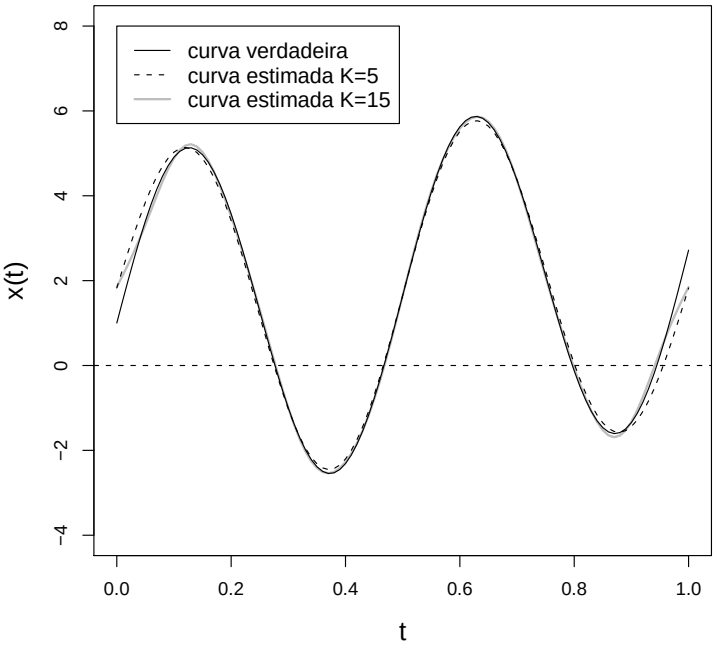


Figura 3.4: Estimação suave por séries de Fourier utilizando $K = 5$ e 15. Curva verdadeira $x(t) = e^t + 4\text{sen}(8\pi t/2)$.

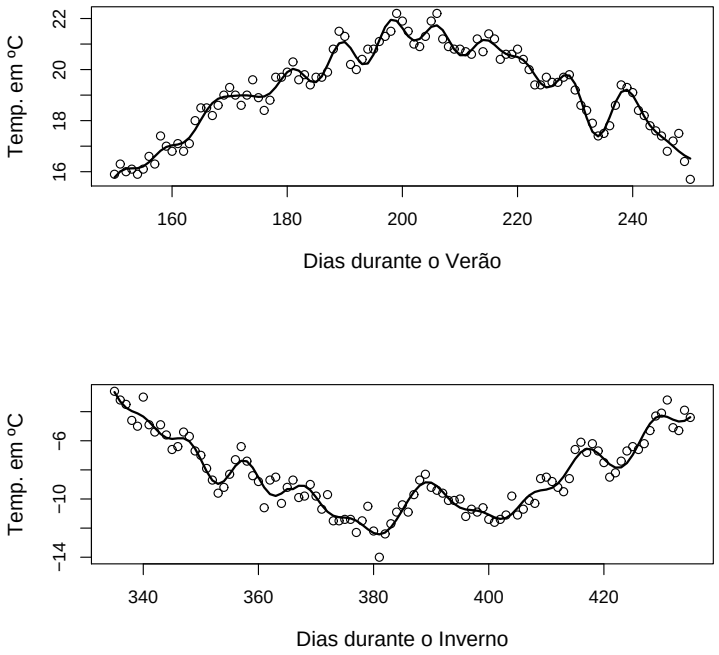


Figura 3.5: O painel superior mostra a temperatura média diária durante 101 dias de verão em Montreal. Já o painel inferior apresenta a temperatura média durante 101 dias de inverno, com os dias se estendendo para o próximo ano. As curvas sólidas são os ajustes suaves ao dados por mínimos quadrados, utilizando 95 funções base de Fourier.

3.2.5 Bases Polinomiais

A base constituída por monômios da forma $\phi_k(t) = (t-\omega)^k$, $k = 0, 1, \ldots, K$, é também muito conhecida. Infelizmente, ela pode produzir matrizes $\mathbf{\Phi}'\mathbf{\Phi}$ quase singulares e o parâmetro de mudança ω deve ser escolhido cuidadosamente. Contudo, se os argumentos t_j forem igualmente espaçados ou forem escolhidos de forma a ter um certo padrão, expansões polinomiais ortogonais podem ser obtidas facilitando assim o cálculo dos coeficientes.

Como a expansão por séries de Fourier, polinômios não podem exibir muitas características locais sem fazer uso de um K grande. Mais ainda, os polinômios tendem a ajustar bem o centro dos dados, mas o seu comportamento nas caudas não é muito bom. Eles geralmente não são uma base muito apropriada quando se quer fazer previsão.

Embora as derivadas de uma expansão polinomial sejam simples de se calcular, elas não são muito satisfatórias como estimadores das verdadeiras derivadas devido à rápida oscilação localizada típica de polinômios de alta ordem.

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

```
library(fda)
t <- seq(from=0,to=1.1,by=0.01)
basesp <- create.polynomial.basis(rangeval=c(min(t),max(t)),
nbasis=10,ctr=0)
MatrizBases <- getbasismatrix(t, basesp, nderiv=0)
plot(t,MatrizBases[,1],type='l',ylab=expression(phi[k](t)),xlab='t',
cex.lab=1.3,ylim=c(0,2.1))
for (i in 1:4){
lines(t,MatrizBases[,2*i],lty=2)
lines(t,MatrizBases[,2*i+1],lty=1)
}
```

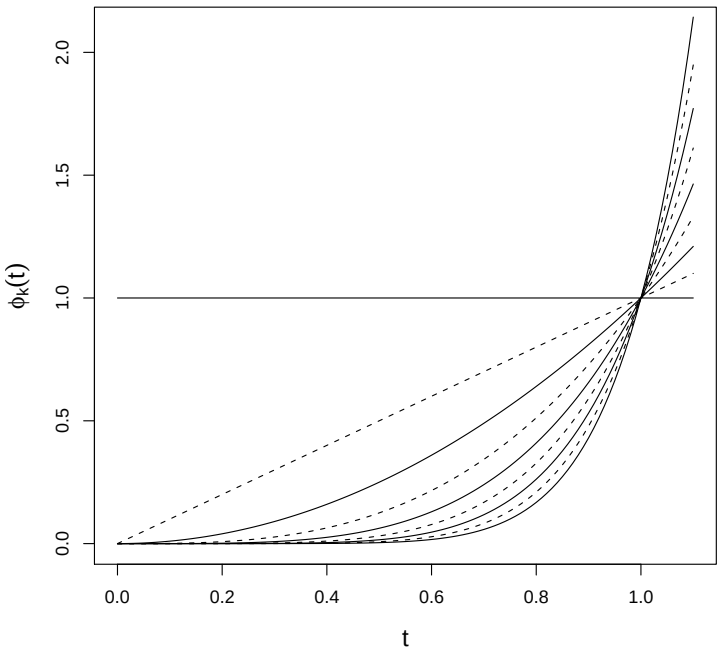


Figura 3.6: Dez funções base polinomiais com $\omega = 0$.

3.2.6 Funções Splines

Devido a sua estrutura simples e as suas boas propriedades de aproximação, os polinômios são amplamente utilizados na prática para aproximar funções. Geralmente o intervalo $\mathcal{T} = [a, b]$ é dividido em subintervalos suficientemente menores da forma $[x_0, x_1], \ldots, [x_k, x_{k+1}]$ e então um polinômio de grau menor p_i é usado para aproximação em cada intervalo $[x_i, x_{i+1}]$, $i = 0, \ldots, k$. Esse procedimento produz uma função de aproximação polinomial por partes $s(\cdot)$, ou seja, $s(t) = p_i(t)$ 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 (em 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 de polinômio $p_i(t)$ são constituídas independentemente umas das outras e, portanto, não formam uma função contínua $s(t)$ em $[a, b]$. Isso não pode ser aceito se alguém deseja, particularmente, aproximar uma função suave. Portanto, é necessário que as partes do polinômio sejam unidas suavemente nos nós interiores $x_1, \ldots, x_k$ 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

$$s(t) = \sum_{i=0}^{m-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i (t - x_i)_+^{m-1}, \tag{3.18}$$

onde os coeficientes $\theta_0, \ldots, \theta_{m-1}, \delta_1, \ldots, \delta_k$ são números reais e, dada uma função u, a *função de potência truncada* de grau r é definida como

$$u_+^r = \begin{cases} u^r & \text{se} \quad u \geq 0 \\ 0 & \text{se} \quad u < 0. \end{cases} \tag{3.19}$$

Assim, pode-se concluir que qualquer função *spline* é uma combinação linear de $m + k$ funções base. Considerando os nós interiores $\{x_1, \ldots, x_k\}$ as funções base são $\{1, t, t^2, \ldots, t^{m-1}, (t - x_1)_+^{m-1}, \ldots, (t - x_k)_+^{m-1}\}$, de acordo com (3.18).

Uma função *spline* $s(t)$ de ordem m com k nós interiores em $x_1, \ldots, x_k$ satisfaz as seguintes condições:

- $s(t)$ é um polinômio por partes, de grau $m - 1$ em qualquer subintervalo $[x_i, x_{i+1})$ para $i = 0, \ldots, k$;
- $s(t)$, $Ds(t)$,...,$D^{m-2}s(t)$ são funções contínuas em $\mathcal{T}$.

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 $\mathcal{S}_m(x_1, \ldots, x_k)$. Mais ainda, o espaço *spline* é um espaço linear de dimensão $m + k$ (Schumaker, 1981).

Seria interessante se existissem funções base que facilitassem o cálculo das funções *spline*. Os chamados B-*splines* formam uma base de espaços *spline* (Schumaker, 1981). Os B-*splines* têm a importante propriedade computacional de ter suporte compacto, ou seja, ele é não nulo (de fato positivo) num intervalo pequeno e zero fora desse intervalo.

3.2.7 B-*splines*

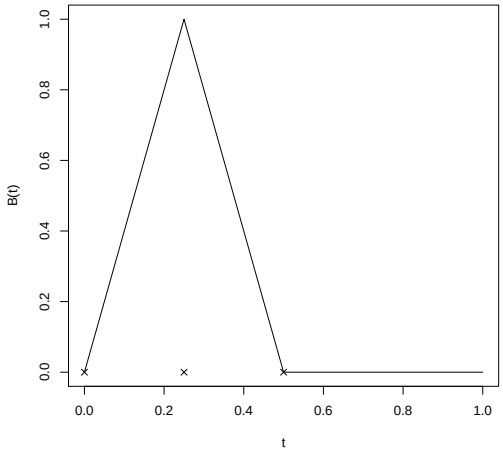
Os B-*splines* são constituídos de pedaços de polinômios unidos de forma especial em certos valores chamados nós. A escolha do número de nós tem sido um assunto de muita pesquisa: muitos nós produzem uma curva sem quase nenhuma suavidade (interpolação), já poucos nós suavizam demais os dados.

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. À esquerda de x_0 e à direita de x_2 esse *B-spline* é zero (ver Figura 3.7-(a)). É claro que é possível construir um conjunto tão amplo de B-*splines* quanto se queira, basta introduzir mais nós. Na Figura 3.7-(b) tem-se todos os B-*splines* possíveis de grau 1 no intervalo $[0, 1]$ com nós em $\{0; 0,25; 0,5; 0,75; 1\}$.

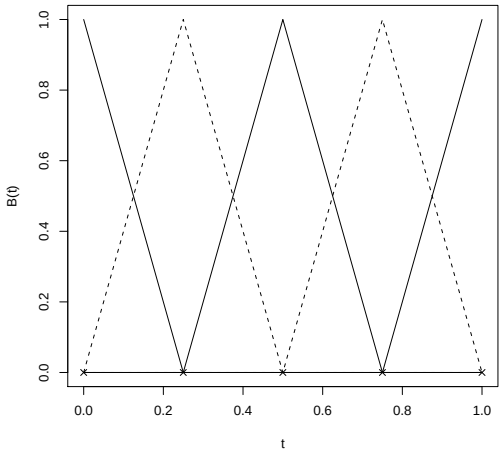
```
library(fda)
baseBspline <- create.bspline.basis(c(0,1),norder=2,nbasis=5)
t <- seq(from=0,to=1,by=0.01)
(knots(baseBspline))
matrizBases <- getbasismatrix(t, baseBspline, nderiv=0)
plot(t,matrizBases[,2],type='l',lty=1,ylim=c(0,1),
 xlab='t',ylab='B(t)')
points(c(0,0.25,0.5),rep(0,3),pch=4)
abline(h=0)
plot(t,matrizBases[,1],type='l',lty=1,ylim=c(0,1),
 xlab='t',ylab='B(t)')
for (i in 1:2){
lines(t,matrizBases[,i*2],lty=2)
lines(t,matrizBases[,(i*2)+1],lty=1)
}
points(seq(from=0,to=1,by=0.25),rep(0,5),pch=4)
```

Por sua vez, a Figura 3.8-(a) apresenta um B-*spline* cúbico (ordem=4) que consiste de quatro pedaços de polinômios cúbicos, unidos em três nós interiores. Nos pontos de união não apenas as ordens dos pedaços de polinômios se encaixam, as suas primeiras e segundas derivadas também são iguais (mas não as terceiras derivadas). Na Figura 3.8-(b) tem-se todos os B-*splines* possíveis de grau 3 no intervalo $[0, 1]$, com nós em $\{0; 0,2; 0,4; 0,6; 0,8; 1\}$. Os B-splines cúbicos são amplamente utilizados em regressão não-paramétrica, uma vez que uma combinação linear deles resulta em uma curva suave.

Os B-*splines* se sobrepõem uns aos outros. B-*splines* de primeiro grau sobrepõem dois vizinhos, B-*splines* cúbicos seis vizinhos e assim por diante.

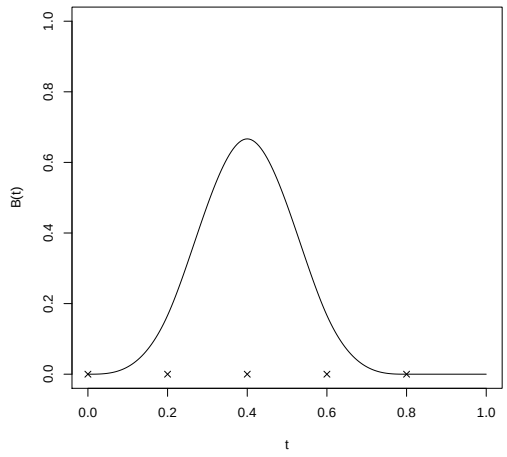


(a)

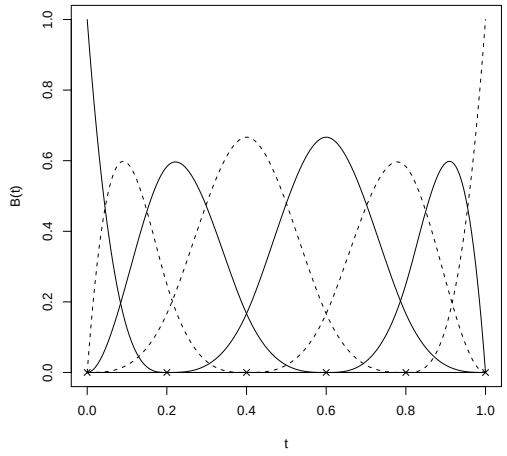


(b)

Figura 3.7: (a) B-*spline* de grau 1 isolado com três nós em "x". (b) B-*splines* de grau 1 com nós posicionados em "x".



(a)



(b)

Figura 3.8: (a) B-*spline* cúbico isolado com cinco nós em "x". (b) B-*splines* cúbicos com nós posicionados em "x".

É claro que, os B-*splines* mais à esquerda e mais à direita possuem menos sobreposições (ver Figuras 3.7-(b) e 3.8-(b)).

De forma geral, um B-*spline* de ordem m possui as seguintes propriedades:

- consiste de m pedaços de polinômio, cada um de grau $m-1$;
- os pedaços de polinômio se juntam em $m-1$ nós internos;
- nos pontos de união, as derivadas de ordem até $m-2$ são contínuas;
- o B-*spline* é positivo no domínio abrangido por $m+1$ nós, fora disso ele é zero;
- exceto nas fronteiras, ele sobrepõem $2(m-1)$ pedaços de polinômio de seus vizinhos;
- dado um certo t, m B-*splines* são não-nulos.

Mais precisamente, um B-*spline* pode ser definido como uma diferença dividida da função de potência truncada. Diferenças divididas surgem da forma de Newton para interpolação, onde para $n+1$ pares $\{t_i, f(t_i)\}_{i=0}^{n}$ um polinômio de grau n é aplicado da seguinte maneira:

$$p_n(t) = c_0 + c_1(t-t_0) + c_2(t-t_0)(t-t_1) + \cdots + c_n(t-t_0)(t-t_1)\cdots(t-t_{n-1}), \tag{3.20}$$

com

$$\begin{aligned} p_0(t_0) &= c_0 = f(t_0) \\ p_1(t_1) &= c_0 + c_1(t_1-t_0) = f(t_1) \\ p_2(t_2) &= c_0 + c_1(t_2-t_0) + c_2(t_2-t_0)(t_2-t_1) = f(t_2) \\ &\text{etc.} \end{aligned} \tag{3.21}$$

O coeficiente c_0 é dado por $c_0 = f(t_0) = [t_0]f$. É dito que $[t_0]f$ é a diferença dividida de ordem zero da função $f(t)$ sobre o ponto t_0. Por sua vez,

$$c_1 = \frac{f(t_1)-f(t_0)}{(t_1-t_0)} = \frac{[t_1]f - [t_0]f}{t_1-t_0} = [t_0, t_1]f$$

é a diferença dividida de ordem 1 da função $f(t)$ sobre os pontos t_0 e t_1.

Dessa forma podemos calcular todos os coeficientes usando a diferença dividida de ordem k da função $f(t)$ sobre os pontos $t_0, t_1, \ldots, t_k$, ou seja,

$$c_k = [t_0, \ldots, t_k]f, \ k = 0, \ldots, n. \tag{3.22}$$

Pode-se mostrar que (3.22) pode ser obtida de forma recursiva (ver Ruggiero, 1997) usando

$$[t_0, t_1, \ldots, t_k]f = \frac{[t_1, \ldots., t_k]f - [t_0, \ldots, t_{k-1}]f}{t_k - t_0} \tag{3.23}$$

para $k = 1, \ldots, n$, já que $[t_0]f = f(t_0)$.

Agora, com a definição de diferenças divididas, é possível definir melhor um B-*spline*. Para uma sequência crescente de nós $x = \{x_i\}$ o i-ésimo B-*spline* de ordem m é dado por:

$$B_{i,m}(t) = (x_{i+m} - x_i)[x_i, \ldots, x_{i+m}](\bullet - t)_+^{m-1} \text{ para todo } t \in \mathbb{R}. \tag{3.24}$$

A notação $(\bullet - t)_+^{m-1}$ é usada aqui para indicar que a m-ésima diferença dividida da função $(x-t)_+^{m-1}$ de duas variáveis x e t deve ser feita fixando t e considerando $(x-t)_+^{m-1}$ como uma função somente de x. O número resultante depende, é claro, do valor particular de t que foi escolhido, ou seja, o número resultante varia conforme t varia e, então, obtém-se por fim, a função $B_{i,m}$ de t.

O fator $(x_{i+m} - x_i)$ é um fator de normalização usado para que se tenha a seguinte propriedade:

$$\sum_i B_{i,m}(t) = 1 \ \forall \ t \in \mathcal{T}. \tag{3.25}$$

A dimensão de uma base B-*spline* é maior que a dimensão de uma base de *splines*.

Seja o intervalo $\mathcal{T} = [a, b]$ dividido em $k+1$ subintervalos menores da forma $[x_0, x_1], \ldots, [x_k, \quad x_{k+1}]$ por $k + 2$ nós, sendo $x_1, \ldots, x_k$ os k nós interiores. 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 Shumaker (1981)). Para isso, $m - 1$ nós são adicionados no início e no final da sequência de tal forma que $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_{m-1} \leq x_0$ e $x_{k+1} \leq \tau_{m+k+2} \leq \tau_{m+k+3} \leq \cdots \leq \tau_{k+2m}$. Os

valores desses nós adicionais são arbitrários. É comum, e em ADF também, fazer com que $\tau_1 = \cdots = \tau_{m-1} = x_0$ e $x_{k+1} = \tau_{m+k+2} = \cdots = \tau_{k+2m}$. Isso foi feito para obter as Figuras 3.7-(b) e 3.8-(b). O número de B-*splines* na regressão é igual ao número de nós interiores mais a ordem do polinômio, ou seja, $K = k + m$.

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. Esse algoritmo funciona quando os nós são igualmente espaçados ou também quando não são.

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*

$$B_{i,m}(t) = \frac{t - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1} + \frac{\tau_{i+m} - t}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(t), \tag{3.26}$$

onde

$$B_{i,1}(t) = \begin{cases} 1 & \text{se} \quad \tau_i \leq t \leq \tau_{i+1} \\ 0 & \text{caso contrário.} \end{cases} \tag{3.27}$$

Os B-*splines* podem ter também múltiplos nós e, assim, é necessário ter um certo cuidado ao usar a relação de recorrência (3.27) 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

$$x(t) = \sum_{i=1}^{K=m+k} c_i B_{i,m}(t). \tag{3.28}$$

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

$$\hat{x}(t) = \sum_{i=1}^{K} \hat{c}_i B_i(t). \tag{3.29}$$

Em geral a ordem do B-*spline* é clara pelo contexto. Dessa forma, no lugar de $B_{i,m}(t)$ pode-se escrever somente $B_i(t)$ como em (3.29).

Assim como foi feito para os B-*splines*, uma relação de recorrência também pode ser utilizada para o cálculo de suas derivadas. Assim, a r-ésima derivada ($r < m$) do B-*spline* de ordem m pode ser computada recursivamente através de

$$D^r B_{i,m}(t) = (m-1)\left\{\frac{D^{r-1}B_{i,m-1}(t)}{\tau_{i+m-1}-\tau_i} - \frac{D^{r-1}B_{i+1,m-1}(t)}{\tau_{i+m}-\tau_{i+1}}\right\}. \tag{3.30}$$

Prova. Usando (3.24) tem-se que

$$DB_{i,m}(t) = -(m-1)(\tau_{i+m}-\tau_i)[\tau_i,\ldots,\tau_{i+m}](\bullet - t)_+^{m-2}. \tag{3.31}$$

De (3.23) sabe-se que

$$[\tau_i,\ldots,\tau_{i+m}](\bullet - t)_+^{m-2} = \frac{[\tau_{i+1},\ldots,\tau_{i+m}](\bullet - t)_+^{m-2} - [\tau_i,\ldots,\tau_{i+m-1}](\bullet - t)_+^{m-2}}{\tau_{i+m}-\tau_i}, \tag{3.32}$$

e usando (3.24) em (3.32), (3.31) se torna

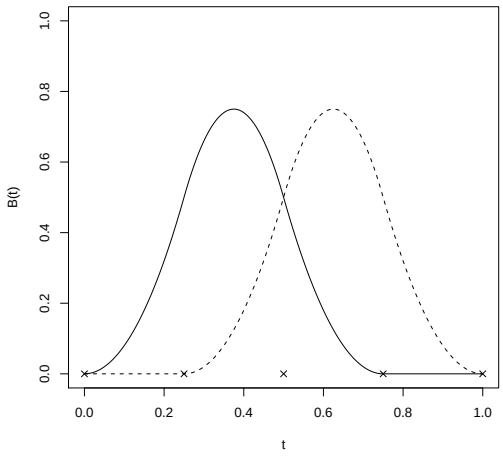
$$DB_{i,m}(t) = (m-1)\left\{\frac{B_{i,m-1}(t)}{\tau_{i+m-1}-\tau_i} - \frac{B_{i+1,m-1}(t)}{\tau_{i+m}-\tau_{i+1}}\right\}.$$

A primeira derivada de um B-*spline* é um outro B-*spline* de ordem uma vez menor. Recursivamente, aplicando essa técnica é possível calcular derivadas de ordem maiores, uma vez que

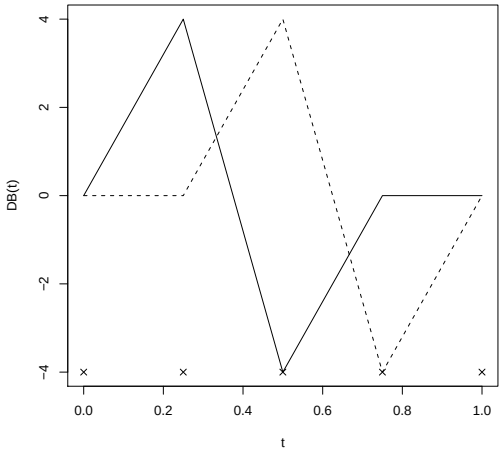
$$D^r B_{i,m}(t) = D\{D^{r-1}B_{i,m-1}(t)\}.$$

□

A Figura 3.9-(a) apresenta dois B-*splines* de ordem 3 e a Figura 3.9-(b) ilustra suas primeiras derivadas. Já na Figura 3.10 é possível observar as suas segundas derivadas. Um B-*spline* de ordem m é formado por pedaços polinomias de ordem m. Portanto, é óbvio que sua r-ésima derivada ($r < m$) também seja uma função polinomial por partes de ordem $m - r$.



(a)



(b)

Figura 3.9: (a) $B_{1,3}(t)$ e $B_{2,3}(t)$, dois B-*splines* de ordem 3 com nós em "x". (b) $DB_{1,3}(t)$ e $DB_{2,3}(t)$. As primeiras derivadas são funções polinomiais por partes de ordem 2.

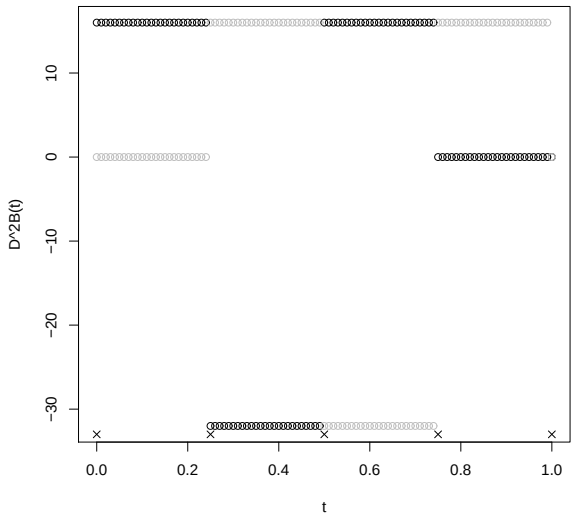


Figura 3.10: $D^2B_{1,3}(t)$ e $D^2B_{2,3}(t)$. As segundas derivadas são funções constantes por partes, ou seja, são funções polinomiais por partes de ordem 1.

```
library(fda)
baseBspline <- create.bspline.basis(c(0,1),norder=3,nbasis=6)
t <- seq(from=0,to=1,by=0.005)
matrizBases <- getbasismatrix(t, baseBspline, nderiv=0)
plot(t,matrizBases[,3],type='l',ylim=c(0,1),xlab='t',ylab='B(t)')
lines(t,matrizBases[,4],lty=2)
points(seq(0,1,by=0.25),rep(0,5),pch=4)
matrizDBases <- getbasismatrix(t, baseBspline, nderiv=1)
plot(t,matrizDBases[,3],type='l',lty=1,xlab='t',ylab='DB(t)')
lines(t,matrizDBases[,4],lty=2)
points(seq(0,1,by=0.25),rep(-4,5),pch=4)
```

Mais detalhes sobre os polinômios *splines*, em geral, e B-splines, em paticular, podem ser encontrados no clássico De Boor (1978), Eubank (1988) e em Green e Silverman (1994).

3.2.8 Bases *Wavelet* (ou ondaleta)

Pode-se construir uma base para todas as funções em $(-\infty, \infty)$ que são de quadrado integráveis, escolhendo uma função *wavelet* mãe ψ e, assim, considerando todas as dilatações e translações da forma

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k), \tag{3.33}$$

para inteiros j e k. A *wavelet* mãe é construída de forma a assegurar que a base é ortogonal. Ela tipicamente possui suporte compacto, e, por consequência, todas as funções base da forma (3.33) também. A ideia da base *wavelet* é facilmente adaptada para lidar com funções definidas em um intervalo limitado.

Suponha que uma função x é observada com ruído. Muitas classes intuitivamente atrativas de funções têm econômicas expansões *wavelet*. Isso leva a uma simples abordagem de suavização não linear. Primeiro se constrói a transformada discreta *wavelet* das observações com ruído e, depois, se faz uma mudança retirando os coeficientes pequenos na expansão e possivelmente reduzindo os grandes. A motivação básica para essa mudança é a noção que qualquer coeficiente pequeno é puramente ruído e não reflete qualquer sinal. Isso indica que os estimadores obtidos dessa forma devem ser muito adaptativos a diferentes graus de suavidade e regularidade na função a ser estimada.

Mais detalhes e outras referências sobre as bases *wavelets* podem ser encontrados em Ramsay e Silverman (1997).

3.2.9 A escolha do número K de funções base

- **O dilema viés/variância**

Para valores grandes de K, o viés na estimação de $x(t)$, ou seja,

$$B[\hat{x}(t)] = x(t) - E[\hat{x}(t)] \tag{3.34}$$

é pequeno. Mas isso é somente uma parte do problema. Uma das razões pelas quais se faz suavização é reduzir a influência de ruído ou variação na estimativa $\hat{x}$. Dessa forma, existe também interesse na variância da estimativa

$$Var[\hat{x}(t)] = E\{[\hat{x}(t) - E(\hat{x}(t))]^2\}. \tag{3.35}$$

Se $K = n$, a variância será quase que certamente muito alta. Reduzir a variância significa optar por valores pequenos de K, mas não tão pequenos de forma que o vício se torne inaceitável. Uma maneira de expressar o que se quer alcançar é através do *erro quadrático médio*

$$EQM[\hat{x}(t)] = E[\hat{x}(t) - x(t)]^2, \tag{3.36}$$

também chamado de função perda $\mathcal{L}^2$. Na maioria das aplicações, não é possível minimizar (3.36) diretamente, pois $x(t)$ não é conhecida. Porém, uma das equações mais importantes da estatística relaciona o erro quadrático médio com o vício e a variância de $\hat{x}(t)$. Tal equação é dada por

$$EQM[\hat{x}(t)] = \{B[\hat{x}(t)]\}^2 + Var[\hat{x}(t)]. \tag{3.37}$$

Essa relação mostra que seria válido tolerar um pouco de vício se o resultado for uma grande redução na variância amostral. Na verdade, isso é o que quase sempre acontece, e é a razão fundamental de suavizar os dados para estimar funções. Essa questão será novamente abordada no Capítulo 4.

O aumento de K nem sempre faz o vício diminuir. Quando a ordem de um *spline* é fixa e os nós são igualmente espaçados, efeitos complicados devido ao espaçamento dos nós em relação aos pontos amostrais podem resultar em

um sistema B-*spline* de menor dimensão, que por sua vez produz melhores resultados do que um sitema de maior dimensão.

Embora a decomposição através do erro quadrático médio (3.37) seja útil para expressar o dilema viés/variância de forma clara, o princípio se aplica de maneira mais ampla. Na verdade, existem situações onde é preferível utilizar outras funções perda. Por exemplo, minimizar $E[|\hat{x}(t) - x(t)|]$ é mais efetivo quando os dados contêm valores aberrantes. Para esse e para quase todo critério de ajuste ou função perda para suavização, pode-se assumir que quando o vício diminui, a variância cresce, e que um pouco de vício deve ser aceito para obter uma estimativa equilibrada da tendência suave dos dados.

- **Algoritmos para escolha de K**

A vasta literatura em regressão múltipla contém muitas ideias para decidir quantas funções base usar. Por exemplo, o método de seleção *stepwise* poderia ser adotado. Funções base seriam adicionadas uma depois da outra, testando em cada passo se a função adicionada melhora significantemente o modelo e, também, se as funções já presentes continuam sendo significantes. No sentido inverso, existem métodos que são usados para modelos de alta dimensão. Eles funcionam começando com uma escolha grande de K e, depois, uma função base que não contribui substancialmente com a quantidade de variação é retirada do modelo em cada passo.

O fato de não existir nenhum método padrão perfeito para o problema de seleção de variáveis leva à difícil tarefa de tentar fixar a dimensão do modelo. O caráter discreto do problema da escolha de K tem parcialmente a culpa por isso. Os métodos descritos no Capítulo 4 provendo níveis de suavização no contínuo serão muito úteis.

Um método adaptativo para a escolha do número de funções de base via penalização estocástica pode ser encontrado em Anselmo et al. (2005).

3.3 Calculando a variância amostral e limites de confiança

3.3.1 Variância Amostral

Primeiramente, considere o vetor aleatório $\mathbf{y}$ com matriz de variância e covariância $\mathbf{\Sigma}_y$. Assim, a variável aleatória $\mathbf{Ay}$ definida por qualquer matriz $\mathbf{A}$

tem a seguinte matriz de variância e covariância

$$\widehat{Var}[\mathbf{Ay}] = \mathbf{A}\hat{\mathbf{\Sigma}}_y\mathbf{A}'. \tag{3.38}$$

A matriz de variância e covariância de $\mathbf{y}$ usando o modelo $\mathbf{y} = x(\mathbf{t}) + \boldsymbol{\epsilon}$ é a matriz de variância e covariância $\mathbf{\Sigma}_e$ do vetor de erros $\boldsymbol{\epsilon}$, já que $x(\mathbf{t})$ é um efeito fixo com variância zero. De alguma forma, a informação presente nos resíduos deve ser usada para substituir a quantidade populacional $\mathbf{\Sigma}_e$ por uma estimativa amostral razoável $\hat{\mathbf{\Sigma}}_e$.

Por exemplo, para calcular as variâncias e covariâncias amostrais dos coeficientes em $\hat{\mathbf{c}}$ pode-se definir

$$\mathbf{A} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{W},$$

e usando (3.38) obtém-se

$$\widehat{Var}[\hat{\mathbf{c}}] = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{W}\hat{\mathbf{\Sigma}}_e\mathbf{W}\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi})^{-1}. \tag{3.39}$$

Quando a suposição padrão para os erros é assumida, ou seja, $\mathbf{\Sigma}_e = \sigma^2\mathbf{I}$ e não for utilizado mínimos quadrados ponderados, um resultado mais simples presente nos livros de análise de regressão é obtido:

$$\widehat{Var}[\hat{\mathbf{c}}] = \hat{\sigma}^2(\mathbf{\Phi}'\mathbf{\Phi})^{-1}. \tag{3.40}$$

Contudo, no contexto de análise de dados funcionais raramente haverá muito interesse no vetor de coeficientes propriamente dito. Ao invés disso, deseja-se estudar a variância amostral das quantidades calculadas a partir desses coeficientes, como, por exemplo, a variância amostral do ajuste dos dados definido por $\hat{x}(t) = \boldsymbol{\phi}(t)'\hat{\mathbf{c}}$. Uma vez que se tem em mãos a variância amostral de $\hat{\mathbf{c}}$ através de (3.39) ou (3.40), basta aplicar mais uma vez o resultado de (3.38) e assim obter

$$\widehat{Var}[\hat{x}(t)] = \boldsymbol{\phi}(t)'\widehat{Var}[\hat{\mathbf{c}}]\boldsymbol{\phi}(t). \tag{3.41}$$

Por sua vez, as variâncias de todos os valores ajustados correspondentes aos valores amostrais t_j estão na diagonal da matriz

$$\widehat{Var}[\hat{\mathbf{y}}] = \mathbf{\Phi}\widehat{Var}[\hat{\mathbf{c}}]\mathbf{\Phi}',$$

e assumindo a suposição padrão e utilizando mínimos quadrados simples, tem-se que

$$\widehat{Var}[\hat{\mathbf{y}}] = \hat{\sigma}^2\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}' = \hat{\sigma}^2\mathbf{S}.$$

3.3.2 Estimando $\mathbf{\Sigma}_e$

É claro que as estimativas das variâncias amostrais descritas anteriormente serão boas se as estimativas das variâncias e covariâncias entre os resíduos ϵ_j forem boas também.

Quando uma única curva é suavizada, a quantidade de informação envolvida é somente suficiente para estimar uma simples variância constante σ^2, assumindo a suposição padrão para os erros, ou no máximo, uma função variância com valores $\sigma^2(t)$ que tenha variações suaves ao longo de t. É importante usar métodos que produzam uma estimativa não viesada da variância para evitar uma subestimação da variância amostral. Por exemplo, se a suposição padrão para os erros for aceitável,

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-K}\sum_{j=1}^{n}[y_j - \hat{y}_j]^2 \tag{3.42}$$

é preferida como sendo uma melhor estimativa de σ^2 do que a estimativa de máxima verossimilhança, que envolve dividir por n.

Uma estratégia comum para estimar pelo menos um número limitado de covariâncias em $\mathbf{\Sigma}_e$ dado um número N pequeno de curvas, ou até mesmo $N = 1$, é assumir um modelo auto-regressivo (AR) para os resíduos. Isso é frequentemente razoável, já que resíduos adjacentes são geralmente correlacionados devido à influência de variáveis não observadas. Para consultar detalhes sobre como estimar a estrutura AR entre os resíduos ver Draper e Smith (1998).

Quando um número substancial de replicações N está disponível, é possível calcular estimativas mais sofisticadas e detalhadas de $\mathbf{\Sigma}_e$. Por exemplo, pode-se optar por estimar toda a matriz de variância e covariância a partir da matriz $N \times n$ $\mathbf{R}$ de resíduos. Assim, tem-se que

$$\hat{\mathbf{\Sigma}}_e = (N-1)^{-1}\mathbf{R}'\mathbf{R}.$$

Contudo, mesmo assim, a estimativa de toda a matriz $\mathbf{\Sigma}_e$ requer a estimação de $n(n+1)/2$ variâncias e covariâncias.

3.3.3 Limites de confiança

Os limites de confiança são geralmente calculados somando e subtraindo um múltiplo dos desvios padrão, que são a raiz quadrada das variâncias amostrais do ajuste em que se está trabalhando. Por exemplo, limites de confiança de 95% correspondem a aproximadamente dois desvios padrão acima e abaixo do ajuste suave. Esses desvios padrão são estimados usando (3.41). Os limites de confiança calculados dessa forma são chamados de *pontuais*, pois eles refletem regiões de confiança para valores fixos de t, ao invés de regiões para a curva como um todo.

A Figura 3.11 mostra as temperaturas durante os 16 dias de degelo em Montreal no mês de janeiro, junto com a suavização dos dados e os limites de confiança pontuais de 95% para o ajuste. Os desvios padrão estimados para os valores ajustados foram de aproximadamente 0,26 grau Celsius.

```
phi <- getbasismatrix(t, basesFourier, nderiv=0)
sigma2hat <- (sum((montreal-montrealfit)^2))/(365-95)
varcoef <- sigma2hat*solve(t(phi)%*%phi)
varyfit <- phi%*%varcoef%*%t(phi)
sdmontrealfit<-sqrt(diag(varyfit))
limiteSup <- montrealfit+2*sdmontrealfit
limiteInf <- montrealfit-2*sdmontrealfit
```

É importante prestar atenção a dois casos nos quais os limites de confiança calculados dessa forma podem ser problemáticos. Primeiro, está implícito que o número de coeficientes K é tido como uma constante fixa, mas na realidade K poderia ser mais um parâmetro a ser estimado nos problemas de suavização, e, consequentemente, os tamanhos desses limites de confiança não refletem a incerteza no conhecimento de K. Segundo, a própria curva suave na qual se adiciona e se subtrai múltiplos dos desvios padrão está sujeita a vício, especialmente nas regiões de maior curvatura. Assim, os limites de confiança calculados dessa forma serão também viciados e a região coberta por eles pode não ser totalmente da forma como é apresentada.

É possível obter uma região de confiança para toda a função através de métodos computacionalmente intensivos como *bootstrap* (Efron e Tibshirani, 1993).

Mais detalhes sobre limites de confiança podem ser encontrados em Ramsay e Silverman (2006).

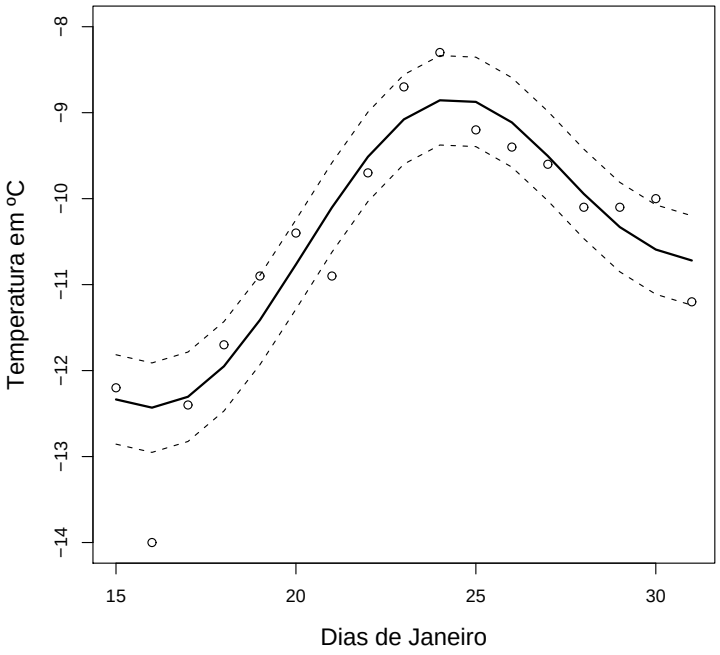


Figura 3.11: Os pontos correspondem as temperaturas de inverno durante o período de degelo em Montreal. A linha sólida é a curva estimada suave referente à Figura 3.5. As linhas tracejadas correspondem aos limites de confiança pontuais de 95%.

Tabela 3.1: *Kernels*

Kernel	$Kern(u)$
Uniforme	$\frac{1}{2}I(\|u\| \leq 1)$
Epanechnikov	$0.75(1-u^2)I(\|u\| \leq 1)$
Gaussiano	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$

3.4 Suavização por ponderação local

3.4.1 Funções *kernel* (ou núcleo)

Para algum método de suavização fazer sentido, o valor da estimativa da função em um ponto t deve ser influenciado pelas observações próximas de t. Essa característica é uma propriedade implícita dos estimadores que foram considerados até aqui. Nessa seção, serão considerados estimadores onde a dependência local é feita de forma mais explícita através de médias ponderadas com funções peso locais. Dentro do contexto de suavização linear, isso significa que $\hat{x}(t) = \sum_j w_j(t)y_j$.

Os pesos w_j são simplesmente construídos através de uma mudança de escala e locação de uma função *kernel* com valores $Kern(u)$. A função *kernel* é construída de tal forma que a maior parte da sua massa esteja concentrada perto de 0, e que decaia rapidamente ou desapareça por completo quando $|u| \geq 1$. A Tabela 3.1 apresenta três *kernels* que são muito utilizados.

É possível definir os seguintes valores de peso:

$$w_j(t) = Kern\left(\frac{t_j - t}{h}\right). \tag{3.43}$$

Valores substanciais de $w_j(t)$ como uma função de j estão agora concentrados para t_j aos arredores de t, e o grau de concentração é controlado pelo tamanho de h. O parâmetro de concentração h, também denominado como parâmetro de suavização, é comumente chamado de janela (em inglês, *bandwidth*). Valores pequenos de h implicam que somente observações perto de t recebem algum peso, enquanto que h grande significa que mesmo valores a uma distância considerável de t serão utilizados. Note que se o *kernel* é uma

função densidade de probabilidade, então h é o parâmetro de escala no sentido estatístico do termo.

3.4.2 Suavização por *kernel*

O caso mais simples e clássico de um estimador que usa pesos locais é o estimador por *kernel*. Como visto em (3.5), a estimativa em um dado ponto é uma combinação linear das observações locais, ou seja,

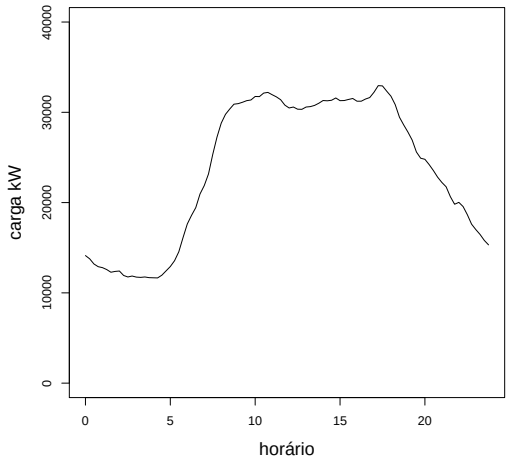
$$\hat{x}(t) = \sum_{j=1}^{n} S_j(t) y_j,$$

onde S_j são funções peso apropriadas. O estimador de Nadaraya-Watson é calculado usando

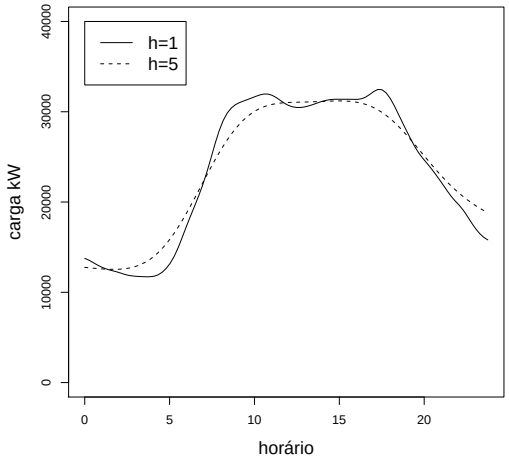
$$S_j(t) = \frac{Kern[(t_j - t)/h]}{\sum_r Kern[(t_r - t)/h]}, \tag{3.44}$$

ou seja, os pesos $w_j(t)$ são normalizados para ter a soma igual à um. A Figura 3.12-(a) apresenta o consumo de energia elétrica da região central de Campinas em kW ao longo de um dia. Por sua vez, a Figura 3.12-(b) mostra os ajustes suaves obtidos através do estimador de Nadaraya-Watson para diferentes valores do parâmetro de suavização h.

```
x <- read.table('campinas1.txt'); t <- x[,1]
plot(t,x[ ,15],ylim=c(0,40000),xlab='horário',ylab='carga kW',
 col=1,type='l')
fit1 <- ksmooth(t,x[,15],"normal",bandwidth=1)
plot(fit1,ylim=c(0,40000),xlab='horário',ylab='carga kW',
     col=1,type='l')
fit2 <- ksmooth(t,x[,15],"normal",bandwidth=5)
lines(fit2,col=1,lty=2)
legend(0,40000,c("h=1","h=5"),lty=c(1,2),col=c(1,1))
```



(a)



(b)

Figura 3.12: (a) Consumo de energia elétrica da região central de Campinas em kW ao longo de um dia. (b) Curvas suaves estimadas utilizando o estimador de Nadaraya-Watson para $h = 1$ e $h = 5$.

Existem também os pesos desenvolvidos por Gasser e Müller (1979, 1984). Eles são construídos da seguinte forma:

$$S_j(t) = \frac{1}{h} \int_{\bar{t}_{j-1}}^{\bar{t}_j} Kern\left(\frac{u-t}{h}\right) du, \tag{3.45}$$

onde $\bar{t}_j = (t_{j+1} + t_j)/2$, $1 < j < n$, $\bar{t}_0 = t_1$ e $\bar{t}_n = t_n$. Esses pesos são calculados de forma mais rápida, lidam de forma mais sensível com argumentos não igualmente espaçados e possuem boas propriedades assintóticas.

3.4.3 Estimadores por função base localizados

As ideias dos estimadores de *kernel* e dos estimadores que usam funções base podem, de certo modo, serem combinadas para produzir os chamados estimadores por função base localizados, que abrangem uma classe ampla de estimadores para funções e derivadas. A ideia básica é estender o critério de mínimos quadrados (3.9) de forma a obter uma medida de erro local. Assim, é possível definir

$$SQE_t = \sum_{j=1}^{n} w_j(t) \left[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2, \tag{3.46}$$

onde as funções peso w_j são construídas usando a função *kernel* (3.43). Em termos matriciais,

$$SQE_t = (\mathbf{y} - \mathbf{\Phi c})' \mathbf{W}(t) (\mathbf{y} - \mathbf{\Phi c}), \tag{3.47}$$

onde $\mathbf{W}(t)$ é uma matriz diagonal contendo os valores $w_j(t)$ na sua diagonal. O vetor de coeficientes $\hat{\mathbf{c}}$ que minimiza SQE_t é dado por

$$\hat{\mathbf{c}} = [\mathbf{\Phi}' \mathbf{W}(t) \mathbf{\Phi}]^{-1} \mathbf{\Phi}' \mathbf{W}(t) \mathbf{y}.$$

Dessa forma, $\hat{x}(t) = \sum_{k=1}^{K} \hat{c}_k \phi_k(t)$ produz um estimador de suavização linear da forma (3.5), com os pesos $S_j(t)$ sendo os elementos do vetor

$$\mathbf{s}(t) = \mathbf{W}(t) \mathbf{\Phi} [\mathbf{\Phi}' \mathbf{W}(t) \mathbf{\Phi}]^{-1} \boldsymbol{\phi}(t), \tag{3.48}$$

onde $\boldsymbol{\phi}(t)$ é o vetor com elementos $\phi_k(t)$.

Os pesos $w_j(t)$ em (3.46) são construídos de forma a possuir valores consideravelmente diferentes de zero somente para observações próximas do argumento t, no qual a função será estimada. Isso implica que somente os elementos de $\mathbf{s}(t)$ em (3.48) associados com valores de t_j próximos do valor t são significantemente diferentes de zero. Consequentemente, $\hat{x}(t)$ é essencialmente uma combinação linear de observações y_j na vizinhança de t.

Uma vez que a base tem que aproximar somente um segmento limitado de dados ao redor de t, a base pode fazer um melhor trabalho na aproximação de características locais dos dados, e, ao mesmo tempo, espera-se obter uma boa estimação usando um número pequeno K de funções base. Porém, o preço pago por essa flexibilidade é que a expansão deve ser calculada para cada ponto de avaliação t.

É interessante notar que a estimativa de Nadaraya-Watson pode ser obtida como um caso especial desse método considerando $K = 1$ e $\phi_1(t) = 1$ em (3.47).

3.4.4 Seleção de h

A janela h controla o equilíbrio entre o viés e a variância na estimação por *kernel*. Valores pequenos de h implicam que o valor esperado do estimador $\hat{x}(t)$ deve ser próximo do valor verdadeiro $x(t)$, mas, em compensação, haverá uma alta variabilidade na estimativa. Por outro lado, a variabilidade pode ser sempre reduzida aumentando o valor de h. Nesse caso, quanto maior o valor de h, mais suave será a estimativa e, assim, mais viés ela também apresentará. O erro quadrático médio (3.37) do estimador em t é uma maneira mais adequada de medir o desempenho do estimador.

Há uma variedade de técnicas automáticas de seleção de h, geralmente baseadas na minimização do erro quadrático médio (ver Härdle, 1990). Porém, nenhuma dessas técnicas é sempre confiável. Algoritmos de seleção de h continuam sendo objeto de muitos estudos. Muitas vezes na prática o que se faz é escolher dentre uma variedade de valores de h aquele que produz graficamente o melhor resultado.

Capítulo 4

O método de penalização da não suavidade

4.1 Introdução

Nesse capítulo será apresentada uma outra opção para a aproximação de dados discretos por uma função. O método de penalização da não suavidade ou método de regularização possui as mesmas vantagens das técnicas de suavização apresentadas no Capítulo 3, mas evita algumas limitações presentes nessas técnicas. Tal método também se adapta a uma variedade maior de problemas que envolvem análise de dados funcionais.

Um bom ajuste aos dados não é o único objetivo ao estimar uma curva. Existe um outro objetivo, geralmente conflitante, que é obter uma estimativa que não oscile muito rapidamente. O princípio básico do método de penalização da não suavidade é quantificar a noção de quão rapidamente uma curva oscila e, então, representar o problema de estimação de maneira que o compromisso entre esses dois objetivos seja explícito.

4.2 A suavização *spline*

O objetivo dessa seção é estudar como a regularização funciona no caso funcional mais simples, quando o objetivo é estimar uma função não periódica x com base em observações discretas e com ruído dispostas em um vetor $\mathbf{y}$. O

problema de suavização de dados apresentado no Capítulo 3 continua, porém o termo "suavização *spline*" será agora reservado para os casos onde a penalização da não suavidade é aplicada. Diferentes penalizações serão descritas nessa seção.

Estudos assintóticos sobre o estimador obtido utilizando o método de suavização *spline* podem ser encontrados em Eubank (1988). Nesse caso, é possível, entre outras coisas, verificar a consistência desse estimador.

Silverman (1984) 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.

4.2.1 Dois objetivos na estimação de uma função

O método de suavização *spline* estima uma curva x através das observações $y_j = x(t_j) + \epsilon_j$ 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 aos elementos do erro quadrático médio (EQM = Vício^2+Variância). Na suavização *spline*, como em outros métodos de suavização, o EQM é uma das maneiras de obter o que geralmente se quer dizer com qualidade da estimativa. Nota-se na Seção 3.2.9 que outras funções perda podem ser escolhidas em certas situações. A noção de um dilema entre viés e variância se aplica também a esses casos, embora não com a mesma decomposição do erro quadrático médio.

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 emprestar informação dos dados vizinhos. Desse modo fica expresso que uma certa regularidade na função latente x é esperada. Esse compartilhar de informação é o que faz a curva estimada ser mais estável, ao preço de um aumento no vício. A penalização da não suavidade torna explícito o que é sacrificado em vício para que o EQM ou outra função perda seja melhorada.

4.2.2 Quantificando a não suavidade

O quadrado da segunda derivada de uma função x em t, $[D^2x(t)]^2$, é frequentemente chamado de curvatura da função no ponto t. Assim, uma medida natural da não suavidade de uma função é a integral do quadrado da sua segunda derivada, ou seja,

$$PEN_2(x) = \int [D^2x(s)]^2 ds. \tag{4.1}$$

Esse critério 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 são grandes sob 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 (4.1) não é adequada, uma vez que ela controla a curvatura de x propriamente dita, e, por conseguinte, 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 $m + 2$, afim de controlar a curvatura de $D^m x$. Por exemplo, a estimativa de uma curva de aceleração será melhor se a seguinte penalização for utilizada:

$$PEN_4(x) = \int [D^4x(s)]^2 ds, \tag{4.2}$$

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

4.2.3 A soma dos quadrados dos erros penalizada

Agora é necessário modificar o critério de mínimos quadrados (3.12) para permitir que a penalização da não suavidade participe do cálculo da estimativa. Seja $x(\mathbf{t}) = [x(t_1), \ldots, x(t_n)]'$, a soma dos quadrados dos erros penalizada é definida como

$$SQEPEN_\lambda = [\mathbf{y} - x(\mathbf{t})]'\mathbf{W}[\mathbf{y} - x(\mathbf{t})] + \lambda PEN_2(x). \tag{4.3}$$

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

$$SQEPEN_\lambda = [\mathbf{y} - x(\mathbf{t})]'[\mathbf{y} - x(\mathbf{t})] + \lambda PEN_2(x). \tag{4.4}$$

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. 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 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.

A Figura 4.1 apresenta um exemplo de suavização *spline* utilizando diferentes parâmetros de suavização e B-*splines* cúbicos. Nota-se que a estimativa obtida com $\lambda = 0,01$ é muito suave e por isso não consegue descrever características importantes da função verdadeira. Já o ajuste obtido com um menor valor de λ apresenta um comportamento mais próximo da verdadeira curva x.

```
library(fda)
t <- seq(from=0.01,to=1,by=0.01) ; n <- length(t)
u <- exp(-4*t)*sin(pi*((4*t)/2))*cos(pi*(4*t))*4
y <- u + rnorm(n,mean=0,sd=0.08)
bases <- create.bspline.basis(range=c(min(t),max(t)),
 norder=4, breaks=t)
yfdPar1 <- fdPar(bases, Lfdobj=2, lambda = 0.01)
yfdPar2 <- fdPar(bases, Lfdobj=2, lambda = 0.0001)
yfd1 <- smooth.basis(t,y,yfdPar1)$fd
yfd2 <- smooth.basis(t,y,yfdPar2)$fd
```

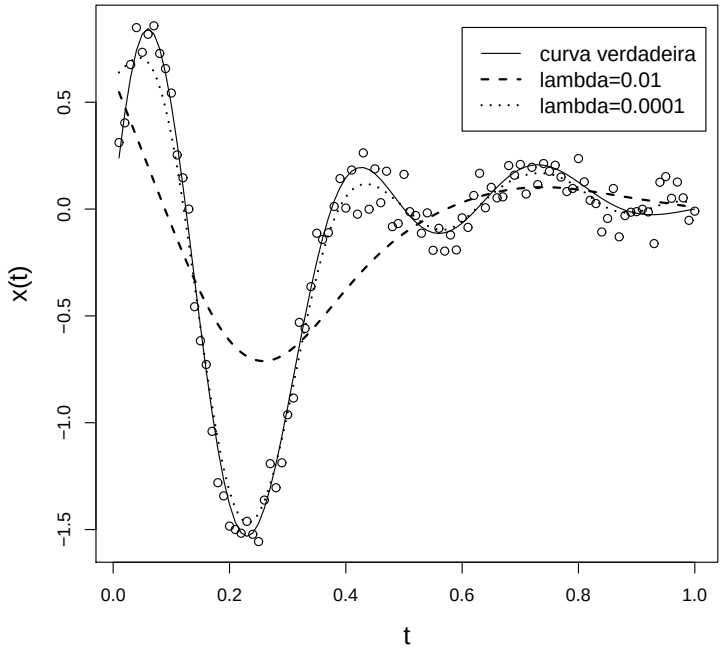


Figura 4.1: Suavização *spline* utilizando diferentes parâmetros de suavização λ. A curva verdadeira é $x(t) = 4\exp(-4t)\text{sen}(2\pi t)\cos(4\pi t)$.

```
tAval <- seq(min(t),max(t),length=500)
yhat1 <- eval.fd(tAval, yfd1)
yhat2 <- eval.fd(tAval, yfd2)

plot(t,y,ylab='x(t)',xlab='t',cex.lab=1.4)
lines(t,u)
lines(tAval,yhat1,lty=2,lw=2)
lines(tAval,yhat2,lty=3,lw=2)
```

4.2.4 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, encontrado em 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).

Posicionar os nós nos pontos de observação elimina uma das questões que envolvem o uso de regressão por *splines*: onde posicionar os nós. A suavização *spline* se adapta naturalmente a 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* de ordem 4 com nós nos pontos de observação e, assim, minimizar o critério (4.3) com respeito aos coeficientes da expansão. Nesse caso, a função ajustada é formada por pedaços polinomiais cúbicos.

Recordando a relação entre o número de nós, a ordem do B-*spline* e o número de funções base que foi descrita no Capítulo 3, o uso de B-*splines* de quarta ordem implica que o número de funções base será $n + 2$, o que é obviamente suficiente para ajustar n pontos amostrais exatamente se $\lambda = 0$ em (4.3).

4.2.5 Como a suavização *spline* é calculada?

No Capítulo 3 foi visto que uma função pode ser descrita como uma combinação linear de funções base, ou seja,

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

onde $\mathbf{c}$ é o vetor dos coeficientes c_k de tamanho K e $\boldsymbol{\phi}(t)$ é o vetor de funções base avaliadas em t, ou seja, $\boldsymbol{\phi}(t) = [\phi_1(t), \phi_2(t), \ldots, \phi_K(t)]'$.

Sem a penalização da não suavidade, o vetor $\mathbf{c}$ que minimiza a soma dos quadrados dos erros é dado por

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y}, \tag{4.5}$$

onde $\boldsymbol{\Phi}$ é uma matriz $n \times K$ contendo os valores das K funções base calculadas nos n pontos de observação $t_1, t_2, \ldots, t_n$, $\mathbf{W}$ é uma matriz de pesos que permite uma possível estrutura de covariância entre os erros e $\mathbf{y}$ é o vetor de dados discretos a serem suavizados. A expressão correspondente para o vetor de valores ajustados é

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y} = \mathbf{S}\mathbf{y}, \tag{4.6}$$

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 funções base.

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

$$\begin{aligned} PEN_m(x) &= \int [D^m x(s)]^2 ds \\ &= \int [D^m \mathbf{c}'\boldsymbol{\phi}(s)]^2 ds \\ &= \int \mathbf{c}' D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) \mathbf{c} ds \\ &= \mathbf{c}' \left[\int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) ds \right] \mathbf{c} \\ &= \mathbf{c}'\mathbf{R}\mathbf{c}, \end{aligned} \tag{4.7}$$

onde

$$\mathbf{R} = \int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) ds \tag{4.8}$$

é uma matriz quadrada de dimensão K cujas entradas são

$$R_{ij} = \int D^m \phi_i(s) D^m \phi_j(s) ds. \tag{4.9}$$

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

$$SQEPEN_\lambda = (\mathbf{y} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{y} - \mathbf{\Phi c}) + \lambda \mathbf{c}'\mathbf{R}\mathbf{c}. \tag{4.10}$$

Agora, calculando a primeira derivada de $SQEPEN_\lambda$ com respeito a $\mathbf{c}$ e igualando a zero obtem-se

$$\begin{aligned} -2\mathbf{\Phi}'\mathbf{W}\mathbf{y} + 2\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi}\mathbf{c} + 2\lambda\mathbf{R}\mathbf{c} &= \mathbf{0} \\ (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda\mathbf{R})\mathbf{c} &= \mathbf{\Phi}'\mathbf{W}\mathbf{y}. \end{aligned}$$

Assim, o vetor de coeficientes estimado que minimiza (4.10) é dado por

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda\mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{y}. \tag{4.11}$$

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

$$\mathbf{S}_\lambda = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda\mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}. \tag{4.12}$$

Esse operador mais geral (4.12) pode ser chamado de operador de subprojeção, 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*,

$$gl_\lambda = \mathrm{tr}(\mathbf{S}_\lambda), \tag{4.13}$$

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 $\mathbf{R}$ geralmente requer a aproximação númerica da integral em (4.8), embora expressões exatas possam ser calculadas quando funções base de Fourier ou B-*splines* estão envolvidos.

4.2.6 A suavização *spline* como um problema de mínimos quadrados estendido

A expressão (4.10) pode ser interpretada como um problema de mínimos quadrados estendido. Primeiro, uma vez que $\mathbf{R}$ é uma matriz semi-definida positiva, pode-se escrever $\mathbf{R} = \mathbf{L}'\mathbf{L}$ aplicando, por exemplo, a decomposição de

Cholesky. Agora seja

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}, \tag{4.14}$$

onde o vetor de zeros $\mathbf{0}$ é do mesmo tamanho do vetor $\mathbf{c}$. Pode-se relacionar esse vetor de respostas estendido com a matriz estendida

$$\tilde{\mathbf{\Phi}} = \begin{bmatrix} \mathbf{\Phi} \\ \sqrt{\lambda}\mathbf{L} \end{bmatrix}. \tag{4.15}$$

Finalmente, estende-se a matriz de pesos $\mathbf{W}$ adicionando uns na diagonal e zeros nos demais locais. Assim, tem-se a matriz de pesos estendida $\tilde{\mathbf{W}}$.

Agora é possível expressar o vetor de coeficientes usando a penalização da não suavidade como sendo a solução de um simples problema de mínimos quadrados ponderados, ou seja, $\hat{\mathbf{c}}$ é o vetor que minimiza

$$SQE = (\tilde{\mathbf{y}} - \tilde{\mathbf{\Phi}}\mathbf{c})'\tilde{\mathbf{W}}(\tilde{\mathbf{y}} - \tilde{\mathbf{\Phi}}\mathbf{c}). \tag{4.16}$$

Essa versão do problema de penalização da não suavidade, deixa claro que o método de mínimos quadrados penalizados pode ser escrito como o método de mínimos quadrados convencional, onde os dados $\mathbf{y}$ são estendidos por um vetor de zeros, sendo esses zeros ajustados usando a extensão $\sqrt{\lambda}\mathbf{L}$ presente em $\tilde{\mathbf{\Phi}}$.

4.3 A escolha do parâmetro de suavização

Existem duas abordagens distintas com relação à escolha do parâmetro de suavização λ. A primeira abordagem considera a livre escolha do parâmetro de suavização como uma importante característica do procedimento. O que se faz é utilizar diferentes parâmetros e 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. Isso 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; por exemplo, eles podem ser 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 freqüentemente em muitos de conjuntos de dados. No último caso, seria talvez interessante utilizar o mesmo parâmetro de suavização nos diferentes conjuntos de dados para efeito de comparação.

Existem diferentes procedimentos automáticos de escolha do parâmetro de suavização. O mais conhecido de todos é o método de *validação cruzada*, que será apresentado na próxima seção.

4.3.1 O método de validação cruzada

A motivação básica para o método de validação cruzada está relacionada com 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 simula "novas observações" através dos próprios dados como descrito a seguir.

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

$$\sum_{j \neq i} \{y_j - x(t_j)\}^2 + \lambda \int [D^2 x(s)]^2 ds. \tag{4.17}$$

A qualidade de predição de $\hat{x}^{(-i)}$ pode ser julgada através de quão bem o valor $\hat{x}^{(-i)}(t_i)$ 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 valicação cruzada

$$VC(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{x}^{(-i)}(t_i; \lambda)\}^2. \tag{4.18}$$

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

À primeira vista, olhando (4.18), 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

$$\hat{x}(\mathbf{t}) = \mathbf{S}_\lambda \mathbf{y}, \tag{4.19}$$

onde a matriz chapéu $\mathbf{S}_\lambda$ está definida na equação (4.12).

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

$$VC(\lambda) = n^{-1} \sum_{i=1}^{n} \left(\frac{y_i - \hat{x}(t_i)}{1 - S_{\lambda_{ii}}} \right)^2, \tag{4.20}$$

onde $\hat{x}$ é a estimativa por suavização *spline* obtida usando todos os dados $\{(t_i, y_i)\}_{i=1}^n$ com parâmetro de suavização λ. Por sua vez, $S_{\lambda_{ii}}$ é o i-ésimo elemento da diagonal de $\mathbf{S}_\lambda$.

A equação (4.20) 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 cojunto de dados. Portanto, não existem problemas de suavização adicionais a serem resolvidos.

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

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

4.3.2 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 (4.20) mostra os resíduos sendo divididos pelo fatores $1 - S_{\lambda_{ii}}$. A ideia básica da validação cruzada generalizada é substituir esses fatores pelo valor médio deles $1 - n^{-1}\text{tr}(\mathbf{S}_\lambda)$. Uma vez que agora tem-se o mesmo fator para todo i, o seguinte critério é obtido:

$$VCG(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}[y_i - \hat{x}(t_i)]^2}{[1 - n^{-1}\text{tr}(\mathbf{S}_\lambda)]^2}. \tag{4.21}$$

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 $VCG(\lambda)$. Para mais detalhes ver Ramsay e Silverman (2006).

4.4 Uma aplicação em quimiometria

Nessa seção será apresentada uma interessante aplicação de análise de dados funcionais em quimiometria desenvolvida por Saraiva (2009).

A área que se refere à aplicação de métodos estatísticos e matemáticos à problemas de origem química é chamada de quimiometria. Devido à grande evolução dos microcomputadores, as análises instrumentais estão em crescente evolução fazendo-se necessário, do ponto de vista estatístico e matemático, o tratamento mais complexo de dados de origem química com o objetivo de se relacionar os sinais obtidos (espectros por exemplo) com os resultados desejados (concentrações).

A maioria das análises quantitativas é realizada por "via úmida", como por exemplo, titulação e precipitação, que são demoradas e geralmente pouco precisas. Estas estão gradativamente sendo substituídas por técnicas instrumentais, como Espectroscopia no Infravermelho e Espectroscopia da Massa, que aliam velocidade na análise com uma boa qualidade dos resultados. Nas técnicas instrumentais se obtém uma grande quantidade de sinais (curvas) que devem ser tratados para possibilitar a quantificação das várias espécies presentes, uma vez que não há uma informação direta do resultado.

Um problema comum é o de como utilizar espectroscopia para se determinar concentrações de vários constituintes em amostras complexas. Considerando que os constituintes não absorvem luz em regiões separadas de frequência, deve-se utilizar uma combinação de várias frequências espectrais para se estimar as concentrações. O problema de como combinar a absorção em várias frequências de forma ótima com o objetivo de aproximar um conjunto de medidas de concentrações é um problema de calibração multivariada.

As técnicas mais utilizadas no momento são de estatística multivariada, como por exemplo PLS (Mínimos Quadrados Parciais) e PCR (Regressão por Componentes Principais), com os quais se pode medir simultaneamente várias variáveis de interesse ao se analisar uma amostra qualquer. Estes métodos têm dificuldades em tratar dados com características funcionais devido, em geral, à alta dimensão das matrizes de dados.

Para resolver o problema de calibração multivariada, Saraiva (2009) utiliza a idéia da lei de Beer-Lambert (definição 4.4.2) e propõe o tratamento do conjunto de dados coletados utilizando análise não-paramétrica de dados funcionais, uma vez que os mesmos têm características funcionais intrínsecas, isto é, são curvas contínuas (espectros). Esta metodologia não apresenta os problemas teóricos com a dimensão dos dados como os métodos comumente utilizados. Além disso, devido à natureza funcional, acredita-se que modelos que levem em conta esta característica terão melhores resultados do que as técnicas mais utilizadas atualmente.

As duas definições a seguir são importantes para o entendimento do problema.

Definição 4.4.1 *A absorbância de um analito em um dado comprimento de onda ou frequência é definida como*

$$x = -\log_{10}\left(\frac{I}{I_0}\right),$$

onde $x > 0$, I é a intensidade da luz transmitida, após a substância ser inserida no feixe de luz, e I_0 é a intensidade de luz incidente, antes da substância ser inserida no feixe de luz. Para espectroscopia de reflexão, $R = I/I_0$ é conhecido como reflectância ($0 < R < 1$).

Definição 4.4.2 *A lei de Beer-Lambert para m constituintes e k comprimentos de onda mais ruído é*

$$x_j = \sum_{l=1}^{m} y_l a_{lj} + \varepsilon_j, \tag{4.22}$$

para $j = 1, ..., k$, onde x_j é a absorbância da amostra no j-ésimo comprimento de onda, y_l é a concentração do l-ésimo constituinte, a_{lj} é a absorbância do l-ésimo constituinte puro no j-ésimo comprimento de onda e ε_j é o erro aleatório no j-ésimo comprimento de onda.

Mais detalhes sobre a lei de Beer-Lambert podem ser encontrados em Jorgensen e Goegebeur (2007).

Aqui Saraiva (2009) propõe um modelo não-paramétrico funcional para a absorbância semelhante à lei de Beer-Lambert, que relaciona as absorbâncias observadas nas amostras com as concentrações dos constituintes, utilizando funções *spline* obtidas por bases B-*spline* . Os coeficientes das funções base são calculados utilizando o método de mínimos quadrados. Além disso, um modelo para a estrutura de covariância dos dados é proposto, também levando em conta suas características funcionais. Observe que este tipo de metodologia difere da usual porque trata o dado de acordo com sua origem funcional e não como um problema multivariado.

4.4.1 Modelo Não-Paramétrico Funcional

Utilizando a idéia da lei de Beer-Lambert (definição 4.4.2), Saraiva (2009) propõe o seguinte modelo de calibração

$$x_i(t) = \sum_{j=1}^{m} y_{ij} \alpha_j(t) + \epsilon(t), \tag{4.23}$$

onde y_{ij} é a concentração do j-ésimo constituinte na i-ésima amostra, t representa o comprimento de onda, $x_i(t)$ é a absorbância da i-ésima amostra e $\epsilon(t)$ é o erro aleatório no comprimento de onda t, com $i = 1, ..., n$. A função $\alpha_j(t)$ representa a absorbância do j-ésimo constituinte puro no comprimento de onda t.

Estimativas suaves das funções α são obtidas utilizando o método de suavização *spline*. O coeficiente de suavização é escolhido através do critério de validação cruzada generalizada e B-*splines* de ordem 4 são usados. Desta forma, 4.23 pode ser escrito como segue.

$$x_i(t) = \sum_{l=i}^{L} \sum_{j=1}^{m} \theta_{jl} y_{ij} B_l(t) + \epsilon_i(t). \tag{4.24}$$

Neste caso, o modelo não apresenta intercepto mas, no entanto, sabemos que não existem elementos químicos com absorbância exatamente igual a zero uma vez que $\frac{I}{I_0} < 1$, como na definição 4.4.2. Assim, o seguinte modelo com intercepto é proposto,

$$x_i(t) = \sum_{l=1}^{L} \left[\theta_{0l} + \sum_{j=1}^{m} \theta_{jl} y_{ij} \right] B_l(t) + \epsilon_i(t) \tag{4.25}$$

Aqui, $B_l(t)$ corresponde à l-ésima base B-*spline* avaliada no ponto t e θ_{jl} é o coeficiente da l-ésima função base do j-ésimo constituinte.

Pode-se representar o modelo na forma matricial da seguinte maneira

$$\mathbf{x(t)} = \mathbf{D(t)}\boldsymbol{\theta} + \boldsymbol{\epsilon}(\mathbf{t}), \tag{4.26}$$

onde $\mathbf{x(t)}$ é o vetor de absorbâncias de tamanho nk dado por

$$\mathbf{x(t)} = (x_1(t_1), \ldots, x_1(t_k), x_2(t_1), \ldots, x_2(t_k), \ldots, x_n(t_1), \ldots, x_n(t_k))',$$

$\mathbf{D(t)}$ é uma matriz $nk \times L(m+1)$ definida por

$$\mathbf{D(t)} = \begin{pmatrix}
B_1(t_1) & y_{11}B_1(t_1) & \ldots & y_{m1}B_1(t_1) & \ldots & B_L(t_1) & y_{11}B_L(t_1) & \ldots & y_{m1}B_L(t_1) \\
B_1(t_2) & y_{11}B_1(t_2) & \ldots & y_{m1}B_1(t_2) & \ldots & B_L(t_2) & y_{11}B_L(t_2) & \ldots & y_{m1}B_L(t_2) \\
\vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\
B_1(t_k) & y_{11}B_1(t_k) & \ldots & y_{m1}B_1(t_k) & \ldots & B_L(t_k) & y_{11}B_L(t_k) & \ldots & y_{m1}B_L(t_k) \\
B_1(t_1) & y_{12}B_1(t_1) & \ldots & y_{m2}B_1(t_1) & \ldots & B_L(t_1) & y_{12}B_L(t_1) & \ldots & y_{m2}B_L(t_1) \\
B_1(t_2) & y_{12}B_1(t_2) & \ldots & y_{m2}B_1(t_2) & \ldots & B_L(t_2) & y_{12}B_L(t_2) & \ldots & y_{m2}B_L(t_2) \\
\vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\
B_1(t_k) & y_{12}B_1(t_k) & \ldots & y_{m2}B_1(t_k) & \ldots & B_L(t_k) & y_{12}B_L(t_k) & \ldots & y_{m2}B_L(t_k) \\
\vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\
B_1(t_1) & y_{1n}B_1(t_1) & \ldots & y_{mn}B_1(t_1) & \ldots & B_L(t_1) & y_{1n}B_L(t_1) & \ldots & y_{mn}B_L(t_1) \\
B_1(t_2) & y_{1n}B_1(t_2) & \ldots & y_{mn}B_1(t_2) & \ldots & B_L(t_2) & y_{1n}B_L(t_2) & \ldots & y_{mn}B_L(t_2) \\
\vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\
B_1(t_k) & y_{1n}B_1(t_k) & \ldots & y_{mn}B_1(t_k) & \ldots & B_L(t_k) & y_{1n}B_L(t_k) & \ldots & y_{mn}B_L(t_k)
\end{pmatrix},$$

$\boldsymbol{\theta}$ é o vetor de coeficientes de tamanho $L \times (m+1)$ dado por

$$\boldsymbol{\theta} = (\theta_{01}, \theta_{11}, \ldots, \theta_{m1}, \theta_{02}, \ldots, \theta_{m2}, \ldots, \theta_{0L}, \ldots, \theta_{mL})$$

e, finalmente, $\boldsymbol{\epsilon}(\mathbf{t})$ é o vetor $nk \times 1$ de erros aleatórios, ou seja,

$$\boldsymbol{\epsilon}(\mathbf{t}) = (\epsilon_1(t_1), \ldots, \epsilon_1(t_k), \epsilon_2(t_1), \ldots, \epsilon_n(t_1), \ldots, \epsilon_n(t_k)).$$

O vetor de coeficientes $\boldsymbol{\theta}$ pode ser estimado utilizando o método de mínimos quadrados, ou seja, resolvendo o seguinte sistema

$$\hat{\boldsymbol{\theta}} = (\mathbf{D}(\mathbf{t})'\mathbf{D}(\mathbf{t}))^{-1}\mathbf{D}(\mathbf{t})'\mathbf{x}(\mathbf{t}). \tag{4.27}$$

Uma vez estimado o vetor $\boldsymbol{\theta}$, é possível calcular as estimativas de $\mathbf{x}(\mathbf{t})$, da seguinte forma:

$$\hat{\mathbf{x}}(\mathbf{t}) = \mathbf{D}(\mathbf{t})\hat{\boldsymbol{\theta}}. \tag{4.28}$$

Pode-se calcular a curva média $\bar{\mathbf{x}}(\mathbf{t})$ conforme 2.3. Assim, obtemos

$$\bar{\mathbf{x}}(\mathbf{t}) = (\bar{x}(t_1), \bar{x}(t_2), \ldots, \bar{x}(t_k))'. \tag{4.29}$$

4.4.2 Função Covariância

No sentido de obter uma melhor explicação da variabilidade dos dados, Saraiva (2009) sugere a seguinte função de covariância:

$$cov_{\boldsymbol{X}}(t_i, t_j) = \eta(t_i)\eta(t_j)\exp(-\phi|t_j - t_i|), \tag{4.30}$$

onde a função η é contínua para todo t. Analogamente ao modelo 4.25, esta função será aproximada utilizando o método de suavização *spline* com parâmetro de suavização determinado pelo método de validação cruzada generalizada. Desta forma, o modelo 4.30 pode ser escrito como

$$cov_{\boldsymbol{X}}(t_i, t_j) = \sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\beta_{l_1}B_{l_1}(t_i)\beta_{l_2}B_{l_2}(t_j)\exp(-\phi|t_j - t_i|), \tag{4.31}$$

onde $B_l(t)$ representa o valor da l-ésima base B-*spline* avaliada no ponto t. Estas bases são as mesmas que já foram calculadas para o modelo 4.25. O valor de ϕ é fixado e é calculado da seguinte forma (Schmidt, et al. 2008):

$$\phi = \frac{-2\log(0,05)}{\max\limits_{\forall t,s}(|t-s|)}.$$

Na forma matricial tem-se

$$\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\beta}, \tag{4.32}$$

onde

$$\boldsymbol{\Sigma} = \begin{pmatrix} cov_{\boldsymbol{X}}(t_1,t_1) \\ cov_{\boldsymbol{X}}(t_2,t_1) \\ \vdots \\ cov_{\boldsymbol{X}}(t_k,t_1) \\ cov_{\boldsymbol{X}}(t_1,t_2) \\ cov_{\boldsymbol{X}}(t_2,t_2) \\ \vdots \\ cov_{\boldsymbol{X}}(t_k,t_2) \\ \vdots \\ cov_{\boldsymbol{X}}(t_1,t_k) \\ cov_{\boldsymbol{X}}(t_2,t_k) \\ \vdots \\ cov_{\boldsymbol{X}}(t_k,t_k) \end{pmatrix},$$

$$\mathbf{P} = \begin{pmatrix} B_{11}B_{11}e_{11} & B_{11}B_{21}e_{11} & \cdots & B_{11}B_{L1}e_{11} & \cdots & B_{L1}B_{11}e_{11} & \cdots & B_{L1}B_{L1}e_{11} \\ B_{12}B_{11}e_{21} & B_{12}B_{21}e_{21} & \cdots & B_{12}B_{L1}e_{21} & \cdots & B_{L2}B_{11}e_{21} & \cdots & B_{L2}B_{L1}e_{21} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ B_{1k}B_{11}e_{k1} & B_{1k}B_{21}e_{k1} & \cdots & B_{1k}B_{L1}e_{k1} & \cdots & B_{Lk}B_{11}e_{k1} & \cdots & B_{Lk}B_{L1}e_{k1} \\ B_{11}B_{12}e_{12} & B_{11}B_{22}e_{12} & \cdots & B_{11}B_{L2}e_{12} & \cdots & B_{L1}B_{12}e_{12} & \cdots & B_{L1}B_{L2}e_{12} \\ B_{12}B_{12}e_{22} & B_{12}B_{22}e_{22} & \cdots & B_{12}B_{L2}e_{22} & \cdots & B_{L2}B_{12}e_{22} & \cdots & B_{L2}B_{L2}e_{22} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ B_{1k}B_{12}e_{k2} & B_{1k}B_{22}e_{k2} & \cdots & B_{1k}B_{L2}e_{k2} & \cdots & B_{Lk}B_{12}e_{k2} & \cdots & B_{Lk}B_{L2}e_{k2} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ B_{11}B_{1k}e_{1k} & B_{11}B_{2k}e_{1k} & \cdots & B_{11}B_{Lk}e_{1k} & \cdots & B_{L1}B_{1k}e_{1k} & \cdots & B_{L1}B_{Lk}e_{1k} \\ B_{12}B_{1k}e_{2k} & B_{12}B_{2k}e_{2k} & \cdots & B_{12}B_{Lk}e_{2k} & \cdots & B_{L2}B_{1k}e_{2k} & \cdots & B_{L2}B_{Lk}e_{2k} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ B_{1k}B_{1k}e_{kk} & B_{1k}B_{2k}e_{kk} & \cdots & B_{1k}B_{Lk}e_{kk} & \cdots & B_{Lk}B_{1k}e_{kk} & \cdots & B_{Lk}B_{Lk}e_{kk} \end{pmatrix},$$

e

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1\beta_1 \\ \beta_1\beta_2 \\ \vdots \\ \beta_1\beta_L \\ \beta_2\beta_1 \\ \beta_2\beta_2 \\ \vdots \\ \beta_2\beta_L \\ \vdots \\ \beta_L\beta_1 \\ \beta_L\beta_2 \\ \vdots \\ \beta_L\beta_L \end{pmatrix}.$$

Para simplificar a notação, usou-se $B_{ij} = B_i(t_j)$ e $e_{ij} = \exp(-\phi|t_j - t_i|)$.

A partir dos dados observados calcula-se a matriz de covariância empírica considerando que a parte aleatória do modelo está nos erros. Desta forma tal matriz é igual à matriz de covariância dos resíduos. Esta matriz pode ser calculada segundo 2.5, e é dada por

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} cov_{\boldsymbol{\epsilon}}(t_1, t_1) \\ cov_{\boldsymbol{\epsilon}}(t_2, t_1) \\ \vdots \\ cov_{\boldsymbol{\epsilon}}(t_k, t_1) \\ cov_{\boldsymbol{\epsilon}}(t_1, t_2) \\ \vdots \\ cov_{\boldsymbol{\epsilon}}(t_k, t_2) \\ \vdots \\ cov_{\boldsymbol{\epsilon}}(t_1, t_k) \\ \vdots \\ cov_{\boldsymbol{\epsilon}}(t_k, t_k) \end{pmatrix}. \tag{4.33}$$

Tal matriz pode ser escrita da seguinte forma

$$\begin{aligned} \boldsymbol{\Sigma}^* &= \boldsymbol{\Sigma} + \boldsymbol{\omega} \\ \boldsymbol{\Sigma}^* &= \mathbf{P}\boldsymbol{\beta} + \boldsymbol{\omega}, \end{aligned}$$

onde $\boldsymbol{\omega}$ é um erro aleatório cuja distribuição é normal com média zero e matriz de variância e covariância $\sigma^2\mathbf{I}$.

Desta maneira, utilizamos os valores da matriz de covariância empírica para estimar os valores de $\boldsymbol{\beta}$ pelo método de mínimos quadrados, isto é,

$$\hat{\boldsymbol{\beta}} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\boldsymbol{\Sigma}^*.$$

Assim, obtemos as estimativas de $\boldsymbol{\Sigma}$ da seguinte forma:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{P}\hat{\boldsymbol{\beta}}. \tag{4.34}$$

É possível verificar em Saraiva (2009) que a função proposta é uma função de covariância válida

Uma vez que temos um modelo funcional para a estrutura de covariância, podemos estimar também intervalos de confiança para a verdadeira função. Um intervalo de 95% é dado por:

$$\left[\bar{\mathbf{x}}(\mathbf{t}) - 2\mathbf{v}; \bar{\mathbf{x}}(\mathbf{t}) + 2\mathbf{v}\right],$$

onde $\mathbf{v}$ é o vetor de desvios padrões.

4.4.3 Conjunto de Dados de "Flow Injection Analysis"

O objetivo é monitorar reações utilizando uma técnica chamada de "Flow Injection Analysis" (FIA) que é usada para gravar o espectro ultravioleta de uma amostra. A reação é da forma $A + B \rightarrow C$, de modo que o objetivo é quantificar os três compostos e produzir um perfil com o tempo. Os dados contém 25 amostras, cada uma com 3 constituintes. Os espectros foram medidos em 22 comprimentos de onda, variando de 234,39nm a 358,86nm. A Figura 4.2 mostra o conjunto de dados.

A Figura 4.3 mostra as curvas obtidas para cada uma das amostras. A curva sólida em destaque representa a função média das absorbâncias por comprimento de onda.

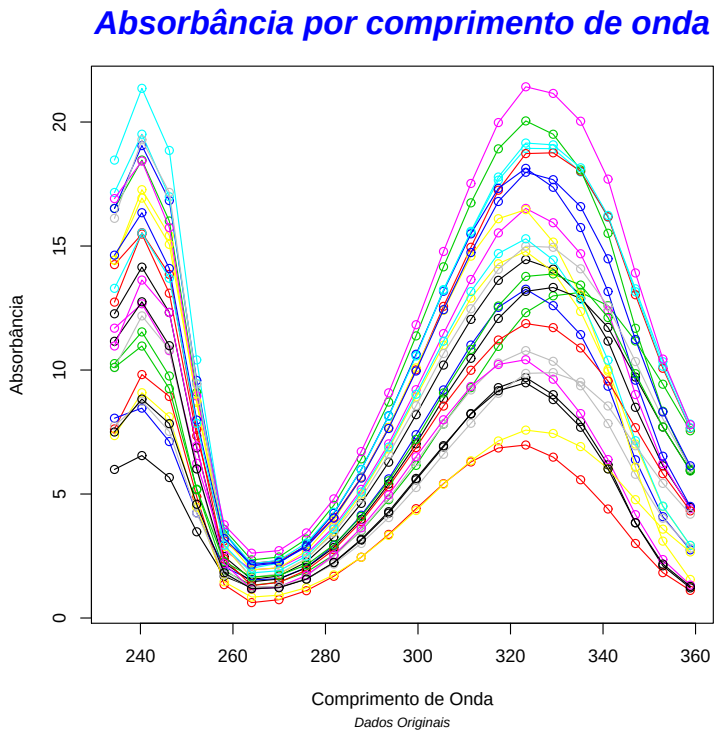


Figura 4.2: Conjunto de dados FIA

Foram construídos intervalos de confiança para a verdadeira função, como mostra o gráfico da Figura 4.4. Nota-se que a amplitude do intervalo de confiança é relativamente pequena em alguns pontos do gráfico, isto porque a variância não é a mesma para todo comprimento de onda.

O código em R utilizado na elaboração desse exemplo encontra-se disponível em Saraiva (2009).

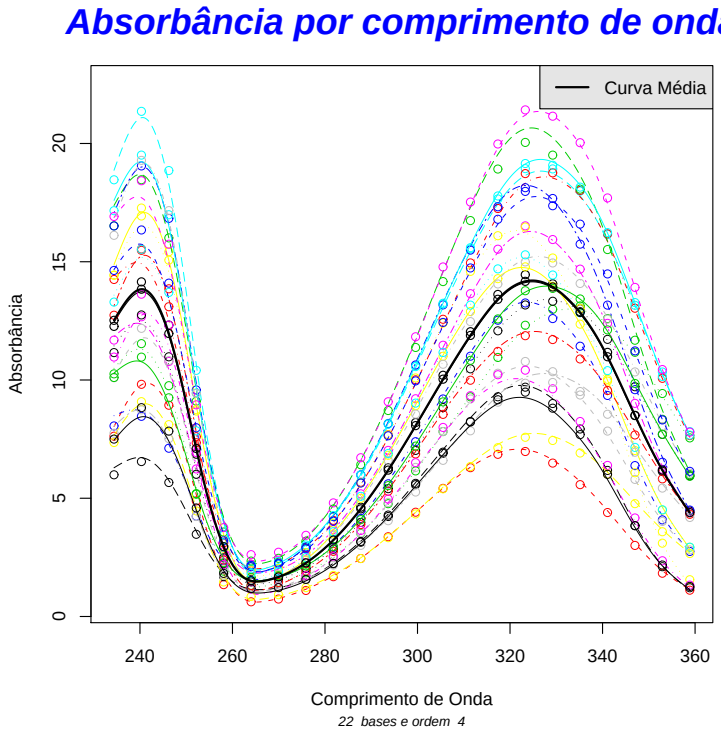


Figura 4.3: Conjunto de dados FIA

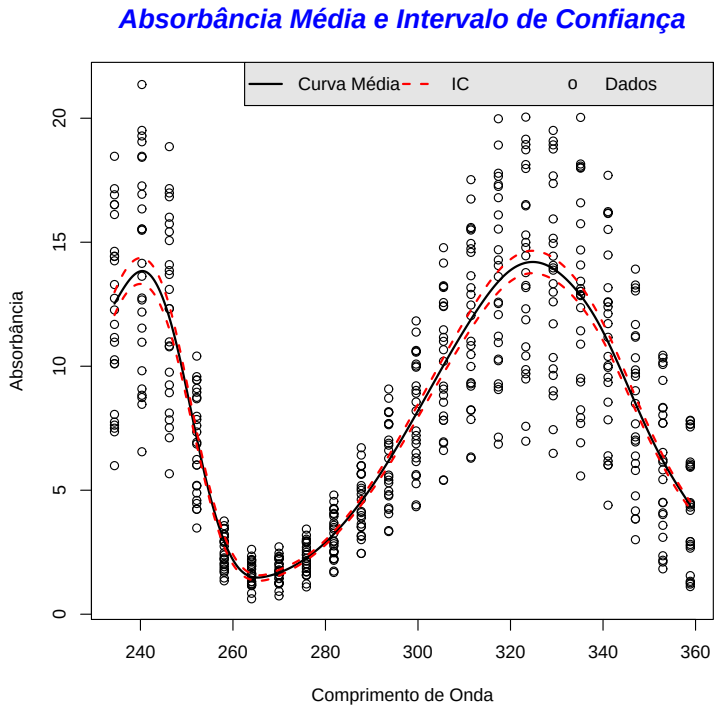


Figura 4.4: Curva média e intervalo de confiança para o conjunto de dados FIA

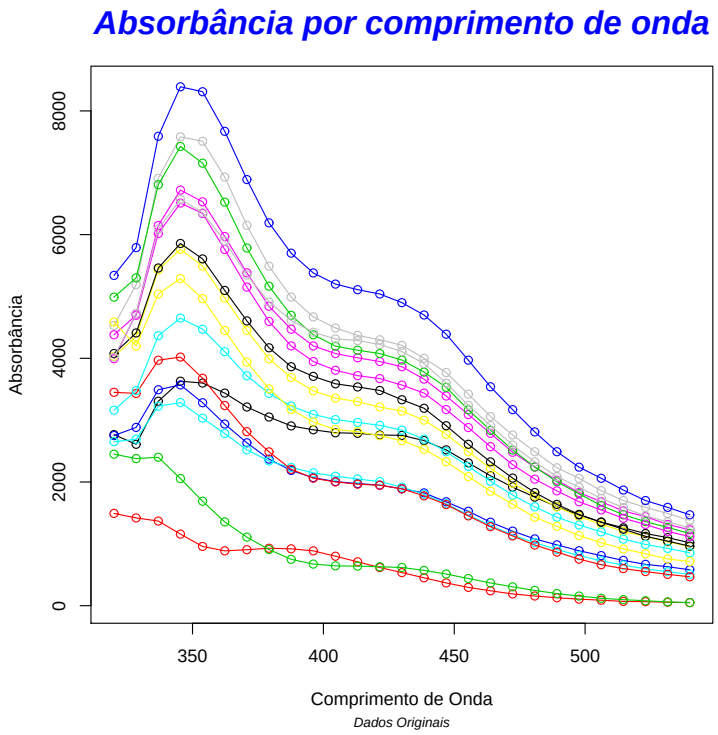


Figura 4.5: Conjunto de dados de Poluição

4.4.4 Dados de Poluição

Um composto despejado na água por fábricas de papel sulfite, chamado de "ligninsulfonate" (sulfonato de lignina), contribui para a poluição das águas do mar e pode ser muito prejudicial para a pesca. Este componente tem sido determinado com algum sucesso através de espectometria de fluorescência. Com este método, as possíveis interferências surgem de ácidos húmicos e detergentes contendo branqueadores óticos. Este conjunto de dados contém 16 amostras, cada uma com 3 elementos: ácido húmico, sulfonato de lignina e detergente. As absorbâncias de cada amostra foram medidas em 27 comprimentos de onda igualmente espaçados entre 320nm a 540nm. A Figura 4.5

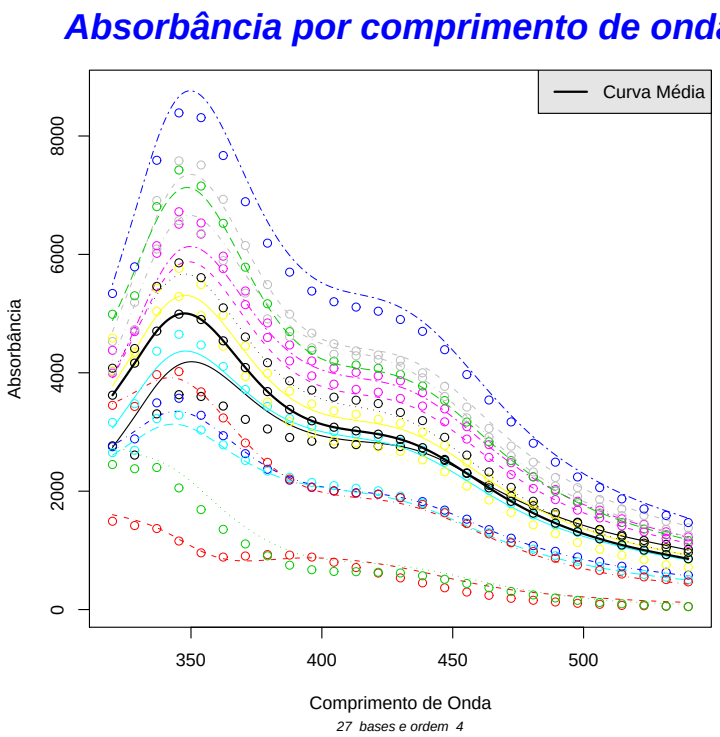


Figura 4.6: Funções estimadas para o conjunto de dados de Poluição

mostra os dados observados.

As curvas estimadas estão mostradas na Figura 4.6. A função médias das absorbâncias é representada pela curva sólida em destaque.

Pode-se ver no gráfico da Figura 4.7 que o intervalo de confiança estimado, como na aplicação anterior, tem amplitude diferente em diferentes comprimentos de onda e a justificativa é a mesma, isto é, a variância se altera de acordo com os comprimentos de onda.

O código em R utilizado na elaboração desse exemplo também encontra-se disponível em Saraiva (2009).

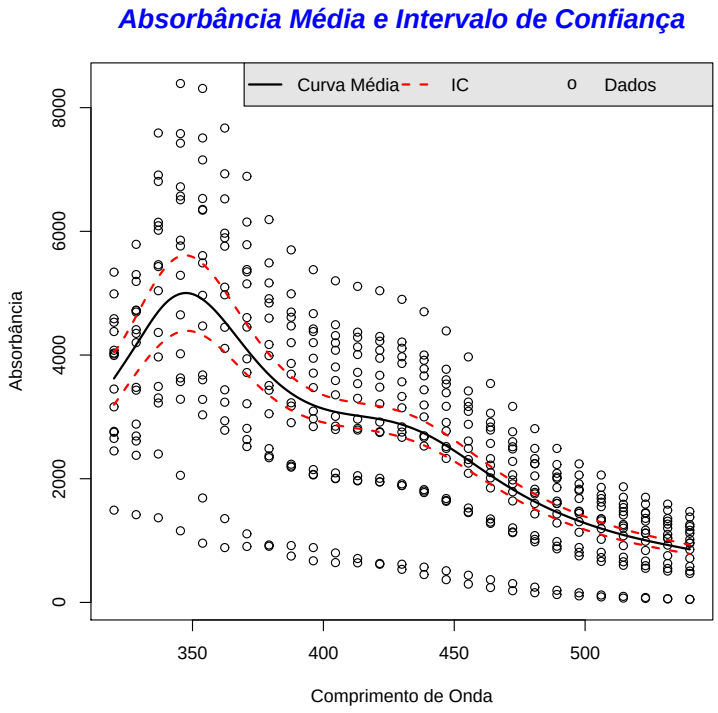


Figura 4.7: Curva média e intervalo de confiança para o conjunto de dados de Poluição

Capítulo 5

O método de suavização monótona

5.1 Introdução

Existem situações onde se exige apenas que as funções estimadas sejam suaves, como foi visto até agora. Contudo, existem casos em que as funções devem satisfazer também outras condições, como, por exemplo, serem estritamente monótonas. Infelizmente, a ideia central de usar uma expansão através de funções base pode causar problemas aqui. É difícil, em geral, forçar funções que são definidas como uma combinação linear a satisfazerem certas restrições.

Nesse capítulo será considerada a seguinte restrição: as funções a serem estimadas devem ser estritamente monótonas. Existem também outras restrições que podem ser feitas às funções, como, por exemplo, serem positivas ou serem funções densidade de probabilidade. Mais detalhes sobre essas outras restrições podem ser encontrados em Ramsay e Silverman (2006).

Muitas técnicas para estimação suave e monótona têm sido desenvolvidas. Elas tendem a ser ou muito complexas em termos de algoritmo ou não muito flexíveis (Ramsay, 1988, Kelly e Rice 1990 e Friedman e Tibshirani 1984). Por outro lado, o método que será apresentado nesse capítulo, desenvolvido por Ramsay (1998), é um procedimento computacionalmente conveniente para estimação de uma função arbitrária, duas vezes diferenciável e estritamente monótona, definida em um intervalo fechado à esquerda que pode ser, sem

perda de generalidade, $[0, \infty)$ ou $[0, 1]$.

A Figura 5.1 apresenta um problema de suavização monótona. Nessa figura é possível ver os registros da altura de um garoto tomados em 83 dias durante um ano escolar, as falhas correspondem às férias. Os dados presentes na Figura 5.1 são os mesmos da Figura 1.2 do Capítulo 1. A Figura 5.1 também apresenta dois ajustes suaves: um ajuste por suavização *spline*, usando o método de validação cruzada generalizada para seleção do parâmetro de suavização; e o ajuste obtido por suavização monótona, que será apresentado nesse capítulo. Embora pareça razoável assumir que a altura cresça de forma monótona, a curva obtida por suavização *spline* está longe de ser monótona. Nota-se que a suavização monótona se comporta de forma mais adequada do que a suavização *spline* nesse exemplo.

```
library(fda)
altGaroto <- onechild$height ; dia <- onechild$day
dia2 <- seq(1,max(dia),length=151)
nBases <- 43
altBases <- create.bspline.basis(range=range(dia), nbasis=nBases,
            norder=4)
cvec0 <- matrix(0,nBases,1) ; Wfd0  <- fd(cvec0, altBases)
garotofdPar <- fdPar(Wfd0, Lfdobj=2, lambda=1)
resultado <- smooth.monotone(dia, altGaroto,garotofdPar,
 active=c(TRUE,rep(TRUE,nBases-1)))
objFuncGaroto  <- resultado$Wfdobj
beta <- resultado$beta
altGarotoSuave <- beta[1] + beta[2]*eval.monfd(dia2, objFuncGaroto)
sp <-smooth.spline(dia,altGaroto,cv=FALSE)
altGarotoSuave2 <- predict(sp,dia2,deriv=0)$y
```

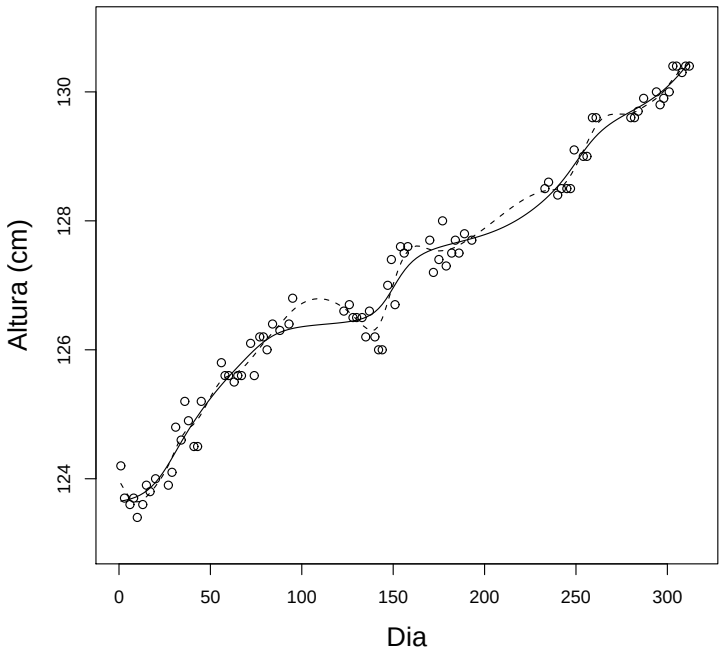


Figura 5.1: Os círculos são 83 medidas de altura de um garoto de 10 anos durante o ano escolar. A curva sólida é o ajuste obtido por suavização monótona dos dados. Já a curva tracejada é a estimativa por suavização *spline*, com o parâmetro de suavização λ escolhido atráves do critério de VCG. Em ambos os ajustes foram utilizados B-*splines* cúbicos.

5.2 Uma equação diferencial para funções monótonas

A notação $D^{-1}x$ será usada aqui para definir o operador de integração, ou seja,

$$D^{-1}x(t) = \int_0^t x(s)ds.$$

A classe de funções monótonas em questão aqui é formada pelas funções x para as quais $\ln(Dx)$ é diferenciável e $D\{\ln(Dx)\} = D^2x/Dx$ é de quadrado integrável à Lebesgue. Essas condições garantem que a função x é estritamente monótona crescente ($Dx > 0$) e que sua primeira derivada é suave e limitada quase que em toda parte. O teorema a seguir, presente em Ramsay (1998), diz que essa classe é identificada com uma simples equação diferencial linear.

Teorema. *Toda função x pertencente a classe de funções monótonas descrita anteriormente é representável como*

$$x = C_0 + C_1 D^{-1}\{\exp(D^{-1}w)\} = C_0 + C_1 \int \exp\left[\int_0^s w(u)du\right] ds \qquad (5.1)$$

ou como uma solução da equação diferencial homogênea linear

$$D^2x = wDx, \qquad (5.2)$$

onde w é uma função de quadrado integrável à Lebesgue e C_0 e C_1 são constantes arbitrárias.

Prova. Se x tem a forma (5.1), note que

$$Dx = C_1 \exp(D^{-1}w)$$

e, sabendo que $Dx > 0$, tem-se que

$$\ln Dx = \ln C_1 + D^{-1}w.$$

Assim,

$$D\{\ln Dx\} = w$$

e, portanto, w é de quadrado integrável à Lebesgue por definição da classe. É fácil ver que x da forma (5.1) também satisfaz a equação (5.2) para $w = D\{ln(Dx)\}$. Como $Dx = C_1 \exp(D^{-1}w)$, então

$$D^2x = C_1 \exp(D^{-1}w)w = wDx$$

Por outro lado, se x tem como representação a equação (5.2), então a equação (5.1) é uma solução e ela é um resultado padrão da teoria de equações diferenciais lineares em que o espaço solução é linear e de dimensão 2 e, assim, toda solução é também dessa forma. Note que é possível escrever (5.2) da seguinte forma:

$$D^2x = D(Dx) = wDx$$

e, assim,

$$\frac{D(Dx)}{Dx} = w.$$

Usando a regra da cadeia e sabendo que $Dx > 0$, tem-se que $D\{\ln(Dx)\} = D(Dx)/Dx$. Dessa forma, pode-se escrever

$$D\{\ln(Dx)\} = w.$$

Integrando ambos os lados da equação acima tem-se que

$$\begin{aligned} D^{-1}\{D[\ln(Dx)]\} &= D^{-1}w, \\ \ln(Dx) + C &= D^{-1}w, \\ \ln(Dx) &= D^{-1}w - C \end{aligned}$$

e agora tomando a exponencial nota-se que

$$Dx = C_1 \exp(D^{-1}w);$$

novamente integrando ambos os lados conclui-se que

$$\begin{aligned} D^{-1}\{Dx\} &= D^{-1}\{C_1 \exp(D^{-1}w)\} \\ x &= C_0 + C_1 D^{-1}\{\exp(D^{-1}w)\}. \end{aligned}$$

Quando x é estritamente monótona decrescente, ou seja, $Dx < 0$, a prova segue de forma semelhante, basta considerar $|Dx|$ e o fato que $D\{\ln(|Dx|)\} = D\{\ln(Dx)\}$. □

A função coeficiente $w = D\{ln(Dx)\} = D^2x/Dx$ mede a curvatura relativa da função monótona. Isso significa que ela calcula o tamanho da curvatura D^2x relativo à inclinação Dx. O caso especial $w = \alpha$ define $x(t) = C_0 + C_1 \exp(\alpha t)$, de tal forma que funções exponenciais tenham curvatura relativa constante. Já quando $w = 0$ define-se uma função linear, ou seja, $x(t) = C_0 + C_1 t$. Portanto, valores pequenos de $w(t)$ ou iguais a zero correspondem localmente a funções lineares, enquanto que valores grandes correspondem a regiões de curvatura acentuada.

5.3 Suavização monótona de dados

Considere mais uma vez o modelo $y_i = x(t_i) + \epsilon_i$, onde os valores ϵ_i seguem a suposição padrão para os erros, ou seja, são independentes e identicamente distribuídos, com média zero e variância constante σ^2, e os valores de argumento t_i estão dentro do intervalo $[0, T]$.

A soma dos quadrados dos erros penalizada nesse caso pode ser escrita como

$$SQEPEN_\lambda = n^{-1} \sum_{i=1}^{n} \{y_i - C_0 - C_1 m(t_i)\}^2 + \lambda \int_0^T [w(t)]^2 dt, \qquad (5.3)$$

onde $m(t) = D^{-1}\{\exp(D^{-1}w)\}(t) = \int_0^t \exp[\int_0^s w(u)du]ds$.

O primeiro termo da equação (5.3) é a soma dos quadrados dos erros usual nos problemas de mínimos quadrados, exceto que os parâmetros de regressão linear C_0 e C_1 são essenciais porque $m(0) = 0$ e $Dm(0) = 1$. O primeiro termo também poderia ser generalizado dando pesos para as observações, mas por simplicidade de exposição será apresentado apenas o caso onde os pesos são todos iguais a um.

O segundo termo é o termo de penalização ou regularização. Ele tem algumas das características da norma da segunda derivada usada na suavização *spline*, mas agora o papel desempenhado pelo denominador em $w = D^2x/Dx$ é importante, uma vez que se deseja que a regularização mantenha a função ajustada longe da condição $Dx = 0$. Da mesma forma que na suavização *spline*, é possível que alguém esteja interessado, por exemplo, em controlar a curvatura da aceleração. Sendo assim, deve-se usar como penalização $\int [D^2w(t)]^2 dt$, ao

invés de $\int [w(t)]^2 dt$. O parâmetro de suavização λ pode ser escolhido por validação cruzada. Porém, essa técnica pode falhar, obtendo-se um parâmetro de suavização muito grande. Como na suavização *spline*, quando $\lambda \to \infty$ a curva ajustada é uma reta.

5.4 Expansão em funções base para w

A principal vantagem trazida pela representação (5.1) é a transformação do problema de estimação. Antes o problema era estimar a função restrita x e, agora, o problema consiste na estimação da função irrestrita w. Por causa da falta de restrição, w pode ser definida como uma combinação linear de algum conjunto de funções base ϕ_k, $k = 1, \ldots, K$, que seja apropriado para o problema em mãos. É claro que o conjunto de funções base B-*splines* é muito utilizado devido sua grande flexibilidade e boas propriedades de aproximação.

Seja $\boldsymbol{\phi}$ o vetor de funções base $(\phi_1, \ldots \phi_K)'$. Dessa forma, tem-se que

$$w(t) = \mathbf{c}'\boldsymbol{\phi}(t), \tag{5.4}$$

onde $\mathbf{c}$ é o vetor de coeficientes que define a combinação linear. Seja $\boldsymbol{\Phi}$ o vetor $(\Phi_1, \ldots, \Phi_K)'$ onde $\Phi_k = D^{-1}\phi_k$, $k = 1, \ldots, K$. Assim, de acordo com a representação (5.1), a função estimada é da forma

$$\hat{x}(t) = \hat{C}_0 - \hat{C}_1 D^{-1}[\exp\{\hat{\mathbf{c}}'\boldsymbol{\Phi}(t)\}], \tag{5.5}$$

onde $\hat{C}_0$, $\hat{C}_1$ e $\hat{\mathbf{c}}$ são obtidos minimizando o critério (5.3). É claro que aqui se tem um problema de regressão não linear e procedimentos apropriados devem ser utilizados para obter as estimativas dos coeficientes. Ramsay (1998) propõe um procedimento satisfatório para a estimação dos coeficientes. Primeiro comece com uma estimativa inicial $\hat{\mathbf{c}}^{(0)}$, que pode ser um vetor de zeros, estime $C_0^{(0)}$ e $C_1^{(0)}$ por regressão linear. Então, para qualquer iteração $\nu > 0$, na qual $\hat{C}_0^{(\nu-1)}$, $\hat{C}_1^{(\nu-1)}$ e $\hat{\mathbf{c}}^{(\nu-1)}$ são as estimativas encontradas na iteração anterior, primeiro otimize (5.3) com respeito a $\mathbf{c}$, utilizando o método de Gauss-Jordan para problemas de mínimos quadrados não-lineares para obter $\hat{\mathbf{c}}^{(\nu)}$ e, depois, calcule $\hat{C}_0^{(\nu)}$ e $\hat{C}_1^{(\nu)}$ por regressão linear. O procedimento de Gauss-Jordan requer que o vetor de atualização

$$\boldsymbol{\delta}^{(\nu)} = \mathbf{c}^{(\nu)} - \mathbf{c}^{(\nu-1)}$$

seja a solução da equação linear

$$\mathbf{R}^{(\nu-1)}\boldsymbol{\delta}^{(\nu)} = -\mathbf{s}^{(\nu-1)},$$

onde

$$\mathbf{R} = n^{-1}\hat{C}_1^2\mathbf{X}'\mathbf{X} + \lambda\mathbf{K},$$

a matriz $\mathbf{X}$ é $n \times K$ e tem as linhas dadas por

$$\mathbf{x}(t_i) = \frac{\partial m(t_i)}{\partial \mathbf{c}} = \int_0^{t_i} \mathbf{\Phi}(s)\exp\{\hat{\mathbf{c}}'\mathbf{\Phi}(s)\},$$

a matriz simétrica $\mathbf{K}$ de ordem K é

$$\mathbf{K} = \int_0^T \phi(s)\phi'(s)ds,$$

o vetor $\mathbf{s}$ de tamanho K é dado por

$$\mathbf{s} = -n^{-1}\hat{C}_1\mathbf{X}'\mathbf{r} + \lambda\mathbf{K}\hat{\mathbf{c}}$$

e onde, finalmente, $\mathbf{r}$ é o vetor de tamanho n contendo os resíduos $r_i = y_i - \hat{C}_0 - \hat{C}_1\hat{m}(t_i)$. A taxa de convergência dessas iterações é somente linear, mas ela pode ser considerada rápida e geralmente é obtida em quatro ou cinco iterações.

Mais detalhes sobre o método de suavização monótona podem ser encontrados em Ramsay (1998).

Capítulo 6

Inferência para dados funcionais

6.1 Introdução

Um problema de grande interesse na área de análise de dados funcionais é o de testes de hipóteses. No caso de estimação de densidades, quando a função de densidade pode ser representada como uma combinação linear de funções base num espaço infinito dimensional apropriado, a distância de Kullback-Leibler simetrizada possui boas propriedades assintóticas e é possível a construção de testes de hipóteses baseados em uma estatística do teste que é assintoticamente normal (ver Dias e Garcia (2007) e Dias e Garcia (2005)). No cenário de regressão, técnicas similares poderiam ser utilizadas.

Nesse capítulo será apresentado um método para testar hipóteses quando os dados são curvas proposto por Souza (2008). As estatísticas dos testes serão baseadas nas distâncias de Hellinger, de Kullback-Leibler, L_1 e na diferença quadrática integrada.

Assuma que as curvas $\{Y_1(t), \ldots, Y_{n_1}(t) : t \in \mathcal{T}\}$ e $\{X_1(t), \ldots, X_{n_2}(t) : t \in \mathcal{T}\}$ são amostras independentes de processos estocásticos $\mathcal{Y}$ e $\mathcal{X}$ respectivamente. Denote as curvas médias dos processos $\mathcal{X}$, $\mathcal{Y}$ por $\mu_X(t)$ e $\mu_Y(t)$ respectivamente. Existe interesse em testar a igualdade dessas curvas médias quando dados funcionais são observados. Pode-se também testar se uma curva média é igual a uma certa função conhecida, ou seja, se $\mu_X(t) = f(t)$. É pos-

sível, também, testar hipóteses com outros funcionais, como, por exemplo, derivadas das curvas médias.

Estimativas suaves das curvas serão obtidas através dos métodos de suavização *spline* ou suavização monótona.

6.1.1 A distância L_1

Um forma muito comum de medir o afastamento entre duas funções x e y é através da distância L_1 dada por

$$L_1(x, y) = \int |x - y|. \tag{6.1}$$

Allen (1997) propõe um teste baseado na distância L_1 entre as estimativas de densidade por *kernel* de duas amostras, com o objetivo de testar a proximidade de suas distribuições. Simulações foram feitas para comparar esse teste com outros através do poder, utilizando o procedimento *bootstrap* de reamostragem. Resultados satisfatórios foram obtidos.

6.1.2 A distância de Hellinger

Seja $\mathcal{L}^2[a, b]$ o conjunto de todas as funções de quadrado integrável num certo intervalo $[a, b]$. Seja $g \in \mathcal{L}^2[a, b]$ e, assim, considere a seguinte transformação:

$$t_g = \frac{g^2}{\int g^2}. \tag{6.2}$$

Dessa forma, $t_g \geq 0$ e $\int t_g = 1$, ou seja, t_g é uma função densidade.

Agora, usando essa transformação tem-se que a distância de Hellinger entre duas funções x e y em $\mathcal{L}^2[a, b]$ é descrita por

$$H(x, y) = \left(\int \left(\sqrt{t_x} - \sqrt{t_y} \right)^2 \right)^{1/2}. \tag{6.3}$$

Com a transformação (6.2), pode-se também definir a partir do quadrado da distância de Hellinger uma outra medida de associação entre as funções x

e y, a chamada afinidade entre x e y, ou, simplesmente, $\rho(x, y)$. Da equação (6.3) tem-se que

$$H^2(x, y) = \int \left(\sqrt{t_x} - \sqrt{t_y}\right)^2 = 2(1 - \rho(x, y)), \tag{6.4}$$

onde

$$\rho(x, y) = \int \sqrt{t_x t_y}. \tag{6.5}$$

Não é difícil ver que $0 \leq \rho(x, y) \leq 1$ para qualquer x, y. Além disso, note que $H(x, y)$ é mínimo quando a afinidade, $\rho(x, y)$, é 1.

6.1.3 A distância de Kullback-Leibler

Considere mais uma vez a transformação $t_g = g^2 / \int g^2$. Assim, a distância de Kullback-Leibler entre as funções x e y é dada por

$$KL(x, y) = \int (\log t_x - \log t_y) t_x. \tag{6.6}$$

6.1.4 A diferença quadrática integrada

Uma medida muito usada para avaliar a proximidade entre duas funções x e y é a diferença quadrática integrada, descrita por

$$DQI(x, y) = \int (x - y)^2. \tag{6.7}$$

Qi Li (1996) utiliza a diferença quadrática integrada entre duas estimativas de densidade por *kernel* para construir um teste não paramétrico para a proximidade de duas distribuições. Considerando certas condições, a estatística do teste proposta por Qi Li é assintoticamente $N(0, 1)$ sob a hipótese nula.

6.2 Estudos simulados

6.2.1 Introdução

Deseja-se testar se a curva média de um conjunto de dados funcionais é igual a uma certa curva conhecida, ou seja, deseja-se testar a hipótese nula $H_0 : \mu = f$ quase certamente contra a alternativa $H_1 : \mu \neq f$.

Seja $\mathcal{W}_2^2[a,b]$ o conjunto formado pelas funções definidas no intervalo $[a,b]$, cuja primeira derivada é absolutamente contínua e cuja segunda derivada é de quadrado integrável. A escolha desse conjunto de funções é importante para que a penalização da não suavidade, utilizada para obter uma estimativa suave dos dados, seja finita. Assim, é importante considerar que as funções μ e $f \in \mathcal{W}_2^2[a,b]$.

Dessa forma, é possível definir algumas estatísticas do teste baseadas nas distâncias descritas anteriormente e na estimativa da curva média $\hat{\mu}$. É importante considerar aqui as transformações $t_{\hat{\mu}}$ e t_f, definidas pela equação (6.2), para garantir que todas as distâncias estejam convenientemente definidas.

- **A estatística L_1:**
$$L_1 = \int |t_{\hat{\mu}} - t_f|. \tag{6.8}$$

- **A estatística KL:**
$$KL = \int (\log t_f - \log t_{\hat{\mu}}) t_f. \tag{6.9}$$

- **A estatística de Hellinger:**
$$H = \left(\int \left(\sqrt{t_{\hat{\mu}}} - \sqrt{t_f} \right)^2 \right)^{1/2}. \tag{6.10}$$

- **A afinidade ρ:**
$$\rho = \int \sqrt{t_{\hat{\mu}} t_f}. \tag{6.11}$$

- **A estatística DQI:**
$$DQI = \int (t_{\hat{\mu}} - t_f)^2. \tag{6.12}$$

Para todas as estatísticas, a integral pode ser aproximada por uma soma multiplicada pelo comprimento Δs dos intervalos igualmente espaçados entre as observações. Por exemplo, no caso da estatística L_1 tem-se que

$$\tilde{L}_1 = \Delta s \left[\sum_{i=1}^{m} |t_{\hat{\mu}}(s_i) - t_f(s_i)| \right]. \tag{6.13}$$

Os objetivos principais dos estudos simulados por Souza (2008) são:

- Avaliar a distribuição de cada estatística do teste sob H_0, já que elas não possuem distribuição de probabilidade conhecida;

- A partir da distribuição das estatísticas sob H_0 verificar o poder do teste proposto para cada uma dessas estatísticas.

6.2.2 Estrutura dos estudos simulados

É importante salientar mais uma vez que o objetivo aqui é estudar as distribuições das estatísticas do teste sob H_0, ou seja, estudar as distribuições das estatísticas quando a curva média do conjunto de dados funcionais é mesmo igual a uma certa curva f.

A seguir, tem-se em ordem a estrutura dos estudos.

1. Uma função $f \in \mathcal{W}_2^2[a,b]$ é escolhida e m pontos dessa função são calculados;

2. A cada ponto da função é adicionado um erro aleatório com distribuição $N(0,\sigma^2)$. Assim, são obtidas as observações $y_i = f(t_i) + \epsilon_i$, $i = 1,\ldots,m$. O vetor de tamanho m formado por essas observações é o que se chama de dado funcional (ver Figura 6.1). É possível também adicionar uma estrutura de covariância em cada curva considerando $y_{ij} = \mu(t_i) + a_j + \epsilon_{ij}$, onde $a_1,\ldots,a_n$ são iid $N(0,\sigma_a^2)$;

3. O passo 2 é repetido n vezes e dessa forma um conjunto com n dados funcionais é obtido. Sendo assim, y_{ij} denota agora a i-ésima observação do dado funcional j, $i = 1,\ldots,m$ e $j = 1,\ldots,n$;

4. Os dados funcionais são então suavizados utilizando um dos procedimentos de suavização discutidos nesse trabalho. Assim, uma amostra de n curvas suaves é obtida.

5. Calcula-se a curva média estimada, ou seja, $\hat{\mu}$ para essa amostra de n curvas suaves obtidas no passo anterior (ver Figura 6.2);

6. As transformações $t_{\hat{\mu}}$ e t_f são calculadas;

7. O valor de cada uma das estatísticas propostas é calculado;

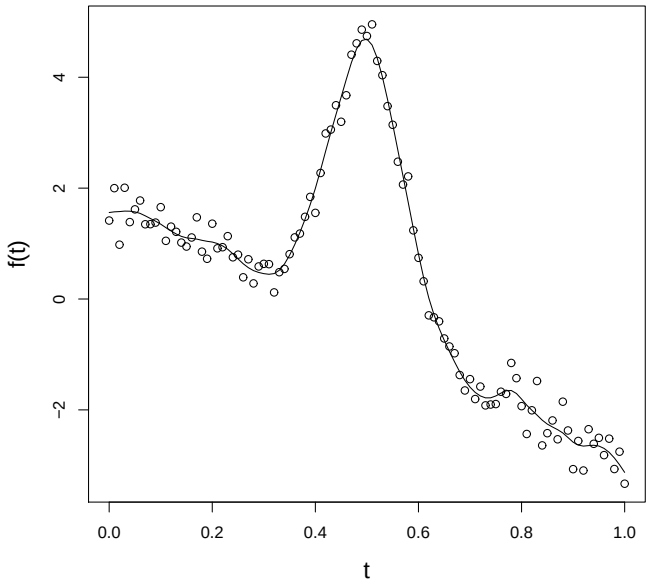


Figura 6.1: Os pontos são as observações $y_i = f(t_i) + \epsilon_i$, $i = 1, \ldots, 101$ e $\epsilon_i \sim N(0; 0,25^2)$ que juntas formam um dado funcional, sendo $f(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$. A linha sólida corresponde ao ajuste suave por suavização *spline*.

8. Os passos (2) até (7) são repetidos um número R de vezes;

9. No final são obtidas amostras aleatórias das estatísticas e, assim, a distribuição de cada uma delas pode ser avaliada através de histogramas e estimativas de função densidade.

Souza (2008) apresenta resultados de diversas simulações utilizando diferentes funções $f(t)$. A seguir um dos exemplos elaborados por Souza (2008) será apresentado. O código em R desenvolvido para as simulações também está disponível em Souza (2008).

Considere a hipótese nula $H_0 : \mu(t) = f(t)$ contra a alternativa $H_1 : \mu(t) \neq f(t)$, sendo $f(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$. Nesse exemplo a distribuição

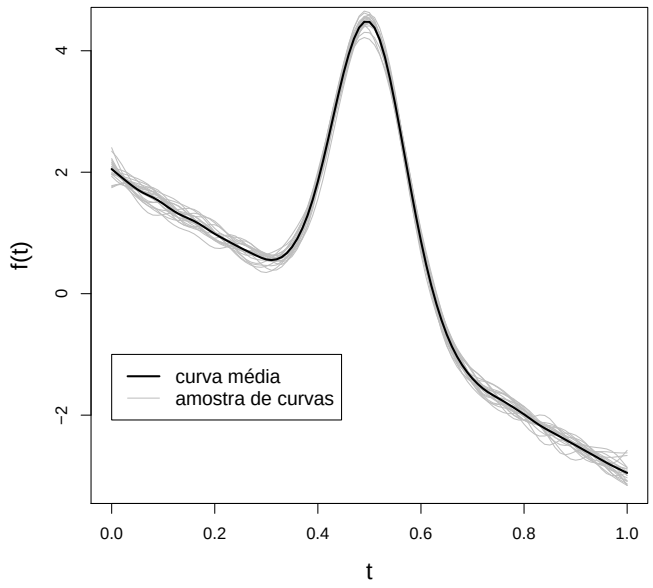


Figura 6.2: Amostra de 20 curvas suaves estimadas através da suavização *spline* e sua correspondente curva média estimada.

de cada uma das estatísticas (sob H_0) foi estudada.

A curva média aqui é a média das $n = 1000$ curvas estimadas através da suavização *spline*, com $\lambda = 1 \times 10^{-5}$ e B-*splines* de ordem 4.

Considerando que os erros têm distribuição $N(0, \sigma^2)$, os próximos resultados mostram o comportamento da distribuição de cada uma das estatísticas do teste, sob H_0, quando $\sigma = 1$. Para isso a estrutura proposta na Seção 6.2.2 foi utilizada, sendo o número de repetições $R = 5000$.

A Figura 6.3 mostra a curva média estimada obtida assumindo que os erros aleatórios têm distribuição normal com média zero e desvio padrão 1.

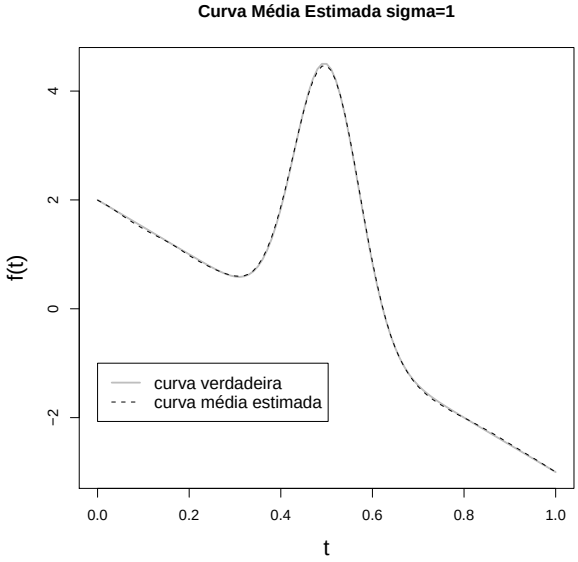


Figura 6.3: A curva sólida cinza é a função $f(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$. Já a curva tracejada é a estimativa da curva média por suavização *spline* para $n = 1000$ e $\boldsymbol{\sigma = 1}$.

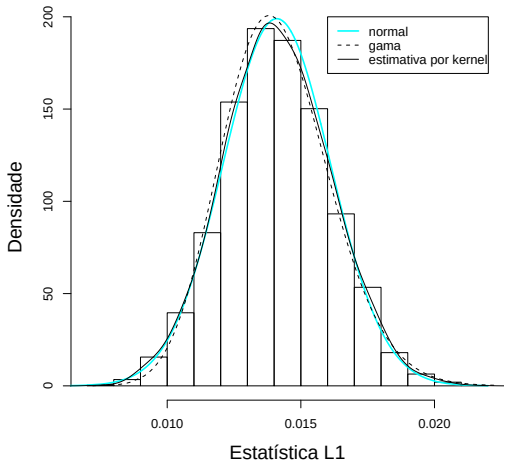
As Figuras 6.4, 6.5 e 6.6 apresentam o histograma de cada uma das estatísticas sob H_0, com $n = 1000$, ϵ_j's $\sim N(0; 1)$ e $R = 5000$. É possível observar também as diferentes estimativas de densidade obtidas. Nota-se que as distribuições das estatísticas DQI e KL são mais assimétricas do que as das

estatísticas $L1$ e Hellinger. Observa-se também que para as estatísticas KL e DQI, as estimativas de densidade de uma normal e de uma gama não são muito semelhantes. A estimativa de densidade uma beta parece ser bastante razoável para a afinidade, já que sabemos que $0 \leq \rho(x, y) \leq 1$.

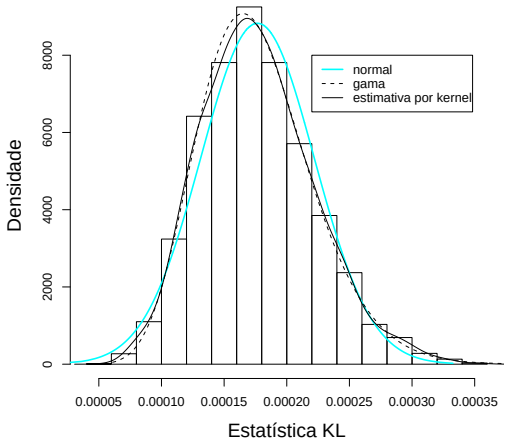
Para as estatísticas $L1$ e Hellinger a hipótese de que sua distribuição seja normal não é rejeitada, considerando o nível de significância $\alpha = 0,05$. Isso já era esperado devido ao comportamento simétrico das distribuições empíricas nesse caso quando $\sigma = 1$. A Tabela 6.1 apresenta os p-valores obtidos atráves do teste de Kolmogorov-Smirnov para duas amostras.

Tabela 6.1: Teste de Kolmogorov-Smirnov usado para testar se a distribuição da estatística sob $H_0 : \mu(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$ é normal; $\sigma = 1$.

Estatística	p-valor
$L1$	0,610
Hellinger	0,744
DQI	3,87 $\times 10^{-7}$
KL	0,0185
Afinidade	1,22 $\times 10^{-5}$

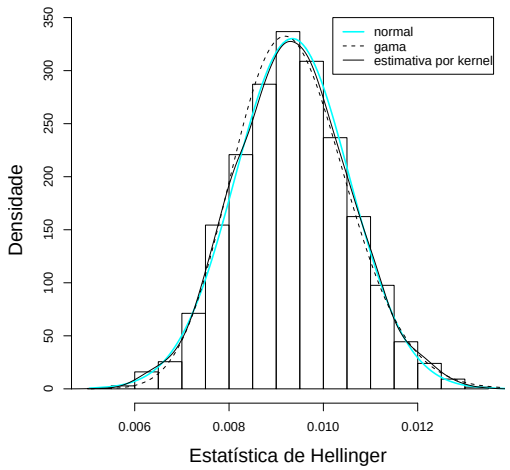


(a)

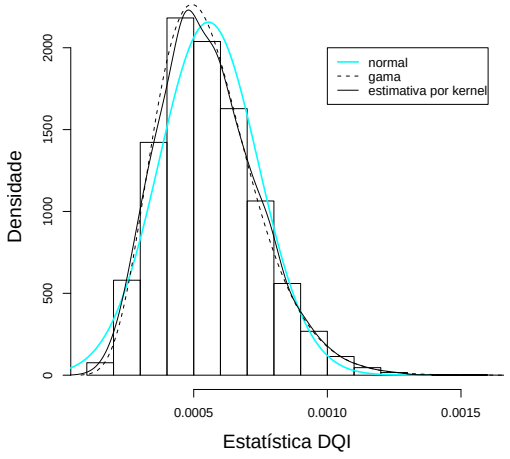


(b)

Figura 6.4: Distribuições das estatísticas $L1$ (a) e KL (b) sob $H_0 : \mu(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$ quando $\sigma = 1$.



(a)



(b)

Figura 6.5: Distribuições das estatísticas de Hellinger (a) e DQI (b) sob $H_0 : \mu(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$ quando $\sigma = 1$.

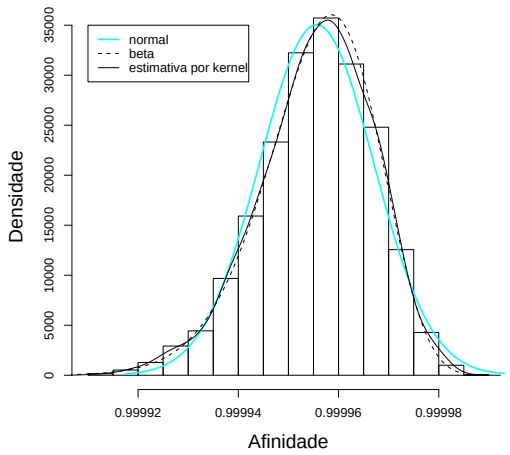


Figura 6.6: Distribuição da afinidade sob $H_0 : \mu(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$ quando $\sigma = 1$.

6.2.3 O poder do teste

Baseado na distribuição estimada sob $H_0 : \mu(t) = f(t)$ de cada uma das estatísticas do teste L_1, KL, DQI, Hellinger e afinidade, Souza (2008) também apresenta um estudo sobre o poder do teste (probabilidade de rejeitar a hipótese nula quando ela é falsa). Para isso a função $f(t) = 2 - 5t + 5\exp[-100(t-\eta)^2]$ foi considerada utilizando diferentes valores de η, sendo que sob H_0 $\eta = 0,5$. Assim, é possível reescrever as hipóteses como $H_0 : \eta = 0,5$ vs $H_1 : \eta \neq 0,5$. O poder do teste é estudado através da função poder $\pi(\eta)$, que é definida como a probabilidade de rejeitar H_0 quando a distribuição da qual a amostra foi obtida foi parametrizada por η. Quando $\eta = 0,5$, $\pi(\eta)$ deve ser igual ou bem próximo ao nível de significância α do teste. Por sua vez, quando $\eta \neq 0,5$, $\pi(\eta)$ corresponde à probabilidade de rejeitar H_0 quando a mesma é falsa, ou seja, ao poder do teste. Nesse caso, não é possível escrever uma fórmula para a função poder, já que as distribuições exatas das estatísticas sob H_0 não são conhecidas, portanto, a probabilidade de rejeição empírica é utilizada para se obter uma estimativa da função poder do teste.

Dessa forma, a distribuição estimada sob H_0 de cada uma das estatísticas obtida assumindo que os ϵ_j's $\sim N(0; 0,25^2)$ é usada para obter um critério de decisão. Como estamos testando a hipótese de igualdade contra diferença, temos um teste bilateral. Assim, são considerados como pontos de corte os valores C_1 e C_2 das estatísticas tais que $P_{H_0}(T \leq C_1) = 0,025$ e $P_{H_0}(T \geq C_2) = 0,025$, ou seja, rejeita-se H_0 se o valor da estatística do teste for maior ou igual que C_2 ou menor ou igual que C_1, considerando o nível de significância $\alpha = 0,05$.

Para cada uma das estatísticas propostas e para cada valor do parâmetro η, a probabilidade de rejeição empírica foi calculada da seguinte forma:

1. Através dos passos de 1 a 5 da estrutura apresentada na Seção 6.2.2, a curva média estimada é calculada usando um parâmetro η para a função $f(t) = 2 - 5t + 5\exp[-100(t-\eta)^2]$ e $n = 1000$;

2. O valor θ da estatística do teste é calculado, lembrando que sob H_0 $\mu(t) = f(t) = 2 - 5t + 5\exp[-100(t-0,5)^2]$;

3. Se $\theta \geq C_2$ ou se $\theta \leq C_1$ rejeita-se H_0;

4. Os passos 1,2 e 3 são repetidos um número $R = 5000$ de vezes;

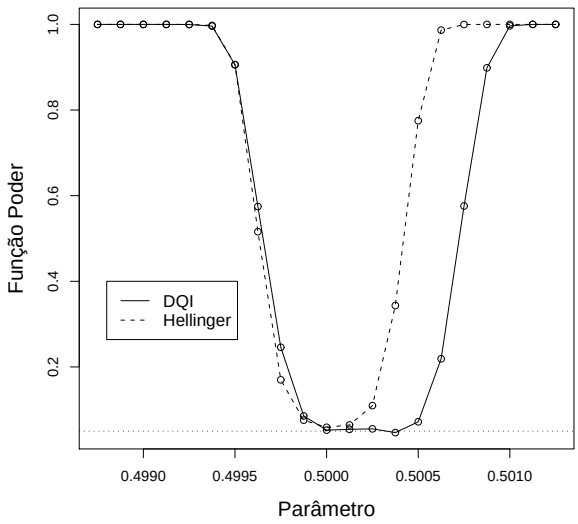


Figura 6.7: Funções poder utilizando as estatísticas DQI e Hellinger. Testando $H_0 : \mu(t) = f(t)\ vs\ H_1 : \mu(t) \neq f(t)$, onde $\mu(t) = 2 - 5t + 5\exp[-100(t - 0,5)^2]$ e $f(t) = 2 - 5t + 5\exp[-100(t - \eta)^2]$ com η variando; $\sigma = 0,25$.

5. A probabilidade de rejeição empírica é dada pelo número de vezes que H_0 foi rejeitada $\div$ 5000.

A Figuras 6.7 apresenta as funções poder estimadas das estatísticas DQI e Hellinger utilizando as respectivas probabilidades de rejeição empíricas.

Mais detalhes e resultados podem ser vistos em Souza (2008).

6.3 Um estudo sobre a distribuição da DQI

Considere a diferença quadrática integrada (DQI) entre duas funções x e y como sendo

$$DQI = \int (x - y)^2. \tag{6.14}$$

Note que agora não será considerada nenhuma transformação.

Lembrando que existe o interesse em testar a hipótese nula $H_0 : \mu(t) = f(t)$ contra a alternativa $H_1 : \mu(t) \neq f(t)$, o objetivo aqui é estudar a distribuição da DQI quando H_0 é verdadeira. A estatística do teste é dada por

$$DQI = \int (\hat{\mu} - f)^2, \tag{6.15}$$

onde $\hat{\mu}$ é a estimativa da curva média. Lembre que os dados funcionais são descritos da seguinte forma:

$$\mathbf{y}_i = f(\mathbf{t}) + \boldsymbol{\epsilon}_i = \boldsymbol{\Phi}\mathbf{c} + \boldsymbol{\epsilon}_i, \;\; i = 1, \ldots, n \tag{6.16}$$

onde $\mathbf{y}_i$ é o vetor $m \times 1$ de observações do i-ésimo dado funcional, $\boldsymbol{\Phi}$ é a matriz $m \times K$ de funções base, $\mathbf{c}$ é o vetor de coeficientes de tamanho K e $\boldsymbol{\epsilon}_i$ é o vetor $m \times 1$ de erros aleatórios. Assuma que $\boldsymbol{\epsilon}_i \sim N_m(\mathbf{0}; \sigma^2\mathbf{I}_m)$. Os dados são suavizados utilizando a técnica de suavização *spline*. Assim, cada curva suave é dada por

$$\hat{\mathbf{y}}_i = \boldsymbol{\Phi}\hat{\mathbf{c}}_i, \tag{6.17}$$

onde $\hat{\mathbf{c}}_i = \mathbf{A}\mathbf{y}_i$, sendo $\mathbf{A} = (\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda\mathbf{R})^{-1}\boldsymbol{\Phi}'$. A matriz $\mathbf{A}$ pode ser facilmente calculada, pois através dos procedimentos `smooth.basis` e `create.bspline.basis` do pacote de ADF do R é possível obter as matrizes $\boldsymbol{\Phi}$ e $\mathbf{R}$, bem como o vetor de coeficientes estimados $\hat{\mathbf{c}}_i$. Como foi assumido que os erros são independentes e distribuídos segundo uma normal com média zero e variância constante, tem-se que

$$\hat{\mathbf{c}}_i \;\sim\; N_K(\mathbf{A}\boldsymbol{\Phi}\mathbf{c}, \mathbf{D}), \tag{6.18}$$

onde $\mathbf{D} = \mathbf{A}\mathbf{A}'\sigma^2$.

A estimativa da curva média $\hat{\mu}$ é a curva média obtida através da amostra de n curvas suaves, ou seja,

$$\hat{\mu}(\mathbf{t}) = \boldsymbol{\Phi}\tilde{\mathbf{c}}, \tag{6.19}$$

onde

$$\tilde{\mathbf{c}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{c}}_i. \tag{6.20}$$

Dessa forma, $\tilde{\mathbf{c}} \sim N_K(\mathbf{A\Phi c}, n^{-1}\mathbf{D})$ e, portanto, tem-se que

$$T = n(\tilde{\mathbf{c}} - \mathbf{A\Phi c})'\mathbf{D}^{-1}(\tilde{\mathbf{c}} - \mathbf{A\Phi c}) \sim \mathcal{X}^2_K, \tag{6.21}$$

ou seja, a forma quadrática (6.21) tem distribuição qui-quadrado com K graus de liberdade, onde K é o número de funções base utilizados no ajuste suave dos dados.

A estatística DQI pode ser aproximada de forma matricial como

$$DQI = \Delta t[(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c})'(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c})], \tag{6.22}$$

onde Δt é o comprimento (sempre igual) do intervalo entre uma observação e outra.

Agora considere que

$$(\tilde{\mathbf{c}} - \mathbf{A\Phi c})' = (\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi A\Phi c})'\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}$$

e seja $\mathbf{W} = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{D}^{-1}(\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'$. Assim, é possível reescrever (6.21) como

$$\begin{aligned} T^* &= n(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi A\Phi c})'\mathbf{W}(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi A\Phi c}) \\ &= n(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c} + \mathbf{\Phi c} - \mathbf{\Phi A\Phi c})'\mathbf{W}(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c} + \mathbf{\Phi c} - \mathbf{\Phi A\Phi c}) \\ &= n(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c}) + 2n(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{\Phi c} - \mathbf{A\Phi c}) + \\ &+ n(\mathbf{\Phi c} - \mathbf{A\Phi c})'\mathbf{W}(\mathbf{\Phi c} - \mathbf{A\Phi c}) \end{aligned} \tag{6.23}$$

Denote o segundo termo em (6.23) por PC (produto cruzado) e o terceiro e último termo por uma constante C. Todas essas quantidades podem ser calculadas, lembrando que $f(\mathbf{t}) = \mathbf{\Phi c}$. Assim, como $T = T^*$, tem-se que

$$T^* = n(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{\Phi}\tilde{\mathbf{c}} - \mathbf{\Phi c}) + PC + C \sim \mathcal{X}^2_K. \tag{6.24}$$

Note que o primeiro termo de T^* é semelhante à estatística DQI em (6.22), bastando multiplicar DQI por $n\mathbf{W}$ e dividir por Δt para obter o primeiro termo. É importante dizer mais uma vez que usando o pacote de ADF do programa R é possível calcular todas as quantidades de T^*.

6.4 Aplicação: Uma extensão para duas amostras

6.4.1 Introdução ao problema e estatísticas do teste

Na Seção 1.6 foram apresentados alguns resultados sobre o crescimento humano. Sabe-se que o estudo do crescimento humano é fundamental para definir o que é um crescimento normal. Registros da altura de 39 garotos e 54 garotas foram coletados durante 18 anos. Esses dados correspondem ao Estudo de Crescimento de Berkeley. Eles estão publicados e, portanto, disponíveis gratuitamente.

Os dados foram suavizados através do método de suavização *spline* utilizando a quarta derivada na penalização (ver Seção 4.2.2). O parâmetro de suavização usado foi $\lambda = 0,1$. Os B-*splines* foram de ordem 6 com nós nos próprios pontos de observação. Os resultados para as 10 primeiras garotas e 10 primeiros garotos podem ser vistos na Figura 6.8.

Após a suavização dos dados, a curva média estimada das garotas e dos garotos foi obtida (ver Seção 2.5). A Figura 6.9 apresenta o resultado obtido.

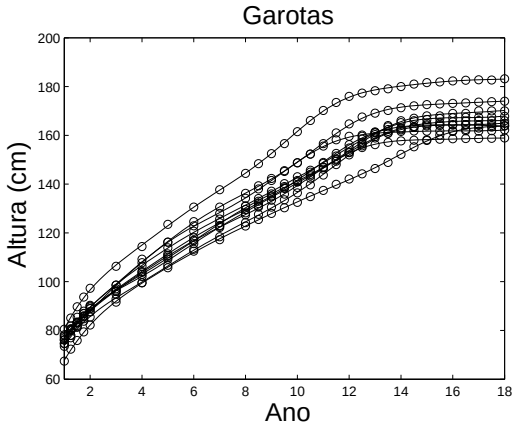
Sejam as curvas $\{X_1(t), \ldots, X_{54}(t) : t \in (0, 18]\}$ e $\{Y_1(t), \ldots, Y_{39}(t) : t \in (0, 18]\}$ amostras independentes de processos estocásticos $\mathcal{X}$ e $\mathcal{Y}$, respectivamente, onde $\mathcal{X}$ representa a altura das garotas e $\mathcal{Y}$ a altura dos garotos. Sabe-se que ambos os processos $\mathcal{X}$ e $\mathcal{Y}$ apresentam $\mu_X(t)$ e $\mu_Y(t)$, respectivamente, como curvas médias. Deseja-se testar a hipótese de que a curva média da altura das garotas é igual a dos garotos, ou seja, deseja-se testar a hipótese nula $H_0 : \mu_X(t) = \mu_Y(t)$ contra a alternativa $H_1 : \mu_X(t) \neq \mu_Y(t)$.

Para isso, é necessário encontrar uma estatística do teste apropriada. Sendo assim, usando as curvas médias estimadas $\hat{\mu}_X$ e $\hat{\mu}_Y$, as transformações $t_{\hat{\mu}_X}$ e $t_{\hat{\mu}_Y}$ e as estatísticas de Hellinger, DQI, KL, $L1$ e afinidade descritas anteriormente, tem-se as seguintes estatísticas do teste:

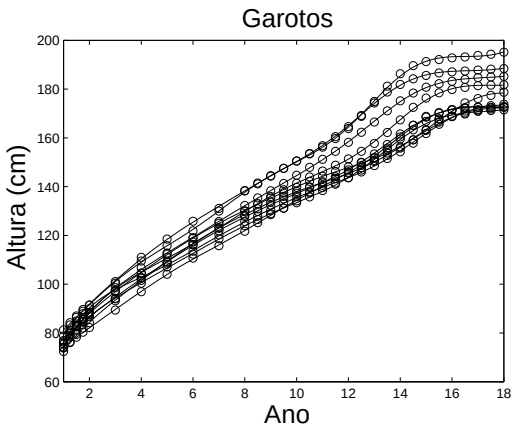
$$\text{Hellinger} = \left(\int \left(\sqrt{t_{\hat{\mu}_X}} - \sqrt{t_{\hat{\mu}_Y}} \right)^2 \right)^{1/2}, \tag{6.25}$$

$$DQI = \int (t_{\hat{\mu}_X} - t_{\hat{\mu}_Y})^2, \tag{6.26}$$

$$KL = \int (\log t_{\hat{\mu}_X} - \log t_{\hat{\mu}_Y}) t_{\hat{\mu}_X}, \tag{6.27}$$



(a)



(b)

Figura 6.8: Alturas das 10 primeiras garotas (a) e dos 10 primeiros garotos (b) do Estudo de Crescimento de Berkeley. Os círculos indicam os dados observados. Cada curva sólida é o ajuste suave aos dados obtido através da penalização da não suavidade usando a quarta derivada.

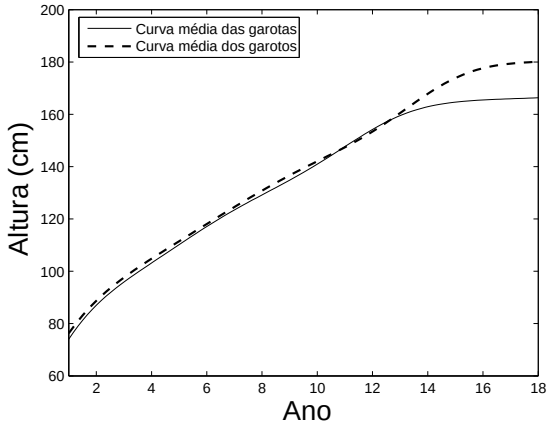


Figura 6.9: A linha sólida corresponde à curva média estimada suave das 54 garotas. Já a linha tracejada corresponde à curva média estimada suave dos 39 garotos.

$$L_1 = \int |t_{\hat{\mu}_X} - t_{\hat{\mu}_Y}| \tag{6.28}$$

e

$$\text{Afinidade} = \int \sqrt{t_{\hat{\mu}_X} t_{\hat{\mu}_Y}}. \tag{6.29}$$

Para decidir se a hipótese nula será ou não rejeitada, o procedimento *bootstrap* de reamostragem será utilizado, porém é importante lembrar que agora as amostras são formadas por curvas.

6.4.2 O procedimento *bootstrap*

O procedimento *bootstrap* de reamostragem é um método computacionalmente intensivo que pode ser utilizado para testar a hipótese de existência de alguma relação estocástica específica entre dois conjuntos de variáveis aleatórias. No procedimento *bootstrap*, amostras artificiais são obtidas a partir das amostras verdadeiras através de reamostragem com reposição. Efron e Tibshirani (1993) e Noreen (1989) apresentam uma discussão mais detalhada sobre testes de hipóteses utilizando o procedimento *bootstrap* para o caso de amostras aleatórias univariadas. Tomando como base o método *bootstrap* para amostras

univariadas, será apresentado a seguir um procedimento *bootstrap* para testar a hipótese de que duas amostras de curvas possuem a mesma curva média.

Sejam as curvas $\{X_1(t), \ldots, X_{n_1}(t) : t \in \mathcal{T}\}$ e $\{Y_1(t), \ldots, Y_{n_2}(t) : t \in \mathcal{T}\}$ duas amostras independentes de processos estocásticos $\mathcal{X}$ e $\mathcal{Y}$, respectivamente e seja θ o valor da estatística do teste a ser utilizada, assim tem-se o seguinte procedimento:

1. O valor θ da estatística do teste é calculado usando as duas amostras;
2. As duas amostras são combinadas formando uma única amostra de tamanho $n_1 + n_2$;
3. Uma nova amostra de curvas de tamanho $n_1 + n_2$ é obtida **com reposição** a partir da amostra combinada. As n_1 primeiras curvas são nomeadas $X_1^*(t), \ldots, X_{n_1}^*(t)$ e as n_2 curvas seguintes são $Y_1^*(t), \ldots, Y_{n_2}^*(t)$;
4. Calcula-se o valor θ^* da estatística utilizando essas duas novas amostras;
5. Os passos 3 e 4 são repetidos um número B vezes. Então B valores de θ^* são obtidos;
6. Para um nível de significância α especificado (por exemplo $\alpha = 5\%$) a hipótese $H_0 : \mu_X(t) = \mu_Y(t)$ é rejeitada se

$$p\text{-valor} = \frac{(\#\{\theta^* \geq \theta\} + 1)}{B + 1} \leq \alpha.$$

   Em particular, para a afinidade, a hipótese nula é rejeitada se

$$p\text{-valor} = \frac{(\#\{\theta^* \leq \theta\} + 1)}{B + 1} \leq \alpha.$$

É importante dizer que o p-valor proposto no item 6 está relacionado com o fato de que os valores das estatísticas KL, L_1, Hellinger e DQI tendem a ser maiores quando a hipótese nula é falsa. Já para a afinidade ocorre exatamente o contrário: valores maiores da afinidade trazem menos evidências contra hipótese nula. Por isso, há um destaque no item 6 para o p-valor quando a estatística utilizada é a afinidade. Para mais detalhes sobre o critério de decisão adotado, ver Noreen (1989).

6.4.3 O poder do teste utilizando *bootstrap*

O poder do teste pode ser estudado através da função poder do teste, que é definida como a probabilidade de rejeitar H_0. Quando a hipótese nula é verdadeira o valor da função poder deve ser igual ou bem próximo ao nível de significância do teste. Por sua vez, quando H_0 é falsa o valor da função poder corresponde à probabilidade de rejeitar H_0 quando a mesma é falsa, ou seja, ao poder do teste. A probabilidade de rejeição empírica é utilizada como uma estimativa da função poder do teste.

Para cada uma das estatísticas do teste propostas a probabilidade de rejeição empírica pode ser calculada da seguinte forma:

1. Duas amostras de curvas $\{X_1(t), \ldots, X_{n_1}(t)$ e $\{Y_1(t), \ldots, Y_{n_2}(t)$ são geradas através da estrutura apresentada na Seção 6.2.2, sendo $\mu_X = f$ e $\mu_Y = g$.

2. As curvas médias estimadas $\hat{\mu}_X$ e $\hat{\mu}_Y$ são calculadas;

3. A hipótese $H_0 : \mu_X = \mu_Y$ é testada utilizando as estatísticas do teste propostas na Seção 6.4.1 e o procedimento *bootstrap* descrito na seção anterior;

4. Os três passos anteriores são repetidos R vezes;

5. A probabilidade de rejeição empírica é dada pelo número de vezes que H_0 foi rejeitada $\div R$.

Souza (2009) apresenta probabilidades de rejeição empíricas para as diferentes estatísticas do teste utilizando diferentes funções f e g e diferentes tamanhos amostras n_1 e n_2. Foi possível verificar que:

- as estatísticas de Hellinger e afinidade apresentam as mesmas probabilidades de rejeição empíricas, o que já era esperado por serem duas estatísticas relacionadas;

- em todos os casos a probabilidade de rejeitar a hipótese nula quando ela é verdadeira ficou bem próxima do nível de significância 0,05 estabelecido para o teste;

- o tamanho amostral é importante para que se tenha um poder maior do teste;
- para amostras maoires ($n_1 = n_2 = 250$) o poder do teste para cada estatística proposta é muito semelhante e satisfatório para pequenas diferenças entre as curvas médias μ_X e μ_Y;
- para um menor tamanho amostral, as maiores probabilidades de rejeição empíricas foram obtidas para as estatísticas L_1 e DQI.

A Figura 6.10 apresenta as funções poder estimadas referentes a um dos exemplos simulados por Souza (2009).

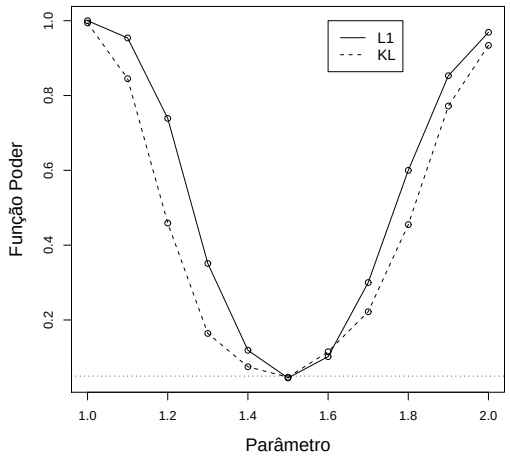
6.4.4 Resultados

A Tabela 6.2 apresenta os valores das diferentes estatísticas do teste obtidos para testar a hipótese de que a curva média da altura das garotas é igual à dos garotos. O procedimento *bootstrap* foi utilizado considerando $B = 2500$. Todas as estatísticas apresentaram p-valor< 0.001. Portanto, considerando o nível de significância $\alpha = 5\%$, a hipótese de que a curva média da altura das garotas é igual à dos garotos é rejeitada com base em todas as estatísticas. Isso já era esperado, pois existe uma conjectura na medicina de que as meninas crescem de forma diferente dos meninos.

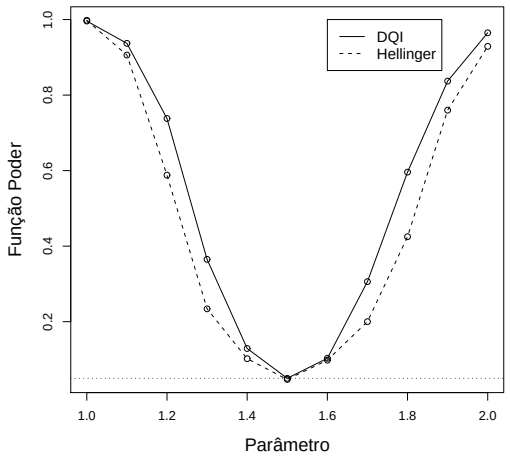
O código em R utilizado nessa aplicação encontra-se disponível em Souza (2008).

Tabela 6.2: Valores das estatísticas encontrados para testar $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X \neq \mu_Y$.

Estatística	Valor obtido
L_1	0,0520
DQI	0,0003
Hellinger	0,0293
Afinidade	0,9996
KL	0,0017



(a)



(b)

Figura 6.10: (a) Funções poder utilizando as estatísticas L_1 e KL. (b) Funções poder utilizando as estatísticas DQI e Hellinger. Testando $H_0 : \mu_X(t) = \mu_Y(t)$ vs $H_1 : \mu_X(t) \neq \mu_Y(t)$, onde $\mu_X(t) = 1,5 \times 0,4(1,5t)^{0,4-1} \exp\{-(1,5t)^{0,4}\}$ e $\mu_Y(t) = \alpha 0,4(\alpha t)^{0,4-1} \exp\{-(\alpha t)^{0,4}\}$ com α variando e $n_1 = n_2 = 250$; $\sigma = 0,25$.

Referências Bibliográficas

[1] Allen, D. L. (1997) Hyphotesis testing using and L_1-distance bootstrap. *The American Statistician*, **51**(2), 145-150.

[2] Anselmo, C. A. F., Dias, R. e Garcia, N. (2005) Adaptive basis selection for functional data analysis via stochastic penalization. *Computational and Applied Mathematics*, **24**(2), 209-229.

[3] 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**, 377-403.

[4] De Boor, C. (2001) *A Practical Guide to Splines*, Revised Edition, New York: Springer-Verlag.

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

[6] Dias, R. (2002) A review of non-parametric curve estimation methods with application to Econometrics. *Economia* **3**(1), 31-75.

[7] Dias, R. e Garcia, N. (2007) Consistent estimator for basis selection based on a proxy of the Kullback-Leibler distance. *Journal of Econometrics*, **141**(1), 167-178.

[8] Dias, R. e Garcia, N. (2005) A spline approach to nonparametric testing of hypothesis. *Brazilian Journal of Probability and Statistics*, **18**, 53-68.

[9] Dias, R. e Garcia, N.L. e Martarelli, A. (2009) Non-parametric estimation for aggregated functional data for electric load monitoring. *Environmetrics*, **20**(2), 111-130.

[10] Draper, N. R. e Smith, H. (1998) *Applied Regression Analysis* Third Edition, New York: Wiley.

[11] Efron, B. e Tibshirani R. J. (1993) *An Introduction to the Bootstrap*, New York: Chapman & Hall.

[12] Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

[13] Fan, J. and Zhang, J. T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc. B*, **62**(2), 303-322

[14] Ferraty, F. e Vieu P. (2006) *Nonparametric Functional Data Analysis*, New York: Springer.

[15] Ferraty, F. e Vieu, P. (2004) Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Journal of nonparametric statistics*, **16**(1), 111-126.

[16] Friedman, J. e Tibshirani, R. J. (1984) The monotone smoothing of scatterplots. *Technometrics*, **26**, 243-250.

[17] Härdle, W. (1991) *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

[18] Härdle, W. (1990) *Smothing Techniques With Implementation in S*, New York: Springer-Verlag.

[19] Gijbels, I. e Heckman, N. (2004) Nonparametric testing for a monotone hazard function via normalized spacings. *Journal of Nonparametric Statistics*, **16**, 463-478.

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

[21] Guo, W. (2004) Functional data analysis in longitudinal settings using smoothing splines. *Statistical Methods in Medical Research. An International Review Journal*, **13**(1), 49-62.

[22] Hall, P., Müller, H. G. e Wang, J. L. (2006) Properties of principal component methods for functional and longitudinal data Analysis. *The Annals of Statistics*, **34**(3), 1493-1517.

[23] Jorgensen, B. e Goegebeur, Y. (2007) *Multivariate Data Analysis and Chemometrics*, `http://statmaster.sdu.dk/courses/ST02`, University of Southern Denmark, Department of Statistics.

[24] Kelly, C e Rice, J. R. (1990) Monotone smoothing with application to dose response curves and the assessment of synergism. *Biometrics*, **46**, 1071-1085.

[25] Nielson, G. M. (1974) Multivariate smoothing and interpolating splines. *SIAM Journal on Numerical Analysis*, **11**, 435-446.

[26] Noreen, E. W. (1989) *Computer Intensive Methods for Testing Hyphotheses: An Introduction*, New York: Wiley.

[27] Qi Li (1996) Nonparametric testing of closeness between two unknown distribution functions. *Econometrics Reviews*, **15**(3), 261-274.

[28] Ramsay, J. O. (1998) Estimating smooth monotone functions. *J. R. Statist. Soc. B*, **60**(2), 365-375.

[29] Ramsay, J. O. (1988) Monotone regression splines in action (with discussion). *Statist. Sci.*, **3**, 425-461.

[30] Ramsay, J. O. e Dalzell, C. J. (1991) Some tools for functional data analysis. *J. R. Statist. Soc. B*, **53**(3), 539-572.

[31] Ramsay, J. O., Hooker, G. e Graves, S. (2009) *Functional Data Analysis with R and MATLAB*, Springer Verlag.

[32] Ramsay, J. O. e Silverman, B. W. (2006) *Functional Data Analysis*, 2nd Edition, New York: Springer.

[33] Ramsay, J. O. e Silverman, B. W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*, New York: Springer.

[34] Ramsay, J. O. e Silverman, B. W. (1997) *Functional Data Analysis*, New York: Springer.

[35] Rice, J. A. (2004) Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, **14**, 613-629.

[36] Ruggiero, M. A. G., Lopes, V. L. R. (1997) *Cálculo Numérico - Aspectos Teóricos e Computacionais*, Segunda Edição, Pearson Makron Books.

[37] Saraiva, M. A. (2009) *Análise Não-Paramétrica de Dados Funcionais: Uma Aplicação à Quimiometria*. Dissertação (Mestrado), Orientador: Prof. Ronaldo Dias, IMECC-UNICAMP, Campinas-SP.

[38] Schmidt, A. M., Conceição, M. de F. ad. G. e Moreira, G. A. (2008) Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications. *Journal of Statistical Computation and Simulation*, **78**, 681-699.

[39] Schoenberg, I. (1964a) Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.*, **52**, 947-950.

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

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

[42] Seber, G. A. F. (1977) *Linear Regression Analysis*, New York: Wiley.

[43] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

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

[45] Souza, C. P. E. (2008) *Testes de hipóteses para dados funcionais baseados em distâncias: um estudo usando splines*. Dissertação (Mestrado), Orientador: Prof. Ronaldo Dias, IMECC-UNICAMP, Campinas-SP.