UNIVERSIDADE
BEIRA INTERIOR

Classificador de Artrite com Dados Funcionais

Versão Final após Defesa

Abrão Beneditus José Irénia Marques

Dissertação para obtenção do Grau de Mestrado em
Matemática e Aplicações
(2º ciclo de estudos)

Orientador: Prof. Doutor Jorge Manuel dos Reis Gama
Coorientador: Prof. Doutor José Aurélio Marques Faria

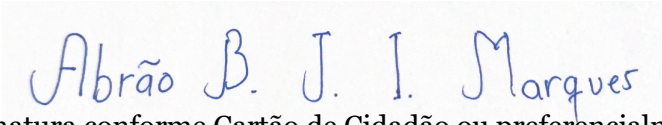
Covilhã, novembro de 2025

Declaração de Integridade

Eu, Abrão Beneditus José Irénia Marques, que abaixo assino, estudante com o número de inscrição M12990 do Mestrado em Matemática e Aplicações da Faculdade de Ciências, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o **Código de Integridades da Universidade da Beira Interior**.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 13/11/2025



(assinatura conforme Cartão de Cidadão ou preferencialmente
assinatura digital no documento original se naquele mesmo formato)

Dedicatória

Dedico este trabalho aos meus pais, Fernando e Luíza, cujo amor, dedicação e fé moldaram o meu caminho e continuam a inspirar-me. À minha esposa, Netty, pelo amor e apoio incansável em cada etapa. E aos meus filhos, Zea e Trezio, minha maior motivação que este percurso vos inspire a sonhar alto e a valorizar sempre o conhecimento.

Agradecimentos

Em primeiro lugar, expresso a minha profunda gratidão a Deus pela vida, proteção, bênçãos e por cada oportunidade que me foi concedida. Reconheço e valorizo imensamente o esforço e as orações incansáveis dos meus pais, que foram pilares fundamentais para o meu crescimento e desenvolvimento ao longo de esta jornada.
Agradeço sinceramente ao Instituto São João de Brito (ISJB), em Timor-Leste, pela valiosa oportunidade de lecionar, experiência que contribuiu de forma significativa para a minha formação pessoal e profissional. Expresso igualmente a minha gratidão ao Governo de Timor-Leste por todo o apoio financeiro que possibilitou a minha vinda e os meus estudos em Portugal.
À Universidade da Beira Interior (UBI), endereço o meu sincero agradecimento por acolher a minha candidatura e por proporcionar-me um ambiente académico rico e estimulante. Agradeço a todos os docentes que me acompanharam e partilharam os conhecimentos necessários à conclusão deste mestrado.
Expresso o meu reconhecimento ao meu coorientador, pela generosa disponibilização dos dados essenciais e pela valiosa colaboração na revisão desta dissertação.
Ao meu orientador, manifesto profunda gratidão pela paciência, dedicação e acompanhamento ao longo de todo o processo. O seu apoio constante foi determinante para a concretização deste percurso.
Por fim, deixo um agradecimento especial a todos que fizeram parte do meu quotidiano em Portugal, oferecendo-me apoio, amizade e companhia ao longo desta etapa. A todos, o meu mais sincero obrigado.

Resumo

A análise de dados funcionais tem ganhado destaque nas últimas décadas como uma abordagem poderosa para lidar com dados que podem ser representados por funções contínuas, como o movimento ao longo do tempo. Este trabalho apresenta uma revisão das técnicas estatísticas para a análise de dados funcionais e consequente aplicação a dados do momento de força do tornozelo direito e esquerdo, obtidos durante o caminhar, em mulheres com e sem artrite reumatoide. Adicionalmente, foi desenvolvido um classificador de artrite reumatoide com um modelo de regressão logística funcional.
Espera-se que os resultados contribuam para o uso de métodos estatísticos funcionais na biomecânica.

Palavras-chave

Análise de dados funcionais, artrite reumatoide, bases de funções, suavização, estatísticas funcionais, modelo linear funcional.

Abstract

Functional data analysis has gained prominence in recent decades as a powerful approach to dealing with data that can be represented by continuous functions, such as movement over time. This work presents a review of statistical techniques for the analysis of functional data and consequent application to force moment data of the right and left ankle, obtained during walking, in women with and without rheumatoid arthritis. Additionally, a rheumatoid arthritis classifier with a functional logistic regression model was developed. It is expected that the results will contribute to the use of functional statistical methods in biomechanics.

Keywords

Functional data analysis, rheumatoid arthritis, basis functions, smoothing, functional statistics, functional linear model.

Índice

1 Introdução **1**

1.1 Contextualização do tema . . . 1
1.2 Problema a investigar . . . 2
1.3 Objetivos . . . 2
1.3.1 Objetivo geral . . . 2
1.3.2 Objetivos específicos . . . 2
1.4 Fundamentação do estudo . . . 2
1.5 Amostra . . . 2
1.6 Avaliação biomecânica do caminhar . . . 3
1.7 Dados utilizados no estudo . . . 3
1.8 Análise estatística . . . 3
1.9 Estrutura deste trabalho . . . 4

2 Bases de Funções **5**

2.1 Notações . . . 5
2.2 Dados funcionais . . . 5
2.3 Representação de uma função por uma base de funções . . . 6
2.3.1 Função spline . . . 8
2.3.2 Função B-spline . . . 8
2.3.3 Séries de Fourier . . . 9
2.4 Derivadas . . . 10

3 Suavização **13**

3.1 Suavização pelo método dos mínimos quadrados . . . 13
3.2 Suavização pelo método dos mínimos quadrados penalizados . . . 13
3.3 A escolha do parâmetro λ . . . 15

4 Estatística de Dados Funcionais **17**

4.1 Média e variância amostral . . . 17
4.2 Covariância e correlação . . . 18
4.3 Consistência da função média e covariância amostral . . . 19
4.4 Intervalo de confiança para uma média . . . 21
4.5 Teste t . . . 21

5 Alinhamento **25**

5.1 Padronização Linear do Tempo . . . 26
5.2 Alinhamento por Pontos de Referência Utilizando Função de Distorção Temporal . . . 27
5.3 Alinhamento Contínuo . . . 28
5.4 Decomposição da Variância em Termos da Amplitude e Fase . . . 31

6 Modelos Lineares Funcionais **33**
6.1 Estimação de $\beta_1(t)$ 34
6.2 Intervalo de confiança para $\beta_1(t)$ 36
6.3 O Modelo de Regressão Logística Funcional 36
6.4 Classificador de Artrite 37

Bibliografia **47**

A O Espaço de Hilbert L^2 **51**

B Cálculo da Média das 6 Réplicas **53**

C Exemplos de Bases de Funções **55**

D Estimação de λ por GCV **57**

E Gráficos das Curvas **59**

F Alinhamentos por Deslocamento e Contínuo **61**

G Regressão Logística Funcional **65**

Lista de Figuras

2.1 Dez funções B-spline de grau 3 (ordem 4) no intervalo $[0, 1]$. 9
2.2 Cinco primeiras funções base de Fourier no intervalo $[0,1]$ (2 senos, 2 cossenos e a constante). 10

3.1 Curvas do momento de força dos grupos artrítico (A) e controlo (B). 16

4.1 Médias funcionais do momento de força de cada grupo. 17
4.2 Superfície das variâncias-covariâncias do momento de força entre os indivíduos do grupo artrítico (lado esquerdo) e respetivas curvas de nível (lado direito). 18
4.3 Superfície das variâncias-covariâncias do momento de força entre os indivíduos do grupo controlo (lado esquerdo) e respetivas curvas de nível (lado direito). 19
4.4 Resultado do teste de permutações para as médias multivariadas do momento de força dos dois grupos. Foram consideradas 200 permutações aleatórias entre os dois grupos. 23

5.1 Curvas do momento de força dos grupos artrítico (A) e controlo (B) no intervalo $[0; 0{,}950]$ segundos. 27
5.2 Curvas do momento de força dos grupos artrítico (A) e controlo (B) no intervalo $[0; 0{,}950]$ segundos após o alinhamento contínuo. 29
5.3 Resultado do teste de permutações para as médias multivariadas do momento de força dos dois grupos após o alinhamento contínuo. Foram consideradas 200 permutações aleatórias entre os dois grupos. 30
5.4 Médias funcionais do momento de força de cada grupo após o alinhamento contínuo dos dados. 30

6.1 $\beta(t)$ do modelo logístico funcional com o momento de força articular do tornozelo direito. Para este modelo a constante, β_0, foi estimada em $4{,}140$. 38
6.2 $\beta(t)$ do modelo logístico funcional com o momento de força articular do tornozelo esquerdo. Para este modelo a constante, β_0, foi estimada em $3{,}705$. 38
6.3 Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo direito. 40
6.4 Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo esquerdo. 40
6.5 Resíduos estandardizados de Pearson do modelo logístico funcional com o momento de força articular do tornozelo direito. 41
6.6 Resíduos estandardizados de Pearson do modelo logístico funcional com o momento de força articular do tornozelo esquerdo. 42

6.7 $\beta(t)$ do modelo logístico funcional, com o momento de força articular do tornozelo direito, obtido com a exclusão dos dois *outliers*. Para este modelo a constante, β_0, foi estimada em $11{,}508$. 43
6.8 $\beta(t)$ do modelo logístico funcional, com o momento de força articular do tornozelo esquerdo, obtido com a exclusão dos dois *outliers*. Para este modelo a constante, β_0, foi estimada em $8{,}726$. 44
6.9 Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo direito, quando excluídos os dois *outliers*. 44
6.10 Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo esquerdo, quando excluídos os dois *outliers*. 45

Lista de Acrónimos

CV	Cross-Validation
FDA	Functional Data Analysis
GCV	General Cross-Validation
MSE	Mean Square Error
ROC	Receiver Operating Characteristic
UBI	Universidade da Beira Interior

Capítulo 1

Introdução

1.1 Contextualização do tema

A análise de dados funcionais (*functional data analysis*, FDA) foi popularizada no início dos anos 2000 por J. O. Ramsay e B. W. Silverman [10]. A FDA consiste em técnicas e métodos avançados de estatística, cujos dados a serem analisados são ou podem ser transformados em curvas. Os conceitos elementares ou mais avançados de estatística são, em geral, adaptados à análise de curvas. Conceitos ou técnicas usuais como média, variância, correlação, análise de componentes principais, regressão (métodos dos mínimos quadrados penalizados ou não penalizados, mínimos quadrados parciais, etc.) têm, na FDA, um análogo para curvas. As curvas são geralmente estimadas ou interpoladas recorrendo-se a bases de funções (Fourier, B-spline, wavelets, etc.). Naturalmente, a FDA atingiu um grau de maturidade que permitiu que esta tenha sido aplicada frequentemente em medicina, ciências, economia e engenharia.

Este trabalho tem como primeiro propósito apresentar os conceitos elementares de estatística para dados funcionais, tais como, média, covariância, variância, intervalo de confiança para uma média, teste t e modelos lineares funcionais. Estes modelos serão abordados no capítulo 6. Nos capítulos 2, 3, e 4 serão abordados os conceitos que permitem a abordagem estatística referida a dados funcionais. No capítulo 5 será abordado o alinhamento de curvas, cuja utilização poderá ser facultativa na FDA, mas poderá ter um grande impacto nas suas aplicações.

Um segundo propósito deste trabalho consistiu em aplicar os conceitos estatísticos referidos a dados do momento de força ou torque[1] dos tornozelos direito e esquerdo, obtidos durante o caminhar, com o objetivo de discriminar dois grupos de mulheres. Um grupo foi constituído por mulheres diagnosticadas com artrite reumatoide e o outro foi constituído por mulheres saudáveis, que serviu de controlo. Estes dados foram gentilmente cedidos pelos Professores João Abrantes e Pedro Aleixo da Universidade Lusófona de Humanidades e Tecnologias, Portugal.

Saliente-se que a artrite é um conjunto de condições inflamatórias que afetam as articulações, resultando em dor, rigidez e limitação de movimentos. A artrite afeta milhões de pessoas em todo o mundo, impactando diretamente na qualidade de vida. Existem vários tipos de artrite, mas, neste trabalho, focámos a atenção numa consequência da artrite do tipo reumatoide em mulheres.

[1] Efeito de rotação de uma força num corpo, em torno de um ponto ou eixo.

1.2 Problema a investigar

O problema de investigação que orienta este trabalho é: em que medida as técnicas de FDA podem auxiliar na distinção entre mulheres com e sem artrite reumatoide?

1.3 Objetivos

1.3.1 Objetivo geral

Aplicar técnicas de FDA para distinguir mulheres com e sem artrite reumatoide, através da análise do momento de força articular dos tornozelos direito e esquerdo durante o caminhar.

1.3.2 Objetivos específicos

1. Ampliar as aplicações da FDA no campo da biomecânica;

2. Explorar e adaptar métodos estatísticos da FDA para o tratamento de dados referentes ao momento de força durante o caminhar;

3. Desenvolver um método que permita discriminar mulheres com artrite reumatoide daquelas sem a doença.

1.4 Fundamentação do estudo

Com as ferramentas da FDA, espera-se facilitar o diagnóstico de artrite reumatoide por especialistas.
A FDA oferece um conjunto de ferramentas estatísticas avançadas para tratar dados que apresentam natureza funcional, como os dados do momento de força de um movimento. Utilizando essas técnicas, é possível capturar as variações do momento de força que poderão distinguir os indivíduos com artrite reumatoide daqueles sem esta condição.

1.5 Amostra

A amostra foi composta por 36 mulheres pós-menopáusicas. Destas, 18 apresentavam diagnóstico de artrite reumatoide e foram recrutadas no Instituto Português de Reumatologia, em Lisboa, enquanto as outras 18, sem histórico da doença, foram selecionadas na comunidade como grupo de controlo. Informações detalhadas sobre as características demográficas, bem como sobre o histórico médico e reprodutivo das participantes, encontram-se descritas em [1].

1.6 Avaliação biomecânica do caminhar

Para avaliar os parâmetros biomecânicos do caminhar, foi realizada uma análise tridimensional utilizando o sistema *Vicon Motion Capture MX (Vicon Motion Systems, Oxford, UK)*, sincronizado com uma plataforma de força (AMTI, modelo BP400600). O sistema, calibrado segundo as especificações técnicas da Vicon, registou dados cinemáticos a 200 Hz e forças de reação do solo a 1000 Hz. O modelo *Plug-in Gait Full-Body* foi aplicado para determinar a posição de 39 marcadores refletivos colocados em pontos anatómicos específicos, permitindo modelar os segmentos da perna e do pé e calcular o centro articular do tornozelo. Foram recolhidos dados antropométricos (altura, massa corporal e diâmetros segmentares) para a construção do modelo. As participantes caminharam descalças, à sua velocidade natural, ao longo de uma passadeira de 7 metros com a plataforma de força embutida. Foram recolhidos sete ensaios válidos por pé, sendo considerados apenas os ensaios em que o pé assentava completamente na plataforma. Para evitar fadiga, as participantes descansavam 2 minutos após cada 20 ensaios. A análise concentrou-se na fase de apoio do caminhar (plano sagital), desde o contacto inicial do calcanhar até à saída dos dedos. Vários parâmetros foram avaliados, destacando-se o momento de força (Nm/kg), utilizado no presente estudo. Mais detalhes metodológicos podem ser encontrados em [1].

1.7 Dados utilizados no estudo

Os dados consistiram no momento de força articular dos tornozelos direito e esquerdo das participantes dos dois grupos. O momento de força de cada mulher foi previamente normalizado pelo respetivo peso, sendo a unidade considerada o Newton-metro por quilograma (Nm/kg). Para cada participante, foram registados seis ensaios por tornozelo, a partir dos quais se determinou a média funcional do momento de força. Antes do cálculo das médias, os dados dos seis ensaios foram transformados em curvas com funções B-spline, aplicando-se suavização pelo método dos mínimos quadrados penalizados com um parâmetro reduzido (definido por validação cruzada generalizada), de modo a eliminar ruídos ou erros de observação. Todas as análises foram realizadas a partir destas médias funcionais. Convém frisar que também foram disponibilizados dados relativos ao ângulo articular do tornozelo (em graus). Contudo, estes não foram utilizados neste trabalho, devido à elevada correlação com os dados do momento de força. Optou-se pela análise do momento de força, por se considerar que este parâmetro possui maior poder de discriminação entre indivíduos com e sem artrite reumatoide.

1.8 Análise estatística

A revisão da literatura foi fundamental para fundamentar a escolha das ferramentas estatísticas e computacionais aplicadas no âmbito da FDA. A interpretação dos resultados centrar-se-á nos padrões encontrados nos dados funcionais, destacando as diferenças en-

tre mulheres com e sem artrite reumatoide. Toda a análise estatística foi efetuada no software R (versão 4.4.2), com recurso aos pacotes *fda* [13], (versão 6.2.0) e *refund* [6], (versão 0.1-37). Os códigos correspondentes encontram-se nos Apêndices.

1.9 Estrutura deste trabalho

Este trabalho divide-se em seis capítulos.
No primeiro capítulo, apresentamos a contextualização do tema, os objetivos, a fundamentação, a metodologia e a estrutura da dissertação. O propósito deste capítulo é fornecer uma visão geral do trabalho, situando o leitor quanto ao conteúdo que será desenvolvido ao longo do estudo.
O segundo capítulo introduz a matéria-prima fundamental que utilizámos para a realização da análise estatística: os dados do tipo funcional. Neste capítulo, abordámos sobre o que são os dados funcionais, como obtê-los e como trata-los, para poderem ser chamados dados funcionais. Neste capítulo também discutimos as bases de funções que utilizámos para transformar os dados discretos em dados funcionais.
Após conhecer as bases de funções, necessitamos que esses dados discretos sejam transformados em curvas contínuas suaves, de modo a garantir que as curvas resultantes tenham derivadas de várias ordens. Este assunto é abordado no capítulo três, onde se discute a suavização pelo método dos mínimos quadrados e método dos mínimos quadrados penalizados.
No capítulo 4, apresentamos uma panóplia de estatísticas elementares funcionais. O capítulo também discute brevemente os teoremas que fundamentam a consistência da função média e da covariância amostral, e a confiabilidade dos intervalos de confiança.
No Capítulo 5, discutimos o alinhamento, uma ferramenta essencial para ajustar diferenças de fase e período entre as curvas.
O capítulo final detalha o modelo linear funcional empregado para discriminar os grupos de artrite e controlo. Concluímos este capítulo com a aplicação do modelo a dados reais, focando especificamente no momento de força dos tornozelos direito e esquerdo.
Após a bibliografia, encontram-se os apêndices. O apêndice A aborda de uma forma sucinta o espaço onde podem ser realizadas as inferências estatísticas, isto é, um espaço de Hilbert, especificamente o espaço L^2. Os demais apêndices contêm os códigos R utilizados nesta pesquisa.

Capítulo 2

Bases de Funções

2.1 Notações

Primeiramente vamos considerar certas notações, que utilizaremos a partir deste capítulo. Iremos representar por letras minúsculas, por exemplo, x, um escalar ou uma função $x(t)$ e a negrito, $\mathbf{x}$, um vetor, sendo x_i os seus elementos, e $\mathbf{x}'$ será a versão transposta do vetor $\mathbf{x}$.
As matrizes serão representadas por letras maiúsculas a negrito, por exemplo, $\mathbf{X}$. Se considerarmos a matriz $\mathbf{M}$, a sua transposta é $\mathbf{M}'$ e o traço da matriz $\mathbf{M}$ é $tr\mathbf{M}$. Em certas ocasiões, utilizamos também letras gregas, que representam as mesmas características das letras usuais. Em relação às derivadas, adotámos a notação $D^m x$ ou $x^{(m)}$ para a m-enésima derivada de x.
Para os vetores $\mathbf{x} = (x_1, x_2, ..., x_n)$, e $\mathbf{y} = (y_1, y_2, ..., y_n)$, o produto interno é denotado por $\langle \mathbf{x}, \mathbf{y} \rangle$. Será também obtido o operador tensorial $\otimes$, que se aplica da seguinte forma: dados $\mathbf{x} \in R^m$ e $\mathbf{y} \in R^n$ (vetores coluna), o produto tensorial, $\mathbf{x} \otimes \mathbf{y}$, é uma matriz $m \times n$ dada por:

$$\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}' \tag{2.1}$$

2.2 Dados funcionais

Na FDA, o foco de estudo são os dados que podem ser representados em forma de função contínua em certo domínio, como por exemplo tempo, espaço, frequência, peso, etc. Os dados funcionais resultantes podem ser curvas, superfícies ou mesmo hipersuperfícies. Embora os dados não sejam mais do que observações discretas sobre cada unidade estatística (indivíduo), estes podem ser ajustados por uma função. Obtém-se assim uma única informação para essa unidade estatística, como se fosse um único dado, uma única observação, na estatística tradicional.
Na prática, os dados funcionais são geralmente observados e registados discretamente como n pares (t_j, y_j), e y_j é uma observação de $x(t_j)$, com $j = 1, 2, 3, ..., n$ e $t \in [a,b]$. O resultado destes pontos observados é uma única função x. As amostras de dados na FDA normalmente são coleções de dados funcionais x_i, com $i = 1,2,3, ..., n$. A função observada x_i é formada por n_i pares $(t_{ij}, y_{ij}), j = 1, ..., n_i$.
Neste contexto, acredita-se que existe uma função x, *suave* (ver capítulo 3), que representa os dados observados y, cujos valores são iguais em cada t observado. Isto é o que

interessa para a FDA. Assim, o conjunto seguinte,

$$\{x_i(t) : t \in [a,b], i = 1,2,3,...,n\}, \tag{2.2}$$

é um conjunto de curvas, onde as funções $x_i(t)$ existem em qualquer ponto t, mas somente foram observadas num número finito de pontos discretos $t_{i,j}$. O conjunto (2.2) é um conjunto de curvas.
Vamos considerar, a partir daqui, uma única curva de x nas nossas abordagens. É comum que erros ocorram durante a obtenção dos dados y, o que pode comprometer a suavização da função. Esse tipo de erro é chamado de erro observacional ε. Assim, a relação entre x e y pode ser expressa como

$$y_j = x_j + \varepsilon_j. \tag{2.3}$$

Uma das tarefas é representar os dados como funções suaves e tentar filtrar estes erros. Na forma matricial, pode ser representada por,

$$\mathbf{y} = \mathbf{x}(\mathbf{t}) + \mathbf{e}, \tag{2.4}$$

onde **y**, **x** e **e** são vetores coluna com n linhas e $\mathbf{x}(\mathbf{t})$ são os valores de x em t.
Segundo Dias e Souza [5], a suavização é importante para dados funcionais, porque, normalmente na FDA, interessa também analisar as suas derivadas. Uma das técnicas para obter uma função x suave é escrever x em combinação linear dos elemento de uma base de funções, utilizando as observações discretas efetuadas. Isso garante que a função x possa ser derivável um certo número de vezes.

2.3 Representação de uma função por uma base de funções

Vimos anteriormente que os dados funcionais são recolhidos discretamente. No entanto, na FDA são tratados como objetos de dimensão infinita. Para que se possa realizar as inferências estatísticas, estes dados têm de ser tratados em certos espaços funcionais, normalmente o espaço de Hilbert L^2. De acordo com Kokoszka e Reimherr [10], os dados funcionais, tipicamente uma coleção de funções $x_n \in L^2$ definidas num intervalo comum, são vistos como objetos de dimensão infinita. É depois comum projetar essas funções em subespaços de dimensão finita para facilitar os cálculos estatísticos. Uma abordagem sucinta ao espaço L^2 encontra-se no apêndice A. O espaço de Hilbert permite que uma função possa ser escrita como combinação linear dos elementos em uma base ortonormada. Observe-se que o intervalo comum referido é usualmente um intervalo $[a,b]$, com a e b finitos, já que, os dados discretos observados então num intervalo deste tipo.

Definição 2.3.1. Seja A o conjunto de índices arbitrários. Dizemos que $\{e_i, i \in A\}$ é um sistema ortonormado se

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

Definição 2.3.2. Um espaço de Hilbert, H, é um espaço separável, se existe um sistema ortonormado numerável $\{e_1, e_2, e_3, ...\}$, tal que, para qualquer $x \in H$ admite a expansão

$$x = \sum_{i}^{\infty} a_i e_i,$$

onde a série converge em norma, isto é,

$$x = \lim_{n\to\infty} \sum_{i=1}^{n} a_i e_i.$$

Diz-se que o sistema ortonormado $\{e_i\}$ é *completo* em H quando

$$\overline{\text{span}}\{e_i\} = H.$$

Um sistema ortonormado completo num espaço de Hilbert separável é chamado de *base ortonormada*.

Teorema 2.3.3 (Teorema de Parseval)**.** *Suponha que H é um espaço de Hilbert separável e* $\{e_j : j = 1,2,\ldots\}$ *um sistema ortonormado completo. Então, para qualquer* $x \in H$*,*

$$x = \sum_{j=1}^{\infty} \langle x, e_j \rangle e_j$$

e

$$\|x\|^2 = \sum_{j=1}^{\infty} |\langle x, e_j \rangle|^2 .$$

Uma função x pode ser representada por uma combinação linear de elementos de uma base de funções. Uma base de funções é um conjunto de funções conhecidas, ϕ_k, independentes, que têm propriedades que permitam aproximar uma função com um grande número K de funções dessa base de funções através de uma soma ponderada ou combinação linear. De acordo com Ramsay e Silverman [11], a forma de representar uma função x em termos de K funções ϕ_k de uma base de funções, é dada por,

$$x_K(t) = \sum_{k=1}^{K} c_k \phi_k(t), \tag{2.5}$$

onde $x_K(t)$ é a função estimada obtida, a partir dos dados observados, y, pelos elementos da base de funções, e c_k são os coeficientes. A interpolação ou representação exata é quando $K = n$; neste caso, $x(t_j) = y_j$, para cada j. A suavização dos dados y_j depende da escolha de K. Não existe um método para se encontrar o número K. A sua escolha depende das características dos dados.

A escolha de base de funções também é importante, em particular, se for necessário o cálculo de derivadas. Uma base de funções pode servir para representar uma função, mas não necessariamente a sua derivada.

Existem vários tipos de bases de funções conhecidas, por exemplo, a base da série de Fou-

rier, B-spline, monómios, exponenciais, wavelets, etc. Não existe uma base de funções universal adequada para todos as funções. Ao se escolher uma base, tem de se verificar se a sua estrutura combina com a função a estimar. Isso ajuda muito para que se consiga uma aproximação satisfatória com o menor número K de funções da base de funções. De acordo com Ramsay e Silverman [11], as funções que desejamos modelar tendem a enquadrarem-se em duas principais categorias: periódicas e não periódicas. Para as funções periódicas a série de Fourier é a melhor escolha, enquanto a base de funções spline é mais para as funções não periódicas.

2.3.1 Função spline

As funções spline formam uma base de funções muitas vezes utilizadas para suavizar as curvas na FDA. É usual utilizarem-se polinómios para se aproximar uma curva em determinado intervalo. Assim, partindo de um intervalo $I = [a, b]$, onde a curva está definida, e considerando-se i partições do intervalo I da forma $I_i = [x_i, x_{i+1}]$, podemos aproximar a curva por um polinómio, p_i, em cada intervalo I_i. Assim, resultam as funções polinomiais por partes, $s(.)$, usualmente independentes entre si, e define-se a função spline, que une esses polinómios por partes, da forma seguinte:

$$s(t) = \sum_{i=0}^{m-1} \alpha_i t^i + \sum_{i=1}^{k} \beta_i (t - x_i)^{m-1}, \tag{2.6}$$

onde $\alpha_0, ..., \alpha_{m-1}$ e $\beta_1, ..., \beta_k$ são números reais, $\{1, t, ..., t^{m-1}, (t-x_1)^{m-1}, ..., (t-x_k)^{m-1}\}$ é a base de funções polinomiais e $x_1, ..., x_k$ são os nós interiores. Observe-se que a expressão (2.6) é uma combinação linear de $m + k$ polinómios. O conjunto de funções spline de ordem m e nós interiores em $x_1, ..., x_k$ é chamado de espaço spline e é denotado por $S_m(x_1, ..., x_k)$. Resumindo, a função spline é determinada pela ordem do polinómio em cada intervalo e os nós.

2.3.2 Função B-spline

O termo "B-spline" é uma abreviatura de "*basis spline*". B-spline é uma extensão da função spline e forma uma base no espaço spline. Assim, como é um spline, os B-spline são formados por partes de polinómios que ligam um certo número de nós entre eles. Assim, verifica-se que um B-spline de ordem m é uma junção de m pedaços de funções polinomiais de ordem $m - 1$, contínua e derivável nos $m - 1$ nós.

Tomando-se num intervalo $I = [a, b]$ uma partição de $k + 1$ intervalos da forma $[t_0, t_1], ..., [t_k, t_{k+1}]$, onde $t_1, ..., t_k$ são nós interiores, para o B-spline o número de nós deve ser $k + 2m$, ou seja, adicionam-se mais $m - 1$ nós em cada extremidade com valores arbitrários ou, usualmente, iguais aos valores dos nós das extremidades. A seguinte expressão iterativa, conhecida por algoritmo de Boor [2], define o *i*-ésimo B-spline de ordem m para

uma sequência de nós t_i:

$$B_{i,m}(t) = \frac{t - t_i}{t_{i+m-1} - t_i} B_{i,m-1}(t) + \frac{t_{i+m} - t}{t_{i+m} - t_{i+1}} B_{i+1,m-1}(t). \tag{2.7}$$

Tomando-se para primeiro elemento da iteração, $B_{i,1}$, a função dada por:

$$B_{i,1}(t) = \begin{cases} 1, & \text{se } t_i \leq t \leq t_{i+1}, \\ 0, & \text{caso contrário}, \end{cases} \tag{2.8}$$

o processo de cálculo torna-se mais eficiente.

Considerando-se o B-spline de ordem m com k nós interiores, podemos aproximar uma função x, num dado intervalo I, da seguinte forma:

$$x_K(t) = \sum_{i=1}^{K=m+k} c_i B_{i,m}(t). \tag{2.9}$$

Como exemplo, na figura 2.1 estão representadas 10 funções B-spline de ordem 4 (cúbicas) no intervalo [0, 1], com nós igualmente espaçados, sendo 6 nós interiores (ver código R no apêndice C).

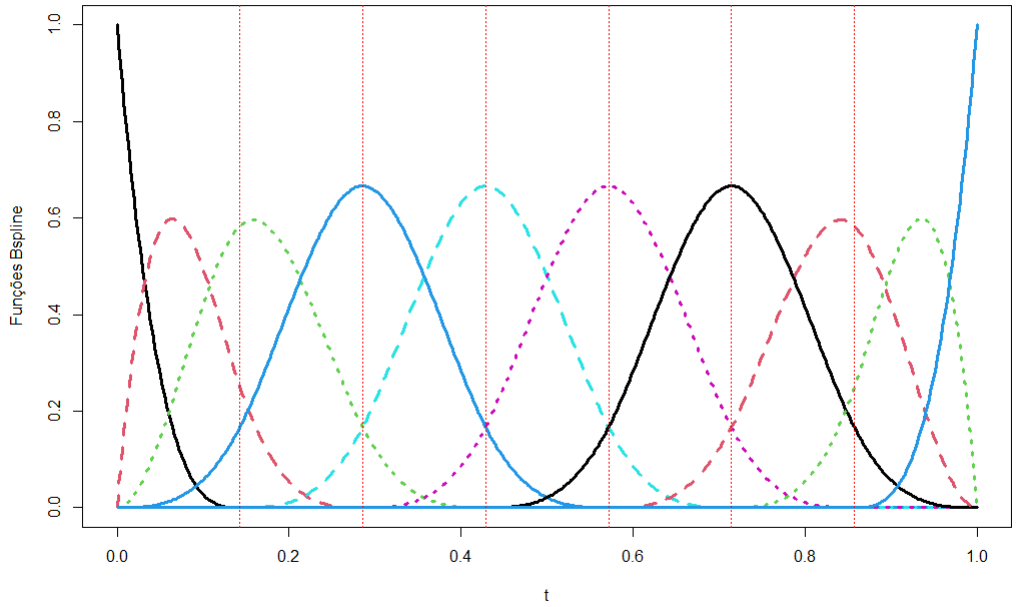


Figura 2.1: Dez funções B-spline de grau 3 (ordem 4) no intervalo $[0, 1]$.

2.3.3 Séries de Fourier

Uma das bases de funções mais conhecidas é a que propicia a bem conhecidas série de Fourier. A base de Fourier é mais utilizada para aproximar funções periódicas, onde o

valor da função no início e fim do intervalo é o mesmo. A sua forma é,

$$x(t) = c_0 + c_1 \sin wt + c_2 \cos wt + c_3 \sin 2wt + c_4 \cos 2wt + \ldots = \sum_{j=0}^{\infty} c_j \phi_j(t), \qquad (2.10)$$

com $\phi_0(t) = 1, \phi_1(t) = \sin wt, \phi_2(t) = \cos wt, \phi_3(t) = \sin 2wt, \phi_4(t) = \cos 2wt$, e assim sucessivamente. O w está relacionado ao período T da seguinte forma:

$$w = \frac{2\pi}{T}. \qquad (2.11)$$

A primeira função na base de Fourier é a função constante, e depois segue com as funções seno e cosseno de período igual à duração do intervalo. Os argumentos dos senos e cossenos estão multiplicados por números naturais, implicando que as sucessivas funções seno e cosseno apresentem período decrescente ou, equivalentemente, de frequência crescente.
Para a aproximação de uma função pela base de Fourier, composta por senos e cossenos, o número de parcelas a considerar em (2.10) será $K = 1 + 2m$, onde m corresponde ao número de pares harmónicos seno-cosseno a incluir.
Na figura 2.2 estão representadas as cinco primeiras funções da base de Fourier (ver código R no apêndice C).

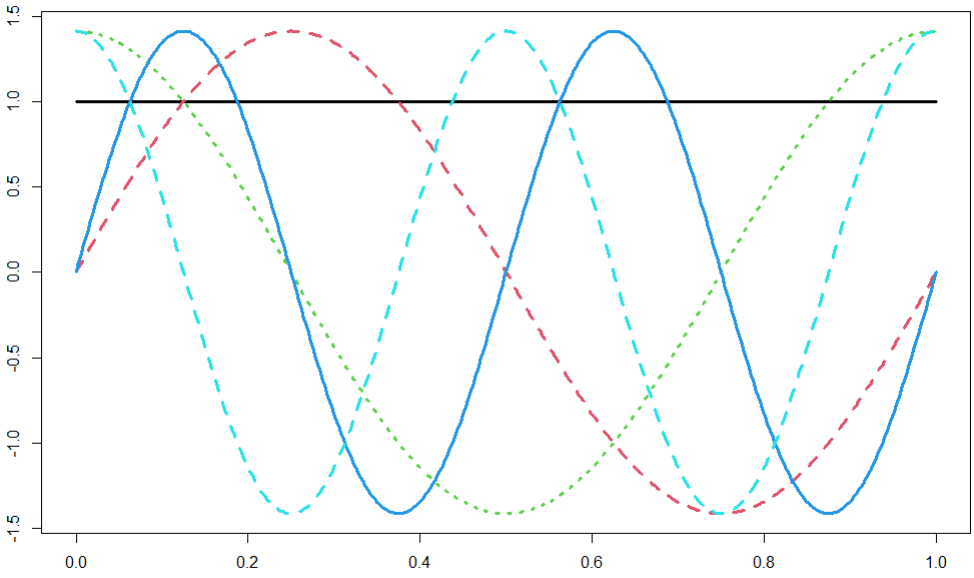


Figura 2.2: Cinco primeiras funções base de Fourier no intervalo $[0,1]$ (2 senos, 2 cossenos e a constante).

2.4 Derivadas

Frequentemente, na FDA as derivadas jogam um papel importante, pois existem muitas aplicações onde interessa estudar alterações ou variações, como no exemplo clássico do

crescimento de crianças [12].
No caso particular do crescimento de crianças, este é a taxa de alteração da altura no instante de tempo t, ou seja, a velocidade, isto é, a primeira derivada da altura. Também pode ter interesse considerar-se a taxa de alteração da velocidade, ou seja, a aceleração, que é a segunda derivada.
Na aproximação de uma função x por uma base de funções $\{\phi_k\}$, a derivada de ordem m da aproximação x_K de x é dada por:

$$x_K^{(m)}(t) = \sum_{k=1}^{K} c_k \phi_k^{(m)}(t). \tag{2.12}$$

Convém salientar que a validade e utilidade da expressão anterior depende da escolha da base de funções. No caso da base de Fourier, a expressão anterior é sempre válida, já que os elementos da base são infinitamente diferenciáveis. Mas, para uma base B-spline de ordem m o segundo membro da expressão (2.12) será nulo a partir da derivada de ordem $m + 1$. Deste modo, em aplicações, a ordem da base B-spline deverá exceder em, pelo menos, duas unidades a ordem da derivada que se pretender considerar.

Capítulo 3

Suavização

3.1 Suavização pelo método dos mínimos quadrados

A suavização pode ocorrer naturalmente quando se utiliza um grande número K de elementos de uma base de funções de modo a que a função aproximada resultante coincida em cada um dos pontos dos dados observados. No entanto, se os dados apresentarem um ruído suficientemente grande ou erros de observação, a utilização de um número grande, K, de elementos da base de funções originará uma aproximação muito ondulada, isto é, a aproximação da função alvo apresentar-se-á sobrestimada.
Segundo Kokoszka e Reimherr [10], uma abordagem simples para a suavização é utilizar um K relativamente pequeno em (2.5). Uma desvantagem desta abordagem é que as funções suaves resultantes são sempre combinações lineares das K funções da base de funções considerada, o que restringe a sua forma.
Para evitar este problema, a suavização penalizada utiliza um grande número de elementos da base de funções, podendo-se usar mais elementos da base do que o número de pontos, t_j, para os quais foram obtidas as observações, e depois suavizam-se as curvas com base em critérios que são adequados para o problema em causa. Os coeficientes em (2.5) podem ser estimados pelo bem conhecido método dos mínimos quadrado da regressão linear, adaptado para dados funcionais. Neste caso, a soma de quadrados a minimizar em relação aos coeficientes (de regressão), c_k, é dada por:

$$SQM = \sum_{j=1}^{n} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2. \tag{3.1}$$

Ou, matricialmente, pode escrever-se da forma:

$$SQM = (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c}) = \| \mathbf{y} - \Phi\mathbf{c} \|, \tag{3.2}$$

onde Φ é a matriz de ordem $n \times K$, relativa aos elementos $\phi_k(t_j)$ da base de funções, e $\mathbf{y}$ e $\mathbf{c}$, representam os vetores de dados observados e coeficientes a estimar, respetivamente. O número de parcelas K pode ser igual ou maior do que o número de pontos t_j.

3.2 Suavização pelo método dos mínimos quadrados penalizados

Se ao método dos mínimos quadrados for adicionado um fator de penalização, por exemplo, na curvatura da curva a estimar, podemos encontrar uma estimativa mais suave.

Em geral, tomando-se o operador diferencial linear L, ou combinação linear de m derivadas de x,

$$L(x)(t) = a_0(t)x(t) + a_1(t)Dx(t) + ... + a_m(t)D^m x(t). \tag{3.3}$$

Por exemplo, L pode ser a segunda derivada, isto é, $L(x)(t) = D^2x(t)$. Podemos estimar a curva x, equivalentemente, estimar os coeficientes c_k, minimizando, em relação a esses coeficientes, a soma dos quadrados penalizado, que é definida por:

$$SQMP_\lambda(c_1, ..., c_K) = \sum_{j=1}^{n}(y_j - x_K(t_j))^2 + \lambda \int_a^b [L(x_K)(t)]^2 dt, \tag{3.4}$$

Ou, fazendo-se $x_K(t) = \mathbf{c}\phi'(t) = \phi(t)\mathbf{c}'$, a expressão anterior fica da forma:

$$SQMP_\lambda(c_1, ..., c_K) = \sum_{j=1}^{n}[y_j - \phi(t_j)\mathbf{c}']^2 + \lambda \mathbf{c} \left[\int_a^b L\phi'(t)L\phi(t)dt\right] \mathbf{c}', \tag{3.5}$$

sendo $\lambda \geq 0$ o parâmetro de penalização. Se $\lambda = 0$, o método dos mínimos quadrados penalizado reduz-se ao método dos mínimos quadrados (não penalizado). Isto é, a expressão (3.4) reduz-se a (3.1).

É usual a penalização ser aplicada à curvatura da função a estimar, isto é, penaliza-se o quadrado da segunda derivada de x em relação a t, ou, mais concretamente, penaliza-se a função:

$$PEN_2(x) = \int_I \left[D^2x(t)\right]^2 dt, \tag{3.6}$$

onde $I = [a, b]$, é o intervalo relativo aos dados observados.

Se houver como objetivo analisar o comportamento até à segunda derivada da função a estimar, também comum na FDA, a penalização será efetuada na quarta derivada, isto é, será penalizada a função:

$$PEN_4(x) = \int_I \left[D^4x(t)\right]^2 dt. \tag{3.7}$$

Em geral, quando se pretende analisar m derivadas da função x a estimar, deve-se penalizar a $m + 2$ derivada, isto é, penaliza-se a função:

$$PEN_{m+2}(x) = \int_I \left[D^{m+2}x(t)\right]^2 dt. \tag{3.8}$$

Consequentemente, segundo a especificação de se analisar m derivadas da função x a estimar, a expressão (3.4) toma a forma:

$$SQMP_\lambda(c_1, ..., c_K) = \sum_{j=1}^{n}[y_j - \phi(t_j)\mathbf{c}']^2 + \lambda \int_I \left[D^{m+2}x(t)\right]^2 dt. \tag{3.9}$$

Uma das principais dificuldades consiste em encontrar-se o parâmetro de penalização λ, que equilibre entre a subestimação e o excesso de suavização. A maioria das vezes a validação cruzada generalizada é o método mais aplicado na determinação do parâmetro λ. Este assunto está abordado na secção seguinte.

3.3 A escolha do parâmetro λ

Existem dois tipos de método na determinação do parâmetro de penalização λ. Um primeiro tipo consiste, simplesmente, em escolher-se o parâmetro de forma livre, isto é, testam-se individualmente vários valores para o parâmetro até se encontrar um que dê um bom ajuste. Obviamente, é um método muito subjetivo. Porém, é muito utilizado, em particular, quando se pretende ajustar uma única curva com grande dispersão dos dados. O segundo tipo é mais automatizado e objetivo, e consiste em utilizar-se algum método de reamostragem. Frequentemente, o método de reamostragem mais recomendado na FDA é a validação cruzada generalizada (*general cross-validation*, GCV).
De acordo com Ramsay e Silverman [11], a GCV é uma forma modificada de validação cruzada (*cross-validation*, CV) com o objetivo de se determinar o parâmetro de penalização, tendo sido desenvolvida em 1978 por Craven e Wahba [4] . Verifica-se que este método é computacionalmente mais eficiente do que a vulgar validação cruzada.
O critério define-se por:

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SQM}{n - df(\lambda)}\right). \tag{3.10}$$

onde $df(\lambda)$ são os graus de liberdade do ajuste definido pelo parâmetro λ e que satisfaz a relação:

$$df(\lambda) = tr\mathbf{H}(\boldsymbol{\lambda}), \tag{3.11}$$

sendo $\mathbf{H}(\lambda)$ a matriz chapéu associada ao método dos mínimos quadrados, que é uma matriz quadrada simétrica de ordem n, que satisfaz a igualdade

$$\hat{\mathbf{x}} = \mathbf{H}(\lambda)\mathbf{y}, \tag{3.12}$$

onde $\hat{\mathbf{x}}$ e $\mathbf{y}$ são os vetores $(x(t_1), \ldots, x(t_n))$ e os correspondentes dados observados $(y_1, \ldots, y_n)$, respetivamente.
De acordo com Dias e Souza [5], não se deve confiar cegamente num método automático para a escolha de λ. Podem surgir certos problemas como, por exemplo, quando há necessidade de ser analisada a derivada, o nível de suavização obtido automaticamente pode gerar derivadas complicadas, isto é, que se afastam das verdadeiras derivadas. Os autores sugerem que se comece por encontrar o valor que minimize $GCV(\lambda)$, e depois tentar experimentar outros valores próximos para visualizar qual o comportamento da suavização.

Como exemplo de aplicação do método dos mínimos quadrados penalizados, onde o parâmetro λ foi escolhido por GCV (não se tomou o valor mínimo, mas um relativamente próximo; ver código R no apêndice D) e a penalização foi efetuada na derivada de ordem 4, considerámos as curvas relativas ao momento de força articular do tornozelo direito de dois grupos de mulheres: 18 mulheres diagnosticadas com artrite reumatoide (grupo artrítico); 18 mulheres consideradas saudáveis (grupo controlo).

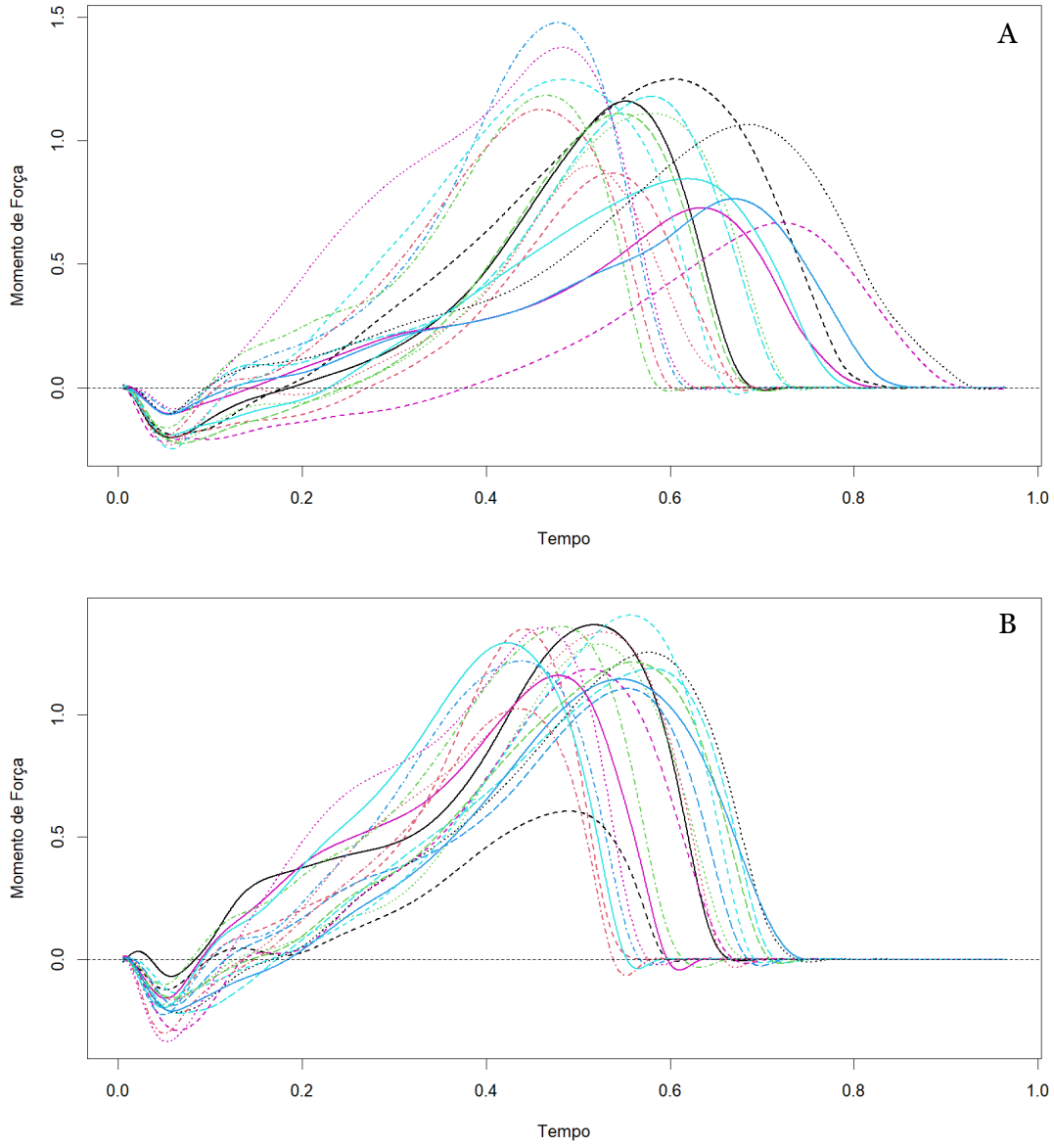


Figura 3.1: Curvas do momento de força dos grupos artrítico (A) e controlo (B).

As curvas foram obtidas com 197 funções B-spline de ordem 6 e foi considerado um parâmetro de penalização muito pequeno ($\lambda = 10^{-15}$) sobre a derivada de ordem 4 (ver código R no apêndice E).

Capítulo 4

Estatística de Dados Funcionais

Neste capítulo, assumimos uma amostra aleatória constituída por n curvas, $X_1, X_2, ..., X_n$, que foram convertidas em objetos funcionais. Cada curva pode ser vista como uma realização de uma função aleatória X em L^2 com a mesma distribuição de cada X_i. Isto é, suporemos o seguinte:

Suposição 4.0.1. *As funções $X_1, X_2, ..., X_n$ são independentes e identicamente distribuídas em L^2, e têm a mesma distribuição que X, que é assumida como quadrado integrável em algum intervalo $[T_1, T_2]$.*

4.1 Média e variância amostral

A estatística para os dados funcionais é obtida de modo similar ao caso de dados univariados. A média para uma amostra de n dados funcionais é definida por:

$$\hat{\mu}(t) = \bar{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t). \tag{4.1}$$

Observe-se a figura 4.1 (ver código R no apêndice E), onde estão representadas as médias do momento de força articular do tornozelo direito dos dois grupos: artrítico e controlo.

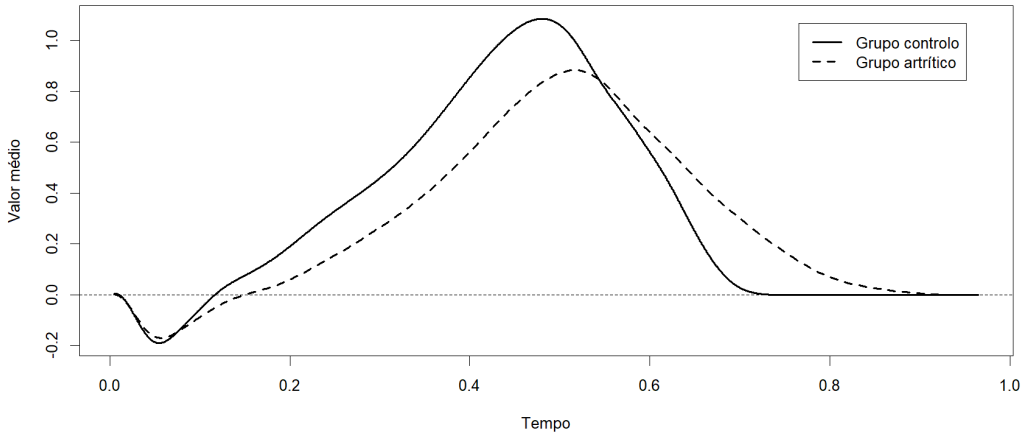


Figura 4.1: Médias funcionais do momento de força de cada grupo.

É notório um momento de força médio progressivamente maior no grupo controlo até próximo de 0,55 segundos. Depois desse instante, a média do momento de força do grupo artrítico foi maior, porque os indivíduos do grupo controlo terminaram, em média, o movimento mais rapidamente.

A função variância amostral corrigida é dada por

$$\hat{var}(t) = \frac{1}{n-1}\sum_{i=1}^{n}[X_i(t) - \hat{\mu}(t)]^2, \tag{4.2}$$

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

4.2 Covariância e correlação

A função covariância amostral é calculada para todos os pontos t e s através da seguinte fórmula:

$$\hat{c}(t,s) = \frac{1}{n-1}\sum_{i=1}^{n}\{X_i(t) - \hat{\mu}(t)\}\{X_i(s) - \hat{\mu}(s)\}. \tag{4.3}$$

Quanto à função correlação amostral, esta é dada por

$$\hat{corr}(t,s) = \frac{\hat{c}(t,s)}{\sqrt{var(t)var(s)}}. \tag{4.4}$$

Novamente, como exemplo, estão representas nas figuras 4.2 e 4.3 as superfícies das variâncias-covariâncias para os grupos artrítico e controlo, respetivamente (ver código R no apêndice E). Observa-se uma maior variação do momento de força do grupo artrítico na primeira e segunda parte do movimento. No grupo controlo a variação é maior na segunda parte do movimento, devido ao facto de os indivíduos deste grupo terminarem o movimento em tempos diferentes. Isto também acontece, e é mais notório, no grupo artrítico.

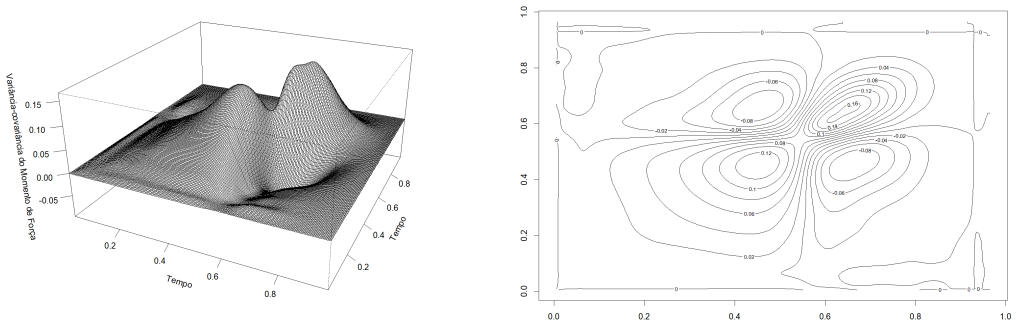


Figura 4.2: Superfície das variâncias-covariâncias do momento de força entre os indivíduos do grupo artrítico (lado esquerdo) e respetivas curvas de nível (lado direito).

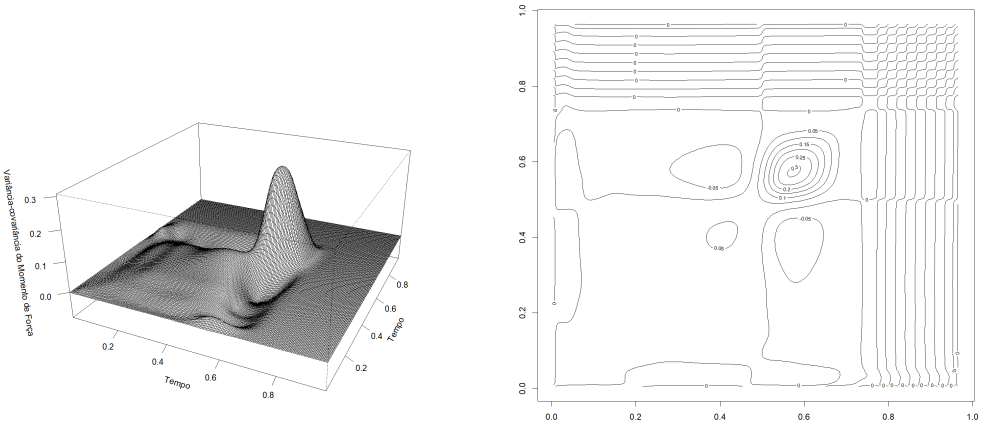


Figura 4.3: Superfície das variâncias-covariâncias do momento de força entre os indivíduos do grupo controlo (lado esquerdo) e respetivas curvas de nível (lado direito).

4.3 Consistência da função média e covariância amostral

A seguir, vamos compreender as propriedades que garantem a consistência das funções média e covariância amostrais. Em procedimentos inferenciais, frequentemente tratamos as curvas x_i como realizações de uma função aleatória X, para a qual definimos as funções média, $\mu(t)$, e covariância, $c(t,s)$, por

$$\mu(t) = E[X(t)] \tag{4.5}$$

$$c(t,s) = E[(X(t) - \mu(t))(X(s) - \mu(s))]. \tag{4.6}$$

As funções amostrais $\hat{\mu}$ e $\hat{c}$ são vistas como estimadores dos parâmetros populacionais μ e c, respetivamente. Defina-se também a covariância tensorial e o operador covariância, respetivamente, da seguinte maneira:

$$C = E[(X - \mu) \otimes (X - \mu)] \tag{4.7}$$

$$C(\cdot) = E[\langle (X - \mu), \cdot \rangle (X - \mu)] \tag{4.8}$$

Lemma 4.3.1. *Se $X_1, X_2 \in L^2$ são independentes e $EX_1 = 0$, então $E[\langle X_1, X_2 \rangle] = 0$.*

Teorema 4.3.2. *Se a suposição 4.0.1 for válida, então $E\hat{\mu} = \mu$ e $E\|\hat{\mu} - \mu\| = O(n^{-1})$.*

Este teorema afirma que $\hat{\mu}$ é um estimador não enviesado e L^2-consistente de μ. Em particular, a L^2-consistência de μ implica que é consistente em probabilidade: $\|\hat{\mu} - \mu\| \xrightarrow{P} 0$.

Demonstração. Para cada n, em quase todos os pontos $t \in [T_1, T_2]$, $EX_n(t) = \mu(t)$, então

resulta de imediato $E\hat{\mu} = \mu$ em L^2. Pelo lema 4.3.1, tem-se, sucessivamente,

$$\begin{aligned} E\|\hat{\mu} - \mu\| &= n^{-2} \sum_{i,j=1}^{n} E[\langle (X_i - \mu), (X_j - \mu) \rangle] \\ &= n^{-2} \sum_{i=1}^{n} E\|X_i - \mu\|^2 \\ &= n^{-1} \sum_{i=1}^{n} E\|X - \mu\|^2 \end{aligned}$$

□

Os teoremas 4.3.3 e 4.3.4, a seguir, são importantes para a demonstração do teorema 4.3.5. As demonstrações destes dois teoremas podem ser encontradas em [10]. O teorema 4.3.5 estabelece a consistência do operador covariância amostral.

Teorema 4.3.3. *Se $E\|X\|^4 < \infty$, $(EX = 0)$ e a suposição 4.0.1 é válida, então,*

$$E\|\widehat{C}\|_S^2 \leq E\|X\|^4.$$

Observe-se que $\|\cdot\|_S$ representa a norma de Hilbert-Schmidt, que é definida por: $\|A\|_S^2 = \sum_{i \in I} \|Ae_i\|^2$, onde A é um operador limitado, tal que, $A : H \to H$, com H um espaço de Hilbert, e $\{e_i : i \in I\}$ é uma base ortonormada de H.

Teorema 4.3.4. *Sejam $\boldsymbol{x}, \boldsymbol{y} \in H$. Então*

$$\|\boldsymbol{x} \otimes \boldsymbol{y}\|_{H \otimes H} = \|\langle \boldsymbol{y}, \cdot \rangle \boldsymbol{x}\|_S.$$

Teorema 4.3.5. *Se $E\|X\|^4 < \infty$, $EX = 0$ e a suposição 4.0.1 é válida, então,*

$$E\|\widehat{C} - C\|_S^2 \leq N^{-1} E\|X\|^4.$$

Demonstração. Pelo teorema 4.3.4,

$$E\|\widehat{C} - C\|_S^2 = E\|\widehat{C} - C\|_{H \otimes H}^2 = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} E\langle X_i \otimes X_i - C, X_j \otimes X_j - C \rangle_{H \otimes H}$$

Desde que as amostras sejam independentes e identicamente distribuídas, todos as parcelas cruzadas no somatório são nulas. Deste modo, o segundo membro da igualdade anterior reduz-se, sucessivamente, a

$$\begin{aligned} & n^{-2} \sum_{i=1}^{n} E\langle X_i \otimes X_i - C, X_i \otimes X_i - C \rangle_{H \otimes H} = \\ &= n^{-1} E\langle X \otimes X - C, X \otimes X - C \rangle_{H \otimes H} \\ &= n^{-1} (E\|X\|^4 - \|C\|_S^2) \\ &\leq E\|X\|^4 \end{aligned}$$

□

4.4 Intervalo de confiança para uma média

Para um estudo inferencial, é útil encontrar as distribuições, muitas vezes assintóticas, dos estimadores. Para o espaço L^2 temos, para quase todos t, que $\hat{\mu}(t)$ é assintoticamente normal com média $\mu(t)$ e variância $n^{-1}c(t,t)$. Temos a seguinte variante do Teorema do Limite Central, cuja demonstração pode ser encontrada em [10].

Teorema 4.4.1. *Seja $X_1, ..., X_n$ uma sequência de elementos num espaço de Hilbert, H, independentes e identicamente distribuídos com $E\|X_i\|^2 < \infty$, para $i = 1,2,\ldots,n$. Então,*

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} n(0, C),$$

em H, onde $C = E[(X_1 - \mu) \otimes (X_1 - \mu)]$. Se, além disso, $E\|X\|^4 < \infty$, então

$$\sqrt{n}(\widehat{C} - C) \xrightarrow{d} n(0, \Gamma),$$

em S , onde $\Gamma = E\left[[(X_1 - \mu) \otimes (X_1 - \mu) - C] \otimes [(X_1 - \mu) \otimes (X_1 - \mu) - C]\right] \in S \otimes S$.

Da propriedade anterior resulta um intervalo de confiança para $\mu(t)$:

$$\left[\hat{\mu}(t) - z_{1-\alpha/2} n^{-1/2} \sqrt{\hat{c}(t,t)}, \hat{\mu}(t) + z_{1-\alpha/2} n^{-1/2} \sqrt{\hat{c}(t,t)}\right] \tag{4.9}$$

onde $z_{1-\alpha/2}$ é o quantil $1 - \alpha/2$ da distribuição normal padrão. De outro modo, para cada t, tem-se que

$$P\left(\mid \hat{\mu}(t) - \mu(t) \mid \leq z_{1-\alpha/2} n^{-1/2} \sqrt{\hat{c}(t,t)}\right) \approx 1 - \alpha. \tag{4.10}$$

Existem outras abordagens para se obter um intervalo de confiança para uma média funcional, em particular, que não dependam de uma distribuição de probabilidade. Em [10] é sugerido que, por exemplo, um intervalo de confiança para uma média funcional seja obtido usando-se o método de reamostragem *bootstrap*.

4.5 Teste t

Uma forma simples e prática para se obter um teste de hipóteses para a comparação de duas médias funcionais a partir de amostras independentes, consiste em aproveitar-se o bem conhecido teste de permutação (ou de aleatorização ou de reamostragem) univariado, que designaremos simplesmente por teste t. Este teste é uma abordagem não paramétrica, isto é, não requer qualquer condição acerca da distribuição nas populações, e é especialmente útil quando os tamanhos amostrais não são grandes.

Para este teste t considere-se a estatística:

$$T(t) = \frac{|\hat{\mu}_1(t) - \hat{\mu}_2(t)|}{\sqrt{\frac{1}{n_1}\hat{var}[X_1(t)] + \frac{1}{n_2}\hat{var}[X_2(t)]}}, \tag{4.11}$$

onde n_1 e n_2 são os tamanhos amostrais. Esta estatística fornece uma noção da separação relativa dos dois grupos de funções. Para um teste de hipótese formal, precisamos de uma estatística para testar e um valor de probabilidade indicando o resultado do teste. Observe-se que a estatística de teste que aqui utilizaremos é o valor máximo do teste t multivariado, isto é, é o máximo de $T(t)$. Para encontrarmos um valor crítico da estatística $T(t)$, usaremos, como referido anteriormente, um número suficientemente grande de permutações aleatórias das curvas de duas amostras e recalcula-se para cada permutação o novo valor máximo de $T(t)$. Consegue-se assim obter a distribuição exata sob a hipótese nula, que consiste na igualdade das duas médias funcionais: $\mu_1(t) = \mu_2(t)$.

Como exemplo de aplicação, recorreremos novamente aos dados funcionais do momento de força articular do tornozelo direito dos dois grupos de mulheres: grupo artrítico e grupo controlo.

Na figura 4.4 encontra-se representado o resultado do teste t para a igualdade das médias do momento de força das duas populações em estudo. Foram consideradas 200 permutações aleatórias entre os dois grupos (ver código R no apêndice E).

A figura 4.1 sugere um momento de força médio menor no grupo artrítico até ao instante $0{,}55$ segundos. O teste t encontrou, para o nível de significância a 5%, diferenças estatisticamente significativas entre as médias pontuais do momento de força no intervalo $[0{,}14; 0{,}50]$ segundos, aproximadamente, sendo menores as médias do momento de força no grupo artrítico. Nos intervalos $[0{,}23; 0{,}27]$ e $[0{,}36; 0{,}45]$ segundos, aproximadamente, as diferenças absolutas máximas entre as médias foram estatisticamente significativas (valor de prova igual a $0{,}04$), para o nível de significância de 5%, onde o quantil da diferença absoluta máxima, dada pela estatística $T(t)$, foi, aproximadamente, igual a $2{,}995$.

No intervalo $[0{,}66; 0{,}80]$ segundos, aproximadamente, foram encontradas também, para o nível de significância a 5%, diferenças estatisticamente significativas entre as médias do momento de força articular dos dois grupos. Mas, neste caso, as médias observadas foram maiores no grupo artrítico, o que é natural que tal tenha acontecido, devido ao facto de os indivíduos do grupo controlo terminarem, em média, o movimento mais rapidamente.

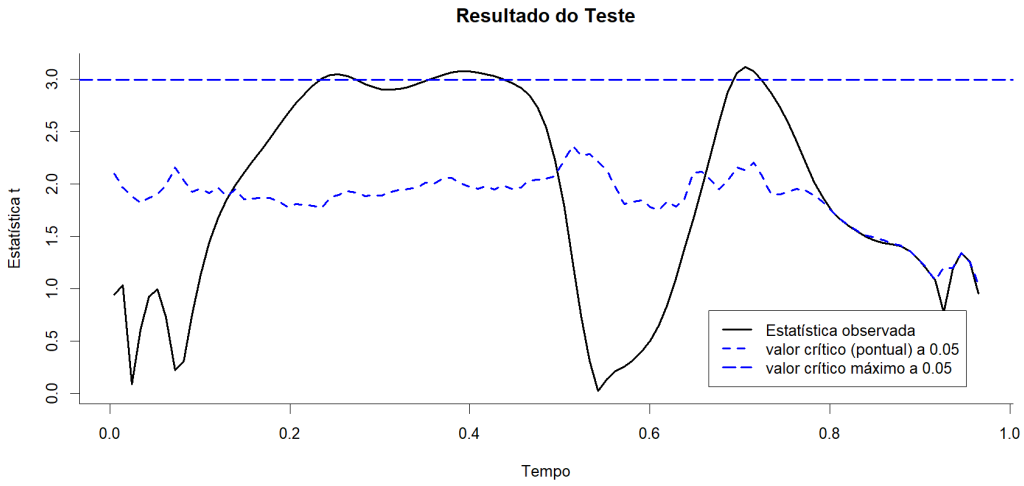


Figura 4.4: Resultado do teste de permutações para as médias multivariadas do momento de força dos dois grupos. Foram consideradas 200 permutações aleatórias entre os dois grupos.

Capítulo 5

Alinhamento

Na análise de dados funcionais, a variabilidade num conjunto de curvas pode provir principalmente de duas fontes. A primeira é a variação de amplitude, que se refere às diferenças na forma ou magnitude das curvas. A segunda é a variação de fase, que se refere às diferenças no tempo ou alinhamento de características nas curvas. As curvas podem apresentar a mesma forma, mas poderão estar desfasadas. Isto é, máximos, mínimos ou até pontos de inflexão das curvas ocorrerem em tempos distintos. Segundo Bates et al., citado por Page et al. [8], essa variabilidade pode contribuir para o efeito de cancelamento estatístico associado à agregação dos dados. As principais características de cada padrão individual podem então ser ocultadas quando padrões médios são obtidos. A variação da fase faz com que esses padrões não possam ser comparados num dado momento porque não apresentam a mesma forma. Por isso, para poderem ser comparados, a escala do tempo tem de ser distorcer ou transformada através do *alinhamento*[1] das curvas.

Na maioria dos casos, os dados funcionais costumam apresentar diferenças de fase e/ou de amplitude, como no exemplo, já clássico, na FDA de curvas relativas do crescimento de crianças [13, Secção 8.1]. As crianças tendem a crescer em ritmos diferentes. Se compararmos essas curvas de crescimento sem alinhamento, a média pode não representar bem os padrões reais. Algo similar acontece com as curvas que provêm dos dados do momento de força, que usámos no presente trabalho. Por observação das figuras 3.1-A e 3.1-B constata-se que todas as curvas do momento de força tendem a ter padrões similares; têm um mínimo no início do movimento e um máximo próximo do final do movimento. No entanto, a fase do movimento nos indivíduos artríticos (figura A) é notoriamente diferente da fase do movimento dos controlos (figura B), onde se observa, tendencialmente, uma maior duração desse movimento, além de uma maior variação da duração.

O alinhamento de curvas é, genericamente, um método que consiste em sincronizar as curvas de forma a reduzir a variabilidade de fase, mas mantendo a forma e a amplitude das curvas. Em particular, o alinhamento fornece uma padronização local como função do tempo. Essa padronização transforma um conjunto de curvas num novo conjunto de curvas que variam apenas em termos da amplitude. Portanto, os eventos principais, tais como máximos, mínimos ou pontos de inflexão, ocorrem ao mesmo tempo para todas as curvas transformadas e as características individuais de interesse permanecem idênticas.

Existem três formas para o alinhamento das curvas: padronização linear do tempo (deslocamento do tempo), alinhamento por pontos de referência (utiliza, por exemplo, máximos ou mínimos locais das curvas) e alinhamento contínuo (utiliza toda a informação contida nas curvas). Observe-se, no entanto, que após a padronização linear as diferenças

[1] Na literatura em inglês é usual utilizar-se a palavra *registration*. Escolhemos utilizar a palavra *alinhamento* porque, pareceu-nos, dá uma melhor ideia do que está em causa.

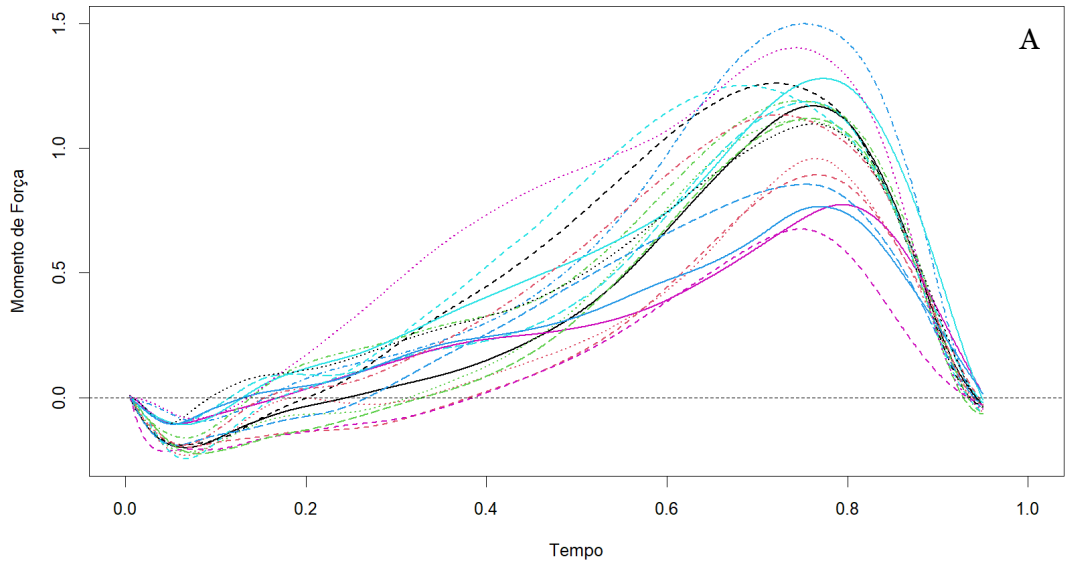
de tempo locais podem permanecer, isto é, a variação de fase pode não alterar. Consequentemente, poderá ser necessário adicionalmente um procedimento não linear como o alinhamento por pontos de referência ou contínuo.

5.1 Padronização Linear do Tempo

O caso mais simples de alinhamento consiste na padronização linear do tempo, que não é mais do que o deslocamento das curvas ao longo do eixo temporal. Ou seja, se os dados das curvas x_i foram obtidos em diferentes intervalos de tempo, por exemplo, $[0, T_i]$, é desejável deslocar-se o tempo com o auxílio de uma constante não negativa, δ_i, de modo a que o tempo máximo, T, seja comum às novas curvas alinhadas. Isto é, cada constante não negativa δ_i deverá ser escolhida de modo que, para cada curva x_i, $[0, T_i] \subseteq [0, T]$ e as novas curvas alinhadas, x_i^*, são dadas por:

$$x_i^*(t) = x_i(t + \delta_i). \tag{5.1}$$

Como exemplo de aplicação do alinhamento pelo deslocamento do tempo, os gráficos da figura 5.1 correspondem às curvas do momento de força articular do tornozelo direito das mulheres artríticas e saudáveis, onde o tempo foi deslocado para o valor máximo observado entre todos os indivíduos da amostra, que foi de 0,950 segundos (ver código R no apêndice F). Por comparação com os gráficos A e B da figura 3.1, é visível que as curvas mantém a forma (inclusive, a escala do momento de força é a mesma), mas agora todas as curvas terminam no mesmo instante de tempo. Observe-se, no entanto, que, em qualquer dos gráficos das novas curvas alinhadas pelo tempo, estas ainda não estão na mesma fase.



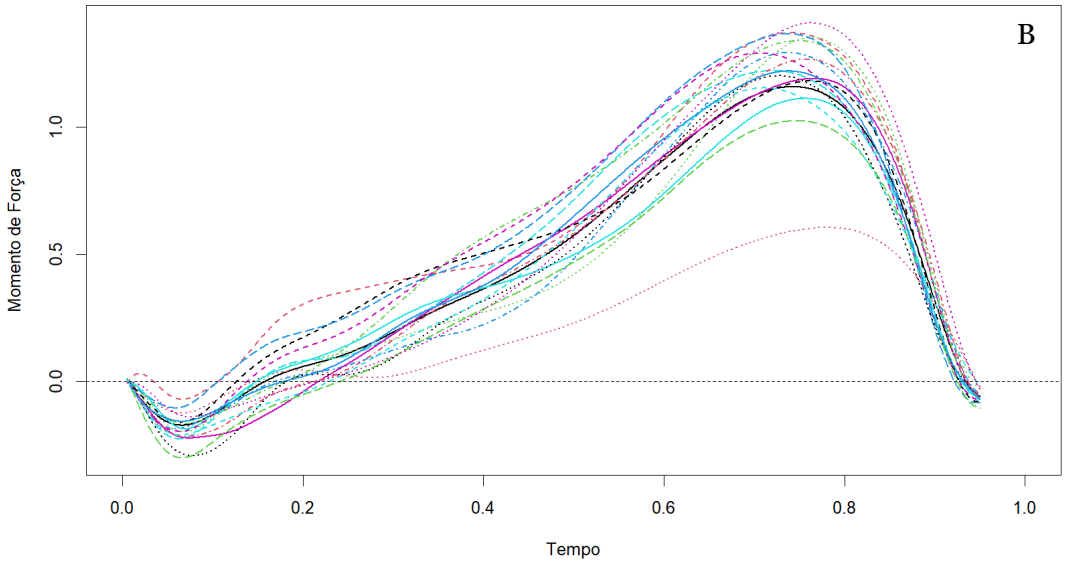


Figura 5.1: Curvas do momento de força dos grupos artrítico (A) e controlo (B) no intervalo $[0; 0{,}950]$ segundos.

5.2 Alinhamento por Pontos de Referência Utilizando Função de Distorção Temporal

Em muitas aplicações reais, como no momento de força articular do tornozelo, podem ocorrer variações no instante em que determinados eventos se manifestam, ainda que todos os dados tenham sido registados no mesmo intervalo temporal. Se essas variações forem desconsideradas, medidas estatísticas como a média e a variância poderão deixar de refletir adequadamente as características dos dados funcionais.

A distorção temporal é uma técnica usada na FDA para corrigir variações de fase quando as curvas têm formas semelhantes, mas ocorrem em momentos diferentes. Esta técnica permitirá alinhar as curvas de modo que os eventos importantes, como máximos e mínimos, coincidam no tempo. A ideia consiste em admitir que existe, para cada curva $x_i(t)$, uma função, $h_i(t)$, que distorça o tempo de modo que as mudanças de fase fiquem alinhadas, isto é, sincronizadas.

Assim, dado um conjunto de n curvas, $x_i(t)$, $i = 1, 2, \ldots, n$, o alinhamento destas curvas corresponde a um novo conjunto de curvas, $x_i^*(t)$, $i = 1, 2, \ldots, n$, tais que,

$$x_i(t) = x_i^*(h_i(t)),$$

de modo a reduzir-se as variações de fase entre as curvas, onde $h_i(t)$ é a função de distorção temporal a estimar para a curva $x_i(t)$. As novas curvas, que depois pretenderemos analisar, são dadas por:

$$x_i^*(t) = x_i(h_i^{-1}(t)),$$

onde h_i^{-1} é a função inversa de h_i, que satisfaz a propriedade: $h^{-1}[h(t)] = t$.
Segundo Ramsay e Silverman [12], a função de distorção temporal deve ser estritamente crescente, pois não é admissível que o tempo volte para trás. A função de distorção temporal também deve ser suave de modo a permitir o cálculo de derivadas até, pelo menos, à ordem de interesse na aplicação das curvas que se desejam alinhar. Quando as curvas forem observadas num intervalo de tempo comum, $[0, T]$, ou previamente alinhadas no mesmo intervalo de tempo, como descrito na secção 5.1, a função de distorção temporal tem ainda de satisfazer as condições: $h(0) = 0$ e $h(T) = T$. Para além destas duas condições, será necessário primeiro escolherem-se, manualmente, os pontos de referência, como máximos, mínimos ou outras características relevantes das curvas. São todos esses pontos que irão permitir estimar por um método de interpolação cada função $h_i(t)$, para uma dada curva $x_i(t)$.
Para ilustrarmos a ideia subjacente à estimação de uma função de distorção temporal, suponhamos que as curvas a alinhar somente tinham uma única característica de interesse. Por exemplo, se, para cada $i = 1, \ldots, n$, t_i é o instante de tempo que torna a curva $x_i(t_i)$ máxima, então interessa alinhar as curvas tendo em conta esses máximos de modo que $h_i(t_i) = t_m$, para todo $i = 1, 2, \ldots, n$, onde $t_m = \frac{1}{n}\sum_{i=1}^{n} t_i$. Assim, para cada curva x_i, temos os pontos $(0, 0)$, (t_i, t_m) e (T, T) e existe uma única parábola $h_i(t)$ que passa por esses três pontos. Consequentemente, a curva alinhada, x_i^*, é definida por $x_i^*(t) = x_i(h_i^{-1}(t))$, para $t \in [0, T]$.

5.3 Alinhamento Contínuo

Enquanto o alinhamento por pontos de referência baseia-se em informações locais das curvas, o alinhamento contínuo utiliza toda a informação contida nas curvas para estimar a função de distorção temporal, sendo assim um procedimento mais refinado. Segundo Ramsay e Silverman [12], o alinhamento por pontos de referência é mais tedioso, em particular numa amostra grande, e pode não alinhar corretamente alguma característica das curvas. Embora o alinhamento contínuo seja computacionalmente bem mais intensivo que o alinhamento por pontos de referência, não necessita qualquer informação inicial acerca das características das curvas a alinhar. O alinhamento contínuo consiste num algoritmo complexo que começa por encontrar as características relevantes das curvas a alinhar. Utiliza, inclusive, as derivadas das curvas a alinhar para detetar, por exemplo, máximos, mínimos ou pontos de inflexão. De seguida estima a função de distorção, em particular a sua inversa, para cada curva por um método de interpolação, como descrito na secção 5.2.
Como exemplo, aplicámos o alinhamento contínuo às curvas do momento de força do tornozelo direito de cada grupo de mulheres, que foram previamente alinhadas para o intervalo $[0; 0{,}950]$ segundos (ver gráficos A e B da figura 5.1). O resultado do alinhamento contínuo encontra-se representado nos gráficos A e B da figura 5.2 (o código R encontra-se disponível no apêndice F). É visível uma diminuição da variação de fase após o alinhamento contínuo, sem que tenha sido alterada a amplitude de cada curva.

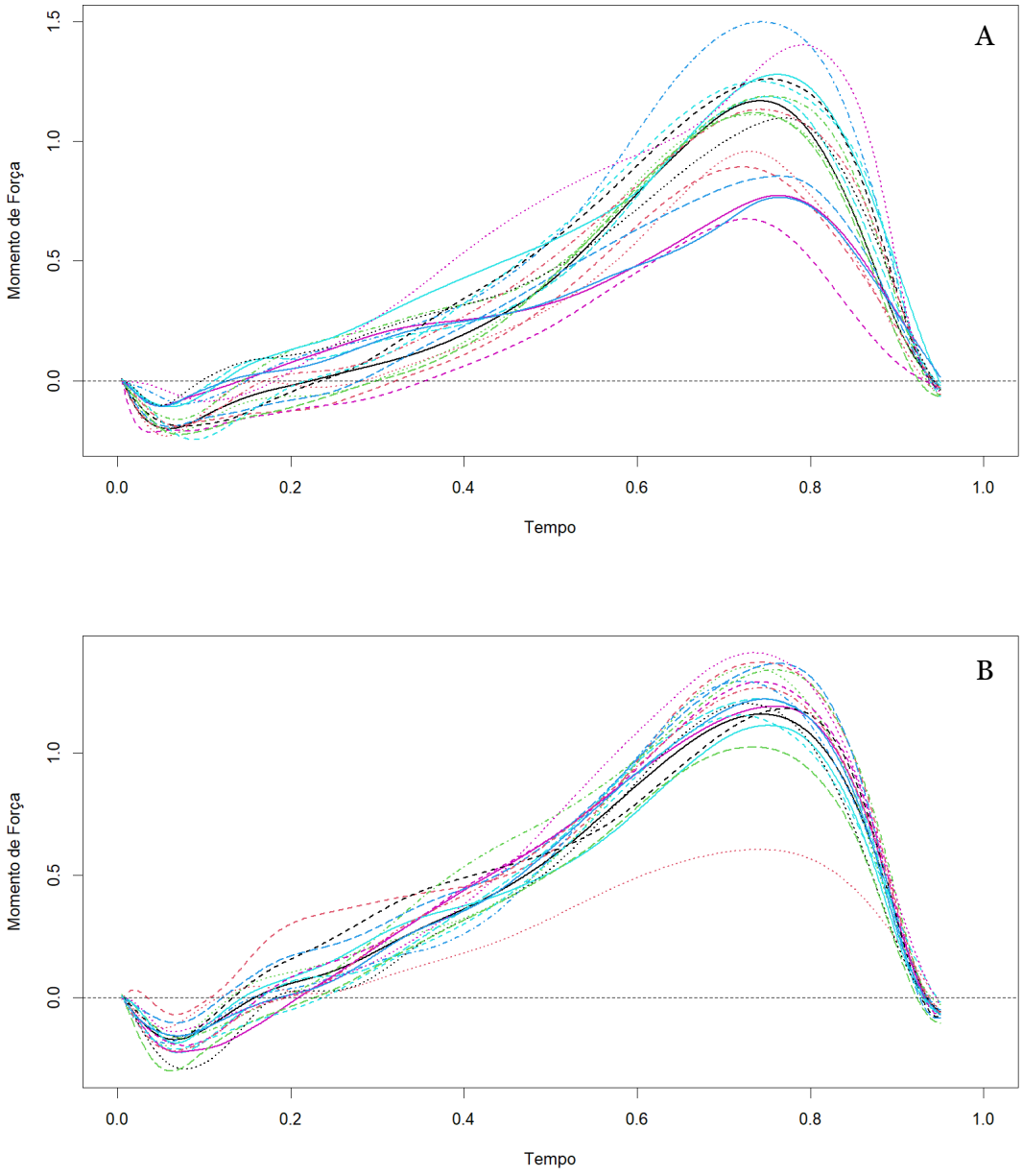


Figura 5.2: Curvas do momento de força dos grupos artrítico (A) e controlo (B) no intervalo $[0; 0{,}950]$ segundos após o alinhamento contínuo.

Se repetirmos o teste t da secção 4.5, mas agora com os dados do momento de força com o alinhamento contínuo, encontra-se um valor de prova igual a $0{,}015$ para a diferença absoluta máxima entre as médias dos momentos de força dos dois grupos, a favor do grupo controlo, sendo o quantil da diferença absoluta máxima, dada pela estatística $T(t)$, igual a 2.876. O intervalo $[0{,}29; 0{,}60]$ segundos foi onde se obtiveram as diferenças absolutas máximas estatisticamente significativas, para o nível de significância de 5%, entre as médias do momento de força dos dois grupos. As diferenças pontuais estatisticamente significativas, para o nível de significância de 5%, entre as médias do momento de força situaram-se, aproximadamente, no intervalo $[0{,}22; 0{,}70]$ segundos, sendo, obviamente, as médias me-

nores no grupo artrítico. Este teste encontra-se representado na figura 5.3.

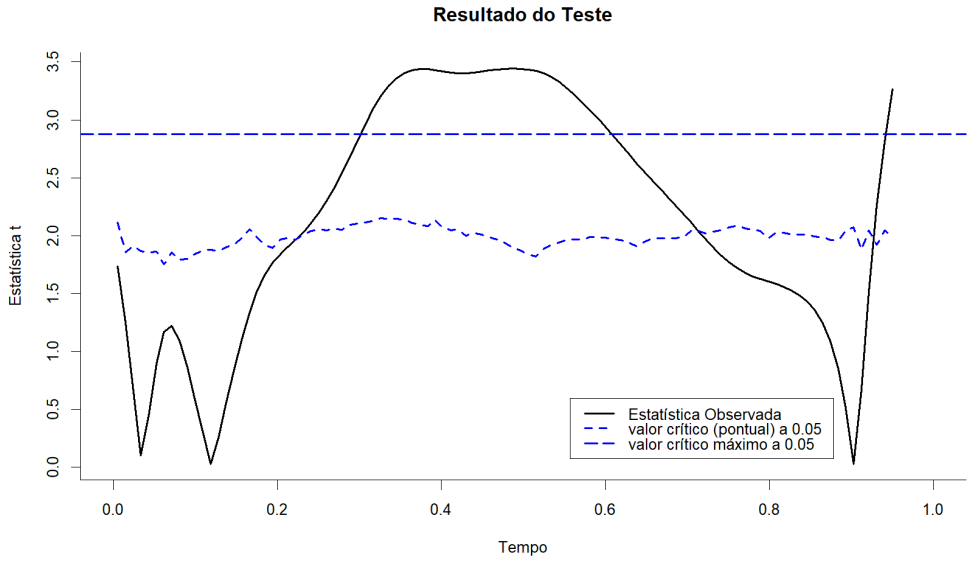


Figura 5.3: Resultado do teste de permutações para as médias multivariadas do momento de força dos dois grupos após o alinhamento contínuo. Foram consideradas 200 permutações aleatórias entre os dois grupos.

É visível na figura 5.3 que, após o alinhamento contínuo, a média do momento de força do grupo artrítico é menor que a respetiva média do grupo controlo durante quase todo o movimento do tornozelo.

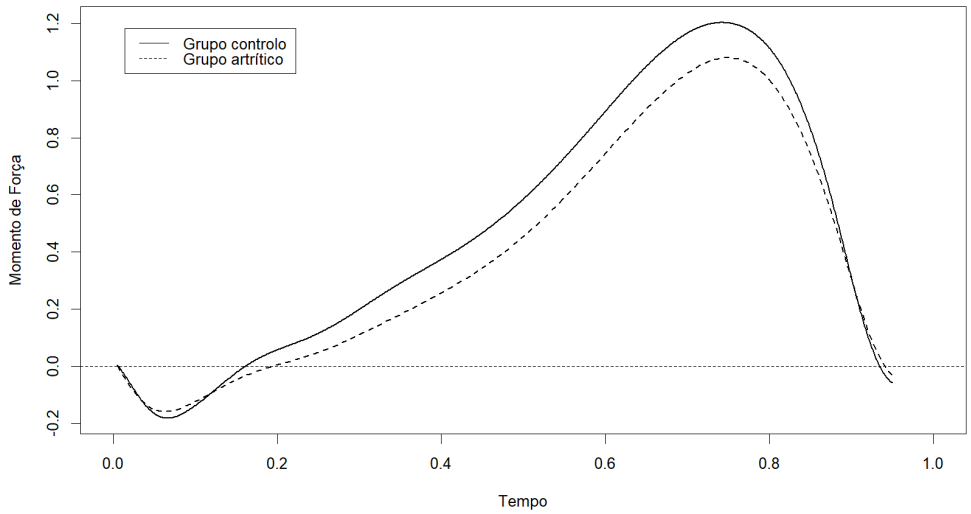


Figura 5.4: Médias funcionais do momento de força de cada grupo após o alinhamento contínuo dos dados.

5.4 Decomposição da Variância em Termos da Amplitude e Fase

Kneip e Ramsay (2008) [9] desenvolveram uma teoria que permite decompor o erro quadrático médio (*Mean Square Error, MSE*) total de n curvas x_i não alinhadas num intervalo comum I,

$$\text{MSE}_{total} = n^{-1} \sum_{i}^{n} \int_I [x_i(t) - \bar{x}(t)]^2 dt \tag{5.2}$$

em termos de duas componentes de variação: aquela que é devida à amplitude, que representaremos por MSE_{amp}, e aquela que é devida à fase, que representaremos por MSE_{fase}. Kneip e Ramsay mostraram que é valida a decomposição:

$$\text{MSE}_{total} = \text{MSE}_{amp} + \text{MSE}_{fase}, \tag{5.3}$$

se for considerada uma constante, C_R, definida por

$$C_R = 1 + \frac{n^{-1} \sum_i^n \int_I \left[Dh_i(t) - n^{-1} \sum_i^n Dh_i(t)\right] \left[x_i^{*2}(t) - n^{-1} \sum_i^n x_i^{*2}(t)\right] dt}{n^{-1} \sum_i^n \int_I x_i^{*2}(t) dt}, \tag{5.4}$$

onde

$$\text{MSE}_{amp} = C_R n^{-1} \sum_{i}^{n} \int_I [x_i^*(t) - \bar{x}^*(t)]^2 dt \tag{5.5}$$

$$\text{MSE}_{fase} = C_R \int_I \bar{x}^{*2}(t) dt - \int_I \bar{x}^2(t) dt, \tag{5.6}$$

e x_i^* e $\bar{x}^*$ representam as curvas alinhadas e a média das curvas alinhadas, respetivamente, pelas funções de distorção h_i.

Verifica-se que a estrutura da constante C_R implica que $C_R - 1$ está relacionada com a covariância entre a função Dh_i e o quadrado da curva alinhada x_i^{*2}. Se as funções $Dh_i(t)$ e as funções alinhadas ao quadrado, x_i^{*2}, são independentes, então a variação de fase capturada por $Dh_i(t)$ é independente da variação de amplitude capturada por x_i^{*2}.

Assim, pode afirmar-se, em termos gerais, que MSE_{fase} é a parte da variância removida pelo processo de alinhamento e será útil considerar-se o coeficiente de determinação seguinte:

$$R^2 = \frac{MSE_{fase}}{MSE_{total}}. \tag{5.7}$$

Este coeficiente dá-nos a proporção de variância devida à variação de fase. Observe-se que, ao contrário do que acontece com o coeficiente de determinação na regressão linear, este coeficiente pode ser negativo, já que, a segunda parcela em 5.6 pode exceder a primeira. Isto significa que houve alguma falha durante o processo de alinhamento ou houve

alinhamento excessivo.
Considerando novamente os dados funcionais do momento de força dos grupos artrítico e controlo, as percentagens de variância devidas à fase que resultaram do alinhamento contínuo foram, aproximadamente, 15,6% ($\text{MSE}_{fase} \approx 0{,}00358$; $\text{MSE}_{amp} \approx 0{,}01944$) e 23,7% ($\text{MSE}_{fase} \approx 0{,}00339$; $\text{MSE}_{amp} \approx 0{,}01090$), respetivamente. Destes valores deduz-se que houve, proporcionalmente, maior variação devida à amplitude no grupo artrítico. Tal é bem visível quando se comparam os respetivos gráficos das figuras 5.1 (sem alinhamento contínuo) e 5.2 (com alinhamento contínuo) dos grupos artrítico e controlo. O código R utilizado nestes cálculos encontra-se no apêndice F.

Capítulo 6

Modelos Lineares Funcionais

No bem conhecido modelo de regressão linear,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, i = 1, 2, \ldots, n, \tag{6.1}$$

todas as variáveis, dependente e independentes, e os coeficientes de regressão são escalares. Na FDA existe o análogo ao usual modelo de regressão linear, mas onde pelo menos uma das variáveis, dependente ou independentes, é uma função. Para esses modelos lineares funcionais, o análogo aos coeficientes de regressão, β_i, terão de ser definidos apropriadamente; tanto podem ser escalares como funções. Os modelos lineares funcionais dividem-se em três grandes categorias, dependendo se a variável dependente e/ou as variáveis independentes são funções.

Variável Dependente Escalar e Independente Funcional: Para este caso o modelo é dado por:

$$y_i = \beta_0 + \int_I \beta_1(t) x_i(t) dt + \varepsilon_i, i = 1, 2, \ldots, n, \tag{6.2}$$

onde y_i é a i-ésima observação da variável dependente escalar, $x_i(t)$ é a i-ésima observação funcional da variável independente funcional, β_0 é a constante escalar, $\beta_1(t)$ é o coeficiente de regressão funcional e ε_i é o resíduo escalar, que são estimativas do erro ε com média zero e variância σ_ε^2 constante. Observe-se que este modelo pode tornar-se múltiplo, se forem incluídas variáveis independentes escalares ou funcionais. Para duas ou mais variáveis independentes funcionais, passa-se a ter múltiplos coeficientes de regressão funcionais. Estes coeficientes têm uma interpretação análoga aos coeficientes de regressão dos modelos lineares não funcionais: variação da variável dependente por unidade aumentada da variável independente, mas agora ao longo do tempo.

Variável Dependente Funcional e Independentes Escalares: Neste caso pretende-se prever uma curva com variáveis independentes escalares. O modelo é dado por:

$$y_i(t) = \beta_0(t) + \sum_{k=1}^{p} x_{ik} \beta_k(t) + \varepsilon_i(t), \tag{6.3}$$

onde o $y_i(t)$ é a i-ésima observação funcional da variável dependente funcional, x_{ik} é a i-ésima observação da variável escalar x_k, $\beta_i(t)$ é o i-ésimo coeficiente de regressão funcional, $\beta_0(t)$ é também funcional e $\varepsilon_i(t)$ é o i-ésimo resíduo funcional.

Variáveis Dependente e Independente Funcionais: Neste caso pretende-se prever curvas com curvas. O modelo é dado por:

$$y_i(t) = \beta_0(t) + \int_I x_i(s)\beta(s,t)ds + \varepsilon_i(t), \tag{6.4}$$

onde $y_i(t)$ é a i-ésima observação funcional da variável independente funcional no tempo t, $x_i(s)$ é a i-ésima observação funcional da variável independente funcional no tempo s, $\beta(s,t)$ é uma função *kernel* bidimensional, $\beta_0(t)$ é também funcional e $\varepsilon_i(t)$ é o i-ésimo resíduo funcional.

No presente trabalho somente focaremos a nossa atenção no modelo em que a variável dependente é escalar e a independente é funcional, isto é, no modelo 6.2. Para este modelo, abordaremos como é estimado o coeficiente de regressão funcional, β_1, e o seu intervalo de confiança. Finalizaremos o estudo do modelo 6.2 adaptando-o ao caso especial em que a variável dependente é dicotómica e aplicaremos este modelo aos dados do momento de força articular dos tornozelos direito e esquerdo dos grupos de mulheres com e sem artrite. Os modelos 6.3 e 6.4 podem ser analisados detalhadamente em [10] ou [3].

6.1 Estimação de $\beta_1(t)$

A estimação do coeficiente de regressão β_1 do modelo 6.2 é similar àquela que se utiliza nos modelos não funcionais, isto é, consiste em utilizar-se o método dos mínimos quadrados:

$$\min_{\beta_0,\beta_1(t)} \sum_{i=1}^{n} \left\{ y_i - \beta_0 - \int_I \beta_1(t)x_i(t)dt \right\}^2 \tag{6.5}$$

No entanto, esta abordagem tem o desafio acrescido de que tanto a variável independente como o coeficiente de regressão são funcionais. Mas, esta abordagem é similar àquela que foi feita na secção 3.1.

Assim, de acordo com Crainiceanu et al. [3], toma-se a expansão do coeficiente β_1 usando-se uma base de funções $\phi_1(s),\ldots,\phi_K(s)$, isto é,

$$\beta_1(t) = \sum_{k=1}^{K} \beta_{1k}\phi_k(t) \tag{6.6}$$

Segundo Kokoszka e Reimherr [10], K deve ser um número relativamente pequeno. Substituindo esta expansão dentro do integral no modelo 6.2, obtemos, sucessivamente,

$$\begin{aligned} E[y_i] &= \beta_0 + \int_I \beta_1(t)x_i(t)dt \\ &= \sum_{k=1}^{K} \left[\int_I \phi_k(t)x_i(t)dt \right] \beta_{1k} \\ &= \mathbf{C}_i'\boldsymbol{\beta}, \end{aligned} \tag{6.7}$$

onde $C_{ik} = \int_I \phi_k(t)X_i(t)dt$, $\mathbf{C}_i = [C_{i1},...,C_{iK}]'$ e $\boldsymbol{\beta} = (\beta_{11},...,\beta_{1K})'$ é o vetor dos coeficientes da expansão 6.6. Consequentemente, a expressão 6.5 fica reduzida em termos matriciais a

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{W}\boldsymbol{\beta}||^2, \tag{6.8}$$

onde $\mathbf{y} = (y_1, ..., y_n)'$ e $\mathbf{W}$ é a matriz cujas linhas são os vetores $[1, C_{i1}, ..., C_{iK}]$.

Devido à dispersão ou, simplesmente, aos erros de observação dos dados, o método dos mínimos quadrados poderá não ser a melhor abordagem para a estimação do coeficiente de regressão funcional $\beta_1(t)$. À semelhança da abordagem efetuada na secção 3.2, interessa acrescentar um fator de penalização ao critério dos mínimos quadrados 6.8 para se garantir suavidade à estimativa da função $\beta_1(t)$. Usualmente, como referido também na secção 3.2, costuma-se penalizar a curvatura da função a estimar, ou seja, será adicionado ao critério dos mínimos quadrados uma penalização da forma $\int_I \{D^2\beta(t)\}^2 dt$. Obtém-se assim o critério dos mínimos quadrados penalizado:

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{W}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}, \tag{6.9}$$

onde $\mathbf{W}$ é a matriz de ordem $n \times (K+1)$, com cada linha i dada por $[1, C_{i1}, \ldots, C_{ik}]$, $\boldsymbol{\beta}$ é a matriz coluna com $K+1$ linhas, para incluir a constante β_0 e todos os coeficientes de β_1, $\lambda \geq 0$ é o parâmetro de penalização e $\mathbf{D}$ é a matriz

$$\begin{bmatrix} \mathbf{0_{1\times 1}} & \mathbf{0_{1\times K}} \\ \mathbf{0_{K\times 1}} & \mathbf{P} \end{bmatrix},$$

sendo $\mathbf{P}$ a matriz quadrada de ordem K com entradas (i,j) iguais a $\int_I D^2\beta_i(t)D^2\beta_j(t)ds$ e $\mathbf{0_{l\times c}}$ são matrizes nulas com l linhas e c colunas. O aparecimento da matriz $\mathbf{P}$ deve-se ao facto de se ter, sucessivamente, que

$$\begin{aligned} \int_I \{D^2\beta(t)\}^2 dt &= \int_I D^2\beta(t)D^2\beta(t)dt \\ &= \int_I \beta_1' D^2\phi(t)D^2\phi(t)'\beta_1 dt \\ &= \beta_1' \left[\int_I D^2\phi(t)D^2\phi(t)'dt\right]\beta_1 \\ &= \beta_1'\mathbf{P}\beta_1. \end{aligned} \tag{6.10}$$

É usual que a base de funções seja com funções B-spline, pois, de acordo com Crainiceanu et al. [3], as funções B-spline são mais flexíveis e numericamente estáveis. Quanto ao fator de penalização na estimação do coeficiente funcional, este permite controlar o grau de suavidade: uma penalização maior leva a um coeficiente funcional mais achatado, enquanto uma penalização menor permite um coeficiente funcional mais flexível, mas mais ondulado. Uma outra vantagem é que, na expressão 6.6, ao contrário do método sem penalização, permite agora utilizar-se um número K maior de elementos da base de funções.

6.2 Intervalo de confiança para $\beta_1(t)$

À semelhança da regressão linear não funcional, a construção de um intervalo de confiança para $\beta_1(t)$ parte de um seu estimador pontual, $\hat{\beta}_1(t)$, e será necessário conhecer-se o erro padrão desse estimador para cada ponto t, que, em aplicações, terá de ser um número finito de pontos. Sendo, obviamente, o erro padrão a raiz quadrada da variância do estimador, esta variância para $\hat{\beta}_1(t)$ é dada por

$$var\{\hat{\beta}_1(t)\} = \mathbf{\Phi}(t) var(\hat{\boldsymbol{\beta}}_1) \mathbf{\Phi}'(t), \tag{6.11}$$

tendo-se em conta que $\hat{\beta}_1(t) = \mathbf{\Phi}(t)\boldsymbol{\beta_1}$ é a forma matricial da expansão 6.6. Se assumirmos que o erro ε segue uma distribuição normal de média zero e variância σ_ε^2 constante, deduz-se facilmente que um intervalo de confiança para $\beta_1(t)$ com probabilidade $1-\alpha$ é da forma:

$$\hat{\beta}_1(t) \pm z_{1-\frac{\alpha}{2}} \sqrt{var(\hat{\beta}_1(t))}, \tag{6.12}$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição Normal padrão. Quando se fixa uma base de funções e se estimam todos os coeficientes (β_0 e todos os coeficientes de β_1) com o método dos mínimos quadrados ordinário (isto é, sem penalização), $var(\hat{\boldsymbol{\beta}}_1)$ em 6.11 é determinada a partir da matriz $\mathbf{W}$, à semelhança do que acontece para o modelo linear usual, isto é, tem-se que,

$$var(\hat{\boldsymbol{\beta}}_\mathbf{1}) = \hat{\sigma}_\varepsilon^2 (\mathbf{W}'\mathbf{W})$$

e, para cada ponto t, $\hat{\sigma}_\varepsilon^2$ é a variância amostral dos resíduos

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \sum_{k=1}^{K} \hat{\beta}_{1k}\phi_k(t).$$

Com o método dos mínimos quadrados penalizado a obtenção do intervalo de confiança é conceptualmente similar e, novamente, resulta de 6.11, mas o cálculo de $var(\hat{\boldsymbol{\beta}}_1)$ já é distinto. Este cálculo não será aqui discutido, pois vai além dos objetivos do presente trabalho. Alguns pormenores acerca deste assunto podem ser consultados em [3, secção 4.3].

6.3 O Modelo de Regressão Logística Funcional

Nos casos em que a variável dependente segue uma distribuição binomial, de Poisson ou outra pertencente à família exponencial, o modelo 6.2 terá de ser adaptado, inserindo-se esta adaptação nos modelos lineares funcionais generalizados, que não é mais do que a versão funcional dos modelos lineares generalizados.

No caso particular em que a variável dependente é dicotómica, o modelo 6.2 não é, obviamente, diretamente aplicável, já que, não se pode assumir a linearidade quando a variável dependente é categórica. Mas, se definirmos que $E[Y_i|X_i(t) = x_i(t) : t \in I] = \mu_i$, o mo-

delo 6.2 transforma-se no modelo

$$g(\mu_i) = \beta_0 + \int_I \beta_1(t) X_i(t) dt + \varepsilon_i, \tag{6.13}$$

onde $g(.)$ é uma função de ligação. Existem muitas funções de ligação para se estabelecer a relação entre μ_i e um modelo linear. Mas, para os objetivos deste trabalho, limitar-nos-emos à função de ligação *logit*:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),$$

que é a inversa da função de distribuição logística padrão. Para a função de ligação *logit*, o modelo 6.13 é o bem conhecido modelo de regressão logística, mas agora na sua vertente funcional. À semelhança do modelo logístico usual, o coeficiente $\beta_1(t)$ em 6.13 é interpretado, para cada t, como o logaritmo da razão de chances (*log-odds*) e a sua exponencial a razão de chances (*odds ratio*).

Neste contexto, o método dos mínimos quadrados penalizados não permite estimar os coeficientes em 6.13. É necessário recorrer-se ao método da máxima verosimilhança penalizada. Detalhes sobre a utilização deste método na estimação dos coeficientes em 6.13 podem ser obtidos, por exemplo, em [10, secção 6.1]. Um intervalo de confiança para $\beta_1(t)$ tem a mesma forma do intervalo 6.12, mas agora a variância do estimador de $\beta_1(t)$ também requer ser estimada pelo método da máxima verosimilhança. Detalhes acerca da construção deste intervalo de confiança podem ser encontrados em [6] para uma extensão do modelo 6.13, que adiciona variáveis independentes escalares.

O modo como um modelo de regressão logística usual classifica ou discrimina os indivíduos no contexto de um estudo tem assim um análogo na vertente funcional. Veremos como isto é feito na secção seguinte para dados funcionais reais.

6.4 Classificador de Artrite

Para os dados funcionais do momento de força articular dos tornozelos, direito e esquerdo, dos dois grupos de mulheres (18 mulheres diagnosticadas com artrite reumatoide; 18 mulheres saudáveis, que serviram de controlo) foram obtidos dois modelos logísticos funcionais, um para cada tornozelo, recorrendo-se ao modelo 6.13 com a função de ligação *logit*. Obviamente, a variável dependente foi a variável categórica dicotómica com as categorias "com artrite" e "controlo".

Previamente, os dados do momento de força articular de cada tornozelo foram transformados em curvas suaves com B-splines, como descrito no final da secção 3.3, e todas essas curvas foram alinhadas para o intervalo de tempo $[0; 0{,}950]$ segundos, como descrito na secção 5.1, e alinhadas continuamente para se corrigir as variações de fase (ver secção 5.3). Nas figuras 6.1 e 6.2 estão representados os coeficientes de regressão funcionais que resultaram da aplicação do modelo 6.13 ao momento de força articular dos tornozelos direito

e esquerdo, respetivamente (ver código R no apêndice G)

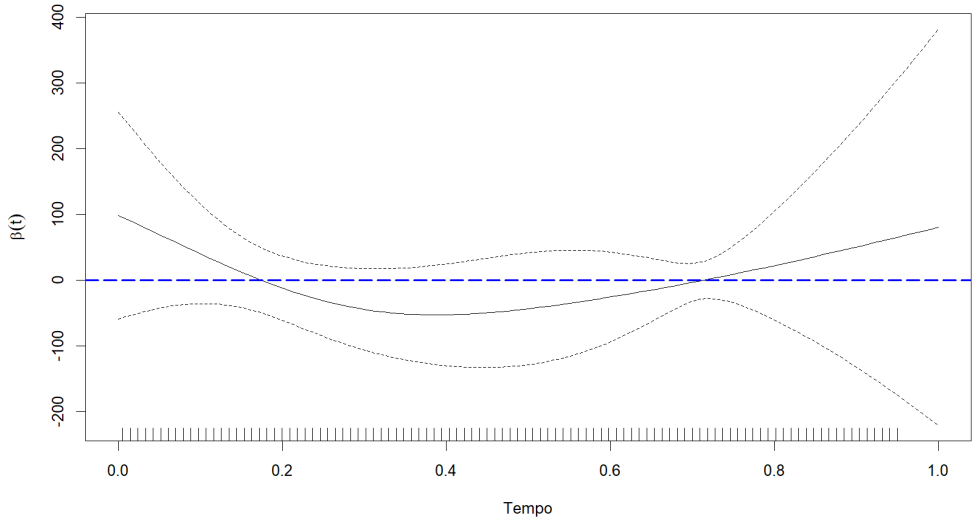


Figura 6.1: $\beta(t)$ do modelo logístico funcional com o momento de força articular do tornozelo direito. Para este modelo a constante, β_0, foi estimada em $4{,}140$.

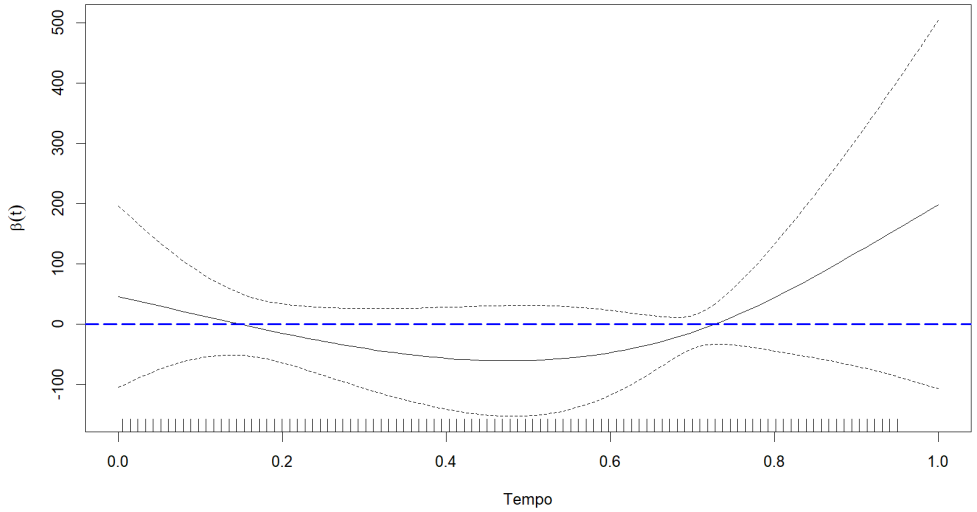


Figura 6.2: $\beta(t)$ do modelo logístico funcional com o momento de força articular do tornozelo esquerdo. Para este modelo a constante, β_0, foi estimada em $3{,}705$.

Destes gráficos constata-se que, para cada instante de tempo considerado, os coeficientes de regressão não foram significativamente diferentes de zero, para o nível de significância de 5%, pois, verifica-se, para cada tornozelo, que a reta horizontal $y = 0$ encontra-se dentro dos limites a 95% do intervalo de confiança para o verdadeiro coeficiente de regressão funcional. Do teste do rácio de verosimilhanças, adaptado para o modelo logístico funcional, cujos valores de prova foram, aproximadamente, iguais $0{,}158$ e $0{,}363$ para os tornozelos direito e esquerdo, respetivamente, obtém-se a mesma conclusão. É previsível que tal

tenha acontecido devido ao facto de o número de indivíduos em cada grupo ser reduzido. Observe-se ainda que a *deviance* explicada foram, aproximadamente, iguais a $16{,}5\%$, para o tornozelo direito, e 12.7%, para o tornozelo esquerdo. Tenha-se em conta que a *deviance* é a diferença entre o logaritmo natural da verosimilhança do modelo saturado (D; modelo com as variáveis independentes) e o logaritmo natural da verosimilhança do modelo nulo (D_0; modelo sem variáveis independentes). A *deviance* explicada é definida por:

$$R^2 = 1 - \frac{D}{D_0},$$

Observe-se este R^2 toma valores entre 0 e 1, mas não tem a mesma interpretação do coeficiente de determinação da regressão linear, e somente indica quanto próximo do ajuste perfeito, isto é, de 1, está o modelo logístico obtido.

Embora os coeficientes de regressão para cada tornozelo não se tenham mostrado significativos, como explicado anteriormente, os respetivos modelos logísticos funcionais com o momento de força mostraram um bom poder de discriminação dos indivíduos, como se pode constatar da respetiva análise ROC, que se encontram representadas nas figuras 6.3 e 6.4. As áreas ROC foram, aproximadamente, iguais a $0{,}790$ e $0{,}781$, para os modelos dos tornozelos direito e esquerdo, respetivamente. Novamente, constata-se que a amostra em cada grupo é pequena, devido ao facto de os intervalos de confiança a 95% para as verdadeiras áreas ROC apresentarem amplitude considerável.

Com as amostras disponíveis, estimaram-se duas probabilidades de corte para o modelo do tornozelo direito para o critério:

$$\max(\text{sensibilidade} + \text{especificidade}).$$

Mas, para o critério que minimiza a distância da curva ROC ao vértice $(1{,}1)$ do quadrado de área total 1, e sendo este critério o mais utilizado, a sensibilidade e a especificidade do modelo logístico com o momento de força articular do tornozelo direito foram $72{,}2\%$ e $83{,}3\%$, respetivamente, para a probabilidade de corte $0{,}473$. Para o modelo logístico com o momento de força articular do tornozelo esquerdo, os valores dessas medidas foram $94{,}4\%$ e $61{,}1\%$, respetivamente, para a probabilidade de corte $0{,}391$. Ambos os modelos, para as probabilidades de corte consideradas, apresentaram a mesma acurácia: aproximadamente $77{,}8\%$ (ver código R para a análise ROC no apêndice G).

Assumindo-se que os modelos propostos são suficientemente bons, o que não é descabido, pelo que foi anteriormente exposto, é visível nos gráficos dos coeficientes de regressão do modelo de cada tornozelo (ver figuras 6.1 e 6.2), que entre, aproximadamente, $0{,}17$ e $0{,}71$ segundos, para o tornozelo direito, e entre, aproximadamente, $0{,}14$ e $0{,}73$ segundos, para o tornozelo esquerdo, a chance de uma mulher apresentar artrite é menor e fora desses intervalos a chance de apresentar artrite é maior.

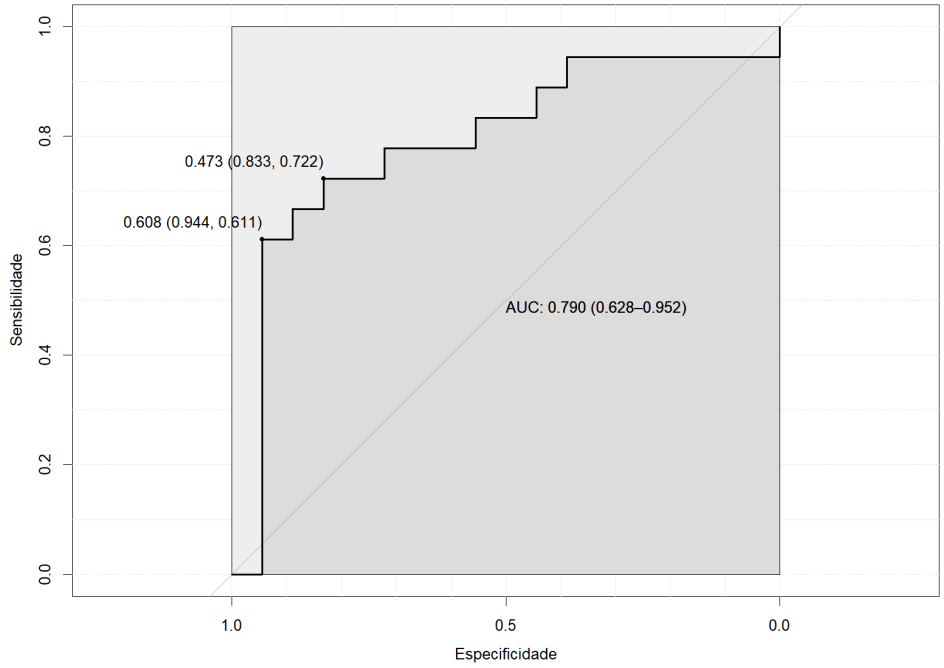


Figura 6.3: Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo direito.

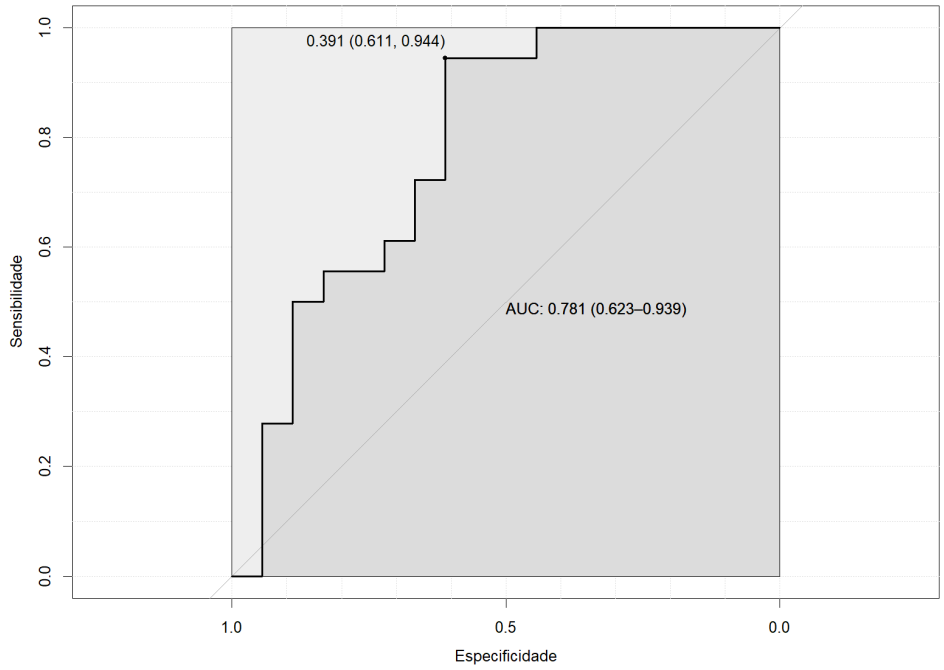


Figura 6.4: Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo esquerdo.

No entanto, os classificadores apresentados podem ser melhorados, se forem excluídas duas observações, que são claramente *outliers*, como se pode observar nos gráficos 6.5 e 6.6 dos resíduos estandardizados de Pearson dos modelos logísticos para os tornozelos

direito e esquerdo. Os resíduos estandardizados de Pearson são definidos por:

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-b_i)}}, i = 1, 2, \ldots, n,$$

onde $\hat{p}_i$ e b_i representa a probabilidade predita e a *leverage* da observação i, respetivamente.

Existem outros *outliers*, dependendo do critério adotado, mas, claramente, as observações identificadas por "ar18" e "sau8", que pertencem aos grupos artrítico e controlo, respetivamente, mostraram-se muito discrepantes e, como mais abaixo veremos, influentes na estimação dos modelos logísticos, em particular para o modelo do tornozelo direito, cujos valores dos resíduos estandardizados de Pearson foram aproximadamente iguais a $21{,}70$ e -32.14, respetivamente.

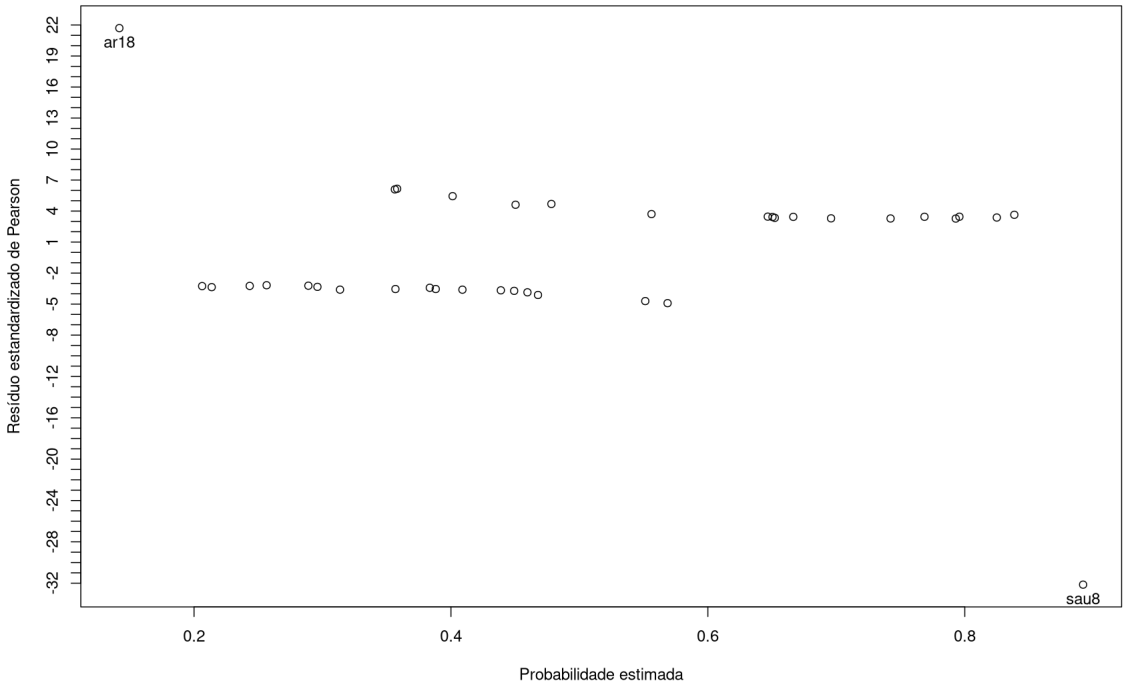


Figura 6.5: Resíduos estandardizados de Pearson do modelo logístico funcional com o momento de força articular do tornozelo direito.

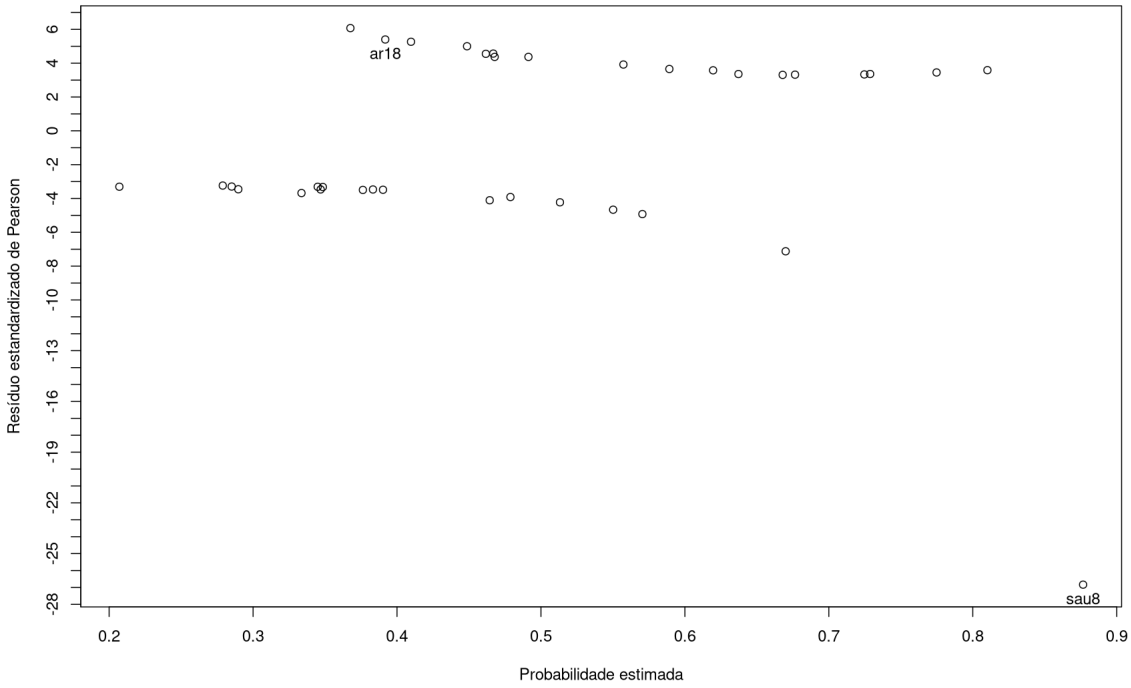


Figura 6.6: Resíduos estandardizados de Pearson do modelo logístico funcional com o momento de força articular do tornozelo esquerdo.

Com a exclusão das duas observações referidas, admitindo-se que, eventualmente, possam ter sido incorretos os valores obtidos para o momento de força, os novos modelos logísticos (ver figuras 6.7 e 6.8) apresentam maior poder de discriminação de artrite, como se pode constatar pelos gráficos da análise ROC dos dois modelos (ver figuras 6.9 e 6.10). Para o tornozelo direito a sensibilidade e especificidade passaram a ser 76,5% e 88,2%, respetivamente, para a probabilidade de corte $0{,}417$ (critério da distância mínima ao vértice $(1{,}1)$), enquanto para o tornozelo esquerdo esses valores passaram a ser $70{,}6\%$ e $88{,}2\%$, respetivamente, para a probabilidade de corte $0{,}411$ (com o mesmo referido critério). As acurácias destes modelos logísticos para os tornozelos direito e esquerdo foram aproximadamente iguais a $82{,}4\%$ e $79{,}4\%$, respetivamente. Verificou-se ainda que a *deviance* explicada aumentou consideravelmente para $43{,}7\%$ e $31{,}8\%$ para os modelos relativos aos tornozelos direito esquerdo, respetivamente.

Observe-se ainda que, com a exclusão dos dois *outliers* referidos anteriormente, o modelo logístico com o momento de força articular do tornozelo direito apresentou um coeficiente de regressão do *logit* significativamente diferente de zero no intervalo aproximado $[0{,}327; 0{,}397]$ segundos, para o nível de significância de 5%. Assim, neste intervalo a chance de uma mulher ser classificada com artrite foi significativamente menor que a chance de ser classificada sem artrite. Para o intervalo aproximado $[0{,}174; 0{,}715]$ segundos a conclusão é a mesma, mas a significância estatística referida somente foi no intervalo $[0{,}327; 0{,}397]$ segundos. Fora do intervalo $[0{,}174; 0{,}715]$ segundos, a chance de uma mulher ser classificada com artrite foi maior do que a chance de ser classificada sem artrite, mas o coeficiente de regressão do *logit* nunca se mostrou significativamente diferente de zero

neste intervalo de tempo, para o nível de significância de 5%.

Para o modelo logístico relativo ao tornozelo esquerdo, o coeficiente de regressão do *logit* nunca se mostrou significativamente diferente de zero, para o nível de significância de 5%. No entanto, embora sem esta significância, estimou-se uma chance menor de uma mulher ser classificada com artrite do que ser classificada sem artrite no intervalo aproximado $[0{,}143; 0{,}736]$ segundos. Fora deste intervalo, a chance foi maior de uma mulher ser classificada com artrite do que ser classificada sem artrite.

Estes modelos logísticos, apresentados no presente trabalho, parecem ser úteis para a classificação de artrite reumatoide em mulheres pós-menopáusicas. No entanto, carecem de validação com uma amostra maior. Poderia também ter sido útil se tivesse sido obtida a informação acerca do pé dominante das participantes do estudo.

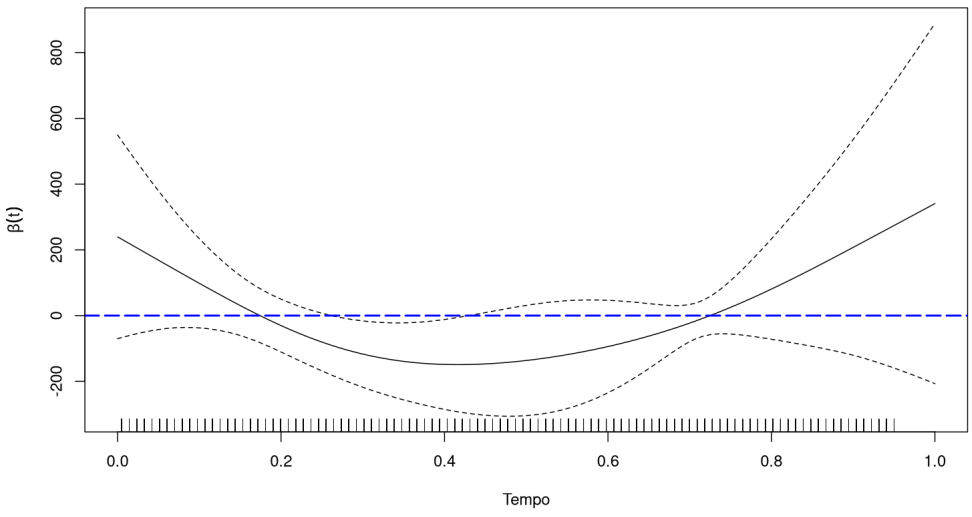


Figura 6.7: $\beta(t)$ do modelo logístico funcional, com o momento de força articular do tornozelo direito, obtido com a exclusão dos dois *outliers*. Para este modelo a constante, β_0, foi estimada em $11{,}508$.

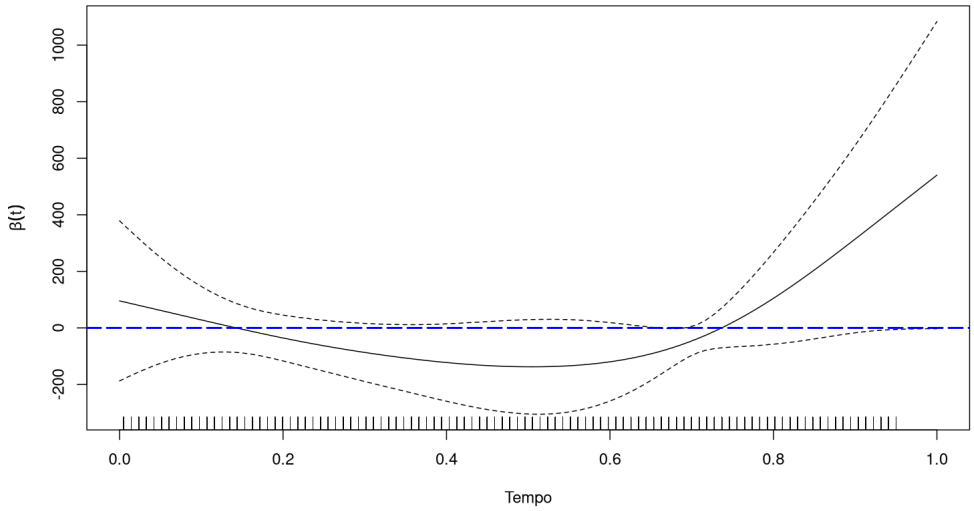


Figura 6.8: $\beta(t)$ do modelo logístico funcional, com o momento de força articular do tornozelo esquerdo, obtido com a exclusão dos dois *outliers*. Para este modelo a constante, β_0, foi estimada em $8{,}726$.

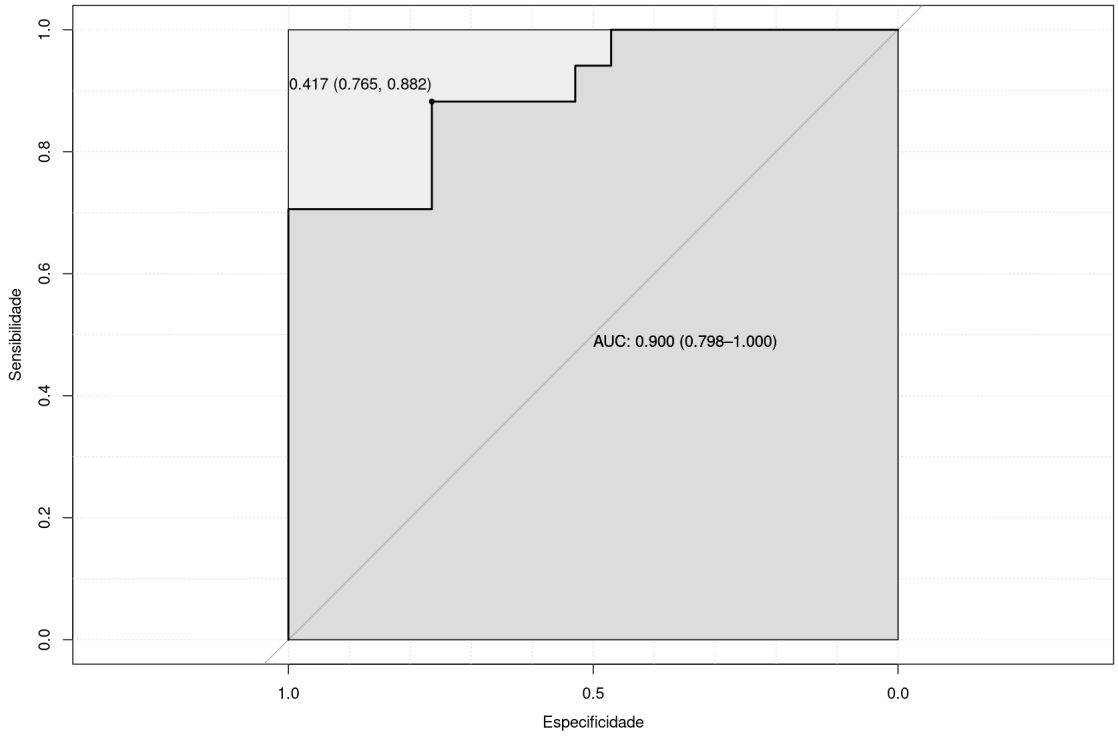


Figura 6.9: Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo direito, quando excluídos os dois *outliers*.

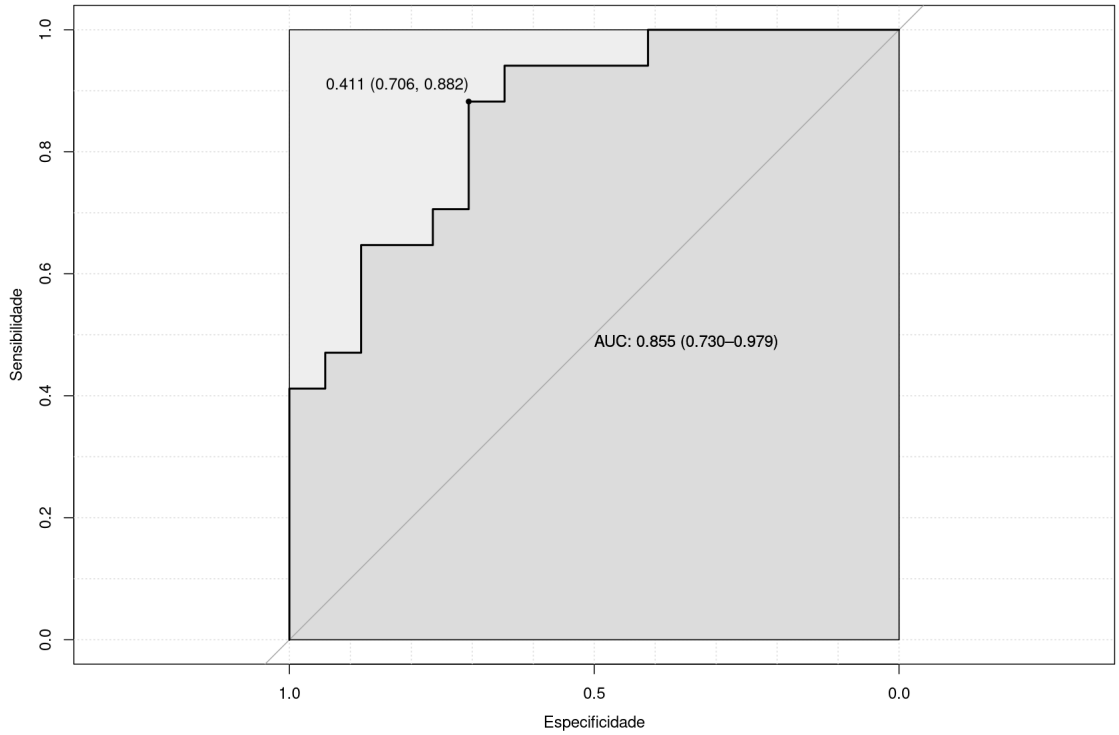


Figura 6.10: Análise ROC do modelo logístico funcional com o momento de força articular do tornozelo esquerdo, quando excluídos os dois *outliers*.

Bibliografia

[1] Aleixo, P. , Potto, J. V. , Cardoso, A., Helena, M., Abrantes, J. (2019). Ankle kinematics and kinetics during gait in healthy and rheumatoid arthritis post-menopausal women. *Somatosensory and Motor Research*, 36(2), 1-8. 2, 3

[2] Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1), 50-62. https://doi.org/10.1016/0021-9045(72)90080-9 8

[3] Crainiceanu, M. C. , Goldsmith, J. , Leroux, A., & Cui, E. (2024). *Functional Data Analysis with R*. CRC Press. 34, 35, 36

[4] Craven, P., & Wahba, G.(1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4), 377-403. https://doi.org/10.1007/BF01404567 15

[5] Dias, R., & Souza, C. P. E. (2010). *Introdução à Análise de Dados Funcionais*. São Pedro - SP - Brazil. 6, 15

[6] Goldsmith, J. , Bobb, J. , Craineceanu, C. M. , Caffo, B., & Reich, D. (2012). Penalized Functional Regression. *Journal of Computacional and Graphical Statistics*, 20(4), 830-851. http://dx.doi.org/10.1198/jcgs.2010.10007 4, 37

[7] Goldsmith, J. , Scheipl, F. , Huang, L. , Wrobel, J. , Di, C. , Gellar, J. , Harezlak, J. , McLean, M. W. , Swihart, B. , Xiao, L. , Crainiceanu, C., & Reiss, P. T.(2024). refund: Regression with Functional Data. *R package version 0.1-37*. https://cran.r-project.org/web/packages/refund/refund.pdf

[8] Page, A. , Ayala, G. , León, M.T. , Peydro, M.F., & Prat, J.M. (2006). Normalizing temporal patterns to analyze sit-to-stand movements by using registration of functional data. *Journal of Biomechanics*, 39(1), 2526-2534. https://doi.org/10.1016/j.jbiomech.2005.07.032 25

[9] Kneip, A., & Ramsay, J. O.(2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483), 1155-1165. 31

[10] Kokoszka, P., & Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Taylor & Francis Group. 1, 6, 13, 20, 21, 34, 37, 51

[11] Ramsay, J.O., & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer Verlag New York, Inc. 7, 8, 15

[12] Ramsay, J.O., & Silverman, B. W. (2005). *Functional Data Analysis*. Springer Science+Business Media, Inc. 11, 28

[13] Ramsey, J. O. , Hooker, G., & Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer Science+Business Media. 4, 25

[14] Wang, J. , Chiou, J., & Muller, H. (2015, 07, 18). Review of Functional Data Analysis. *arxiv*. https://arxiv.org/abs/1507.05135v1

[15] Zhang, J. (2014). *Analysis of Variance for Functional Data*. Taylor and Francis Group.

Apêndice A

O Espaço de Hilbert L^2

Frequentemente considera-se que as funções aleatórias são independentes e partilham a mesma distribuição de probabilidade [10]. Para a realização de inferências estatísticas, por exemplo, testes de hipóteses e estimativas com medidas de incerteza associadas, devemos considerar as funções observadas como elementos de algum espaço. Estas funções, se forem quadrado integráveis, garante-se, em particular, a existência do análogo da variância e da função de covariância. Usualmente, os dados funcionais são funções x_n em L^2 definidas num intervalo comum.
A maioria das ferramentas da FDA utiliza a teoria dos espaços de Hilbert como suporte. Um espaço de Hilbert é um espaço vetorial equipado com um produto interno, que induz uma distância, implicando que seja um espaço métrico completo. O caso particular de um espaço de Hilbert que interessa para o presente trabalho é o conhecido espaço L^2, que é o conjunto de todas as funções quadrado integráveis num intervalo I (simplificaremos a escrita escrevendo L^2 no lugar de $L^2(I)$). Assim, para qualquer $f \in L^2$ tem-se que,

$$\int_I f^2(t)dt = \int_I \{f(t)\}^2 dt < \infty. \tag{A.1}$$

L^2 constitui um espaço vetorial, isto é, satisfaz a propriedade: se $f, g \in L^2$, então $af + bg \in L^2$, para quaisquer números reais a e b, e para o qualquer se define o produto interno,

$$\langle f, g\rangle = \int_I f(t)g(t)dt. \tag{A.2}$$

Se o f e g são ortogonais, então $\langle f, g\rangle = 0$. Este produto interno permite equipar L^2 com uma norma, que é definida por,

$$\|f\| = \sqrt{\langle f, f\rangle} = \left\{\int_I f(t)^2 dt\right\}^{\frac{1}{2}}, \tag{A.3}$$

e induz uma distância entre duas funções de L^2, que é simplesmente a norma da diferença entre as duas: $d(f,g) = \|f - g\|$, para $f,g \in L^2$.
Verifica-se em L^2 a conhecida desigualdade de Cauchy-Schwarz:

$$\left|\int_I f(t)g(t)dt\right| = |\langle f,g\rangle| \leq \|f\|\|g\| = \left\{\int_I f^2(t)dt\right\}^{\frac{1}{2}} \left\{\int_I g^2(t)dt\right\}^{\frac{1}{2}}. \tag{A.4}$$

O espaço L^2 satisfaz também outras propriedades, como a desigualdade triangular:

$$\|f + g\| \leq \|f\|\|g\|. \tag{A.5}$$

ou a linearidade do produto interno:

$$\langle af + bg, h\rangle = a\langle f, h\rangle + b\langle g, h\rangle. \tag{A.6}$$

Sendo as bases de funções importantes na construção de dados funcionais, um conjunto de funções $\{e_1, e_2, e_3, ...\}$ é uma base em L^2, se qualquer $f \in L^2$ tem uma única expansão

$$f(t) = \sum_{j=1}^{\infty} c_j e_j(t). \tag{A.7}$$

Tem-se, portanto, que L^2 é um espaço de dimensão infinita. Adicionalmente, a base $\{e_1, e_2, e_3, ...\}$ é uma base ortonormada, se $\langle e_j, e_{j'}\rangle = 0$, com $e_j \neq e_{j'}$, e $\|e_j\| = 1$. Uma base ortonormada de L^2 satisfaz a igualdade de Parseval:

$$\int_I f^2(t)dt = \| f \|^2 = \sum_{j=1}^{\infty} \langle f, e_j\rangle^2 = \sum_{j=1}^{\infty} \left\{ \int_I f(t)e_j(t)dt \right\}^2. \tag{A.8}$$

Apêndice B

Cálculo da Média das 6 Réplicas

Código R com os pacotes *fda* e *haven* (importação e exportação de dados) para o cálculo da média das 6 réplicas do momento de força de cada indivíduo da amostra.

```
library(fda)
library(haven)

# Função para guardar por coluna as médias das 6 réplicas de cada indivíduo
# da amostra
cbind.fill<-function(...){
 nm <- list(...)
 nm<-lapply(nm, as.matrix)
 n <- max(sapply(nm, nrow))
 do.call(cbind, lapply(nm, function (x)
 rbind(x, matrix(, n-nrow(x), ncol(x)))))
}

# Primeiro foram importados para o R os dados do tempo e das 6 réplicas do
# momento de força de cada um dos indivíduos a partir de uma base de dados
# SPSS. Por exemplo, a base de dados para o indivíduo 1 do grupo artrítico
# tinha o nome: ar01dto.sav
dados=data.frame(ar01dto[c(1,3,5,7,9,11,13)])
attach(dados)
dados[is.na(dados)] <- 0

knots    <- dados$tempo
norder   <- 6
nbasis   <- length(knots) + norder - 2
arbasis <- create.bspline.basis(range(knots), nbasis, norder, knots)
Lfdobj <- 4
lambda <- 1e-15 #2e-26
arfdPar <- fdPar(arbasis, Lfdobj, lambda)

arfd <- smooth.basis(as.vector(knots), as.matrix(dados[,2:7]), arfdPar)$fd

# Guardar a média para o primeiro indivíduo numa base de dados
med_ardto <- data.frame(cbind.fill(mean.fd(arfd)$coefs))
# Para guardar a média dos restantes indivíduos na mesma base de dados
med_ardto <- data.frame(cbind.fill(med_ardto,mean.fd(arfd)$coefs))
```

```
med_ardto[is.na(med_ardto)] <- 0

# Exportar a base de dados para uma base de dados SPSS
write_sav(med_ardto, "med_mf_saudto.sav")
```

Apêndice C

Exemplos de Bases de Funções

Código R com o pacote *fda* para a elaboração dos gráficos das figuras 2.1 e 2.2.

```
library(fda)
exbsline = create.bspline.basis(c(0,1), 10)
plot(exbsline, xlab="t", ylab="Funções Bspline", lwd=2)
exfourier=create.fourier.basis(c(0,1), 5)
plot(exfourier, xlab="t", ylab="Funções de Fourier", lwd=2)
```

Apêndice D

Estimação de λ por GCV

Código R com o pacote *fda* para a procura do parâmetro λ por GCV.

```
# GCV
library(fda)
lambda=seq(1e-30,1e-8,1e-30)
Lfdobj <- 4
gcv1=seq(1,length(lambda),1)
for (i in 1:length(lambda)) {
   arfdPar <- fdPar(arbasis, Lfdobj, lambda[i])
   artofd1 <- smooth.basis(as.vector(tempo), as.matrix(dados[,2:7]),
                           arfdPar)$gcv
   gcv1[i]<-mean(artofd1)
 }
min(gcv1)
list(cbind(lambda,gcv1))
```

Apêndice E

Gráficos das Curvas

Código R com o pacote *fda* para a elaboração dos gráficos das figuras 3.1, 4.1, 4.2 e 4.3.

```
library(fda)
# O ficheiro med_mf_dto continha os dados do tempo e das 36 curvas médias
# do momento de força articular do tornozelo direito: 18 do grupo
# artrítico e 18 do grupo controlo.
dados=data.frame(med_mf_dto)
attach(dados)
dados[is.na(dados)] <- 0

# Criação da base B-spline, com 197 funções B-spline de ordem 6, e
# suavização pelo método dos mínimos quadrados penalizados com parâmetro
# de penalização igual a 10e-15 sobre a derivada de ordem 4.
knots    <- dados$tempo
norder   <- 6
nbasis   <- length(knots) + norder - 2
arbasis <- create.bspline.basis(range(knots), nbasis, norder, knots)
Lfdobj <- 4
lambda <- 1e-15 #2e-26
arfdPar <- fdPar(arbasis, Lfdobj, lambda)
arfd <- smooth.basis(as.vector(knots), as.matrix(dados[,2:19]), arfdPar)$fd
saufd <- smooth.basis(as.vector(knots), as.matrix(dados[,20:37]), arfdPar)$fd

# Gráficos das 18 curvas do momento de força dos grupos artrítico e controlo,
# respetivamente.
plot(arfd, xlab="Tempo", ylab="Momento de Força",lwd=1.5)
plot(saufd, xlab="Tempo", ylab="Momento de Força", lwd=1.5)

# Gráfico das médias do momento de força dos dois grupos.
oldpar<- par(no.readonly=TRUE)
plot(mean.fd(saufd),lwd=2)
lines(mean.fd(arfd),lty=2,lwd=2)
par(oldpar)

# Gráficos da média, superfície das variâncias-covariâncias e respetivas
# curvas de nível do momento de força articular do grupo artrítico.
plot(mean.fd(arfd))
arvar=var.fd(arfd)
```

```
arvar_mat = eval.bifd(as.vector(tempo), as.vector(tempo), arvar)
persp(as.vector(tempo), as.vector(tempo), arvar_mat,
theta=25, phi=25, r=3, expand = 0.5,
ticktype="detailed",
xlab="Tempo",
ylab="Tempo",
zlab="Variância-covariância do Momento de Força")
contour(as.vector(tempo), as.vector(tempo), arvar_mat)

# Gráficos da média, superfície das variâncias-covariâncias e respetivas
# curvas de nível do momento de força do grupo controlo.
plot(mean.fd(saufd))
sauvar=var.fd(saufd)
sauvar_mat = eval.bifd(as.vector(tempo), as.vector(tempo), sauvar)
persp(as.vector(tempo), as.vector(tempo), sauvar_mat,
theta=25, phi=25, r=3, expand = 0.5,
ticktype="detailed",
xlab="Tempo",
ylab="Tempo",
zlab="Variância-covariância do Momento de Força")
contour(as.vector(tempo), as.vector(tempo), sauvar_mat)
```

Código R com o pacote *fda* para se efetuar o teste t e obtenção do respetivo gráfico.

```
oldpar<- par(no.readonly=TRUE)
tres <- tperm.fd(saufd,arfd,nperm=200,plotres=FALSE)
plot(tres$argvals, tres$Tvals, type = "l", main = "Resultado do Teste",
ylab = "Estatística t", xlab = "Tempo", bty = "l",lty=1, lwd=2)
lines(tres$argvals, tres$qvals.pts,lty=2, lwd=2,col="blue")
abline(h=tres$qval, lty=5, lwd=2, col="blue")
legend("bottomright", inset=c(0.05,0.05), c("Estatística Observada",
"valor crítico (pontual) a 0.05", "valor crítico máximo a 0.05"),
lty=c(1,2,5), lwd=2,col=c("black","blue","blue"))
par(oldpar)
```

Apêndice F

Alinhamentos por Deslocamento e Contínuo

Código R com os pacotes *fda* e *haven* para o alinhamento das curvas do momento de força por deslocamento para o tempo máximo de 0,950 segundos.

```
library(fda)
library(haven)

cbind.fill<-function(...){
 nm <- list(...)
 nm<-lapply(nm, as.matrix)
 n <- max(sapply(nm, nrow))
 do.call(cbind, lapply(nm, function (x)
 rbind(x, matrix(, n-nrow(x), ncol(x)))))
}

# Como exemplo, importação dos dados das bases de dados SPSS dos indivíduos
# com artrite.

setwd("~/BD/ar_dto")
files_sav <- list.files(path = "~/BD/ar_dto",
pattern = "\\.sav$", full.names = TRUE)

for (i in 1:18) {
 data<- read_sav(files_sav[i])
 dados=data.frame(data[c(1,3,5,7,9,11,13)])
 attach(dados)

 for (j in 2:7) {
  knots    <- as.vector(tempo[1:length(na.omit(dados[,j]))])
  norder   <- 6
  nbasis   <- length(knots) + norder - 2
  arbasis <- create.bspline.basis(range(knots), nbasis, norder, knots)
  Lfdobj <- 4
  lambda <- 1e-15 #2e-26
  arfdPar <- fdPar(arbasis, Lfdobj, lambda)
  arfd <- smooth.basis(knots, as.matrix(na.omit(dados[,j])), arfdPar)$fd
  #plot(arfd)

  areval=eval.fd(seq(min(knots),max(knots),length=103),arfd)
```

```
  knots0    <- seq(0.005,0.950,length=103)
  norder0   <- 6
  nbasis0   <- length(knots0) + norder - 2
  arbasis0 <- create.bspline.basis(range(knots0), nbasis0, norder0, knots0)
  Lfdobj0 <- 4
  lambda0 <- 1e-15 #2e-26
  arfdPar0 <- fdPar(arbasis0, Lfdobj0, lambda0)
  arfd0 <- smooth.basis(knots0, areval, arfdPar0)$fd
  areval0=eval.fd(knots0,arfd0)

  if (j==2) {
   warpcurves <- data.frame(cbind.fill(areval0))
  } else {
   warpcurves <- data.frame(cbind.fill(warpcurves,areval0))
  }
 }

 knots1    <- seq(0.005,0.950,length=103)
 norder1   <- 6
 nbasis1   <- length(knots1) + norder - 2
 arbasis1 <- create.bspline.basis(range(knots1), nbasis1, norder1, knots1)
 Lfdobj1 <- 4
 lambda1 <- 1e-15 #2e-26
 arfdPar1 <- fdPar(arbasis1, Lfdobj1, lambda1)
 arfd1 <- smooth.basis(knots1, as.matrix(warpcurves), arfdPar1)$fd
 areval1=eval.fd(knots1,mean.fd(arfd1))

 if (i==1) {
  armeans <- data.frame(cbind.fill(areval1))
 } else {
  armeans <- data.frame(cbind.fill(armeans,areval1))
 }
}
write_sav(armeans, "med_mf_ardto_tw.sav")
```

Código R com o pacote *fda* para o alinhamento contínuo das curvas do momento de força articular de cada um dos grupos e respetivas representações gráficas.

```
arregfd <- register.fd(yfd=arfd)
arfd=arregfd$regfd
sauregfd <- register.fd(yfd=saufd)
saufd=sauregfd$regfd
```

```
plot(arfd, xlab="Tempo", ylab="Momento de Força", lwd=1.5, xlim=c(0,1))
plot(saufd, xlab="Tempo", ylab="Momento de Força", lwd=1.5, xlim=c(0,1))
```

Código R com o pacote *fda* para o cálculo do R^2 que resultou do alinhamento contínuo das curvas do momento de força dos dois grupos.

```
AmpPhasList_ar = AmpPhaseDecomp(arfd, arregfd$regfd, arregfd$warpfd)
AmpPhasList_ar$RSQR
AmpPhasList_ar$MS.pha
AmpPhasList_ar$MS.amp

AmpPhasList_sau = AmpPhaseDecomp(saufd, sauregfd$regfd, sauregfd$warpfd)
AmpPhasList_sau$RSQR
AmpPhasList_sau$MS.pha
AmpPhasList_sau$MS.amp
```

Apêndice G

Regressão Logística Funcional

Código R com os pacotes *fda*, *refund* e *pROC* (análise ROC) para toda a análise efetuada na secção 6.4.

```
library(fda)
library(refund)

# Depois da importação dos dados do momento de força, tornozelo esquerdo
# ou direito, faz-se
c1=data.frame(med_mf_ardto_tw,med_mf_saudto_tw)
# ou
c1=data.frame(med_mf_aresq_tw,med_mf_sauesq_tw)

# Comando para a exclusão dos dois outliers
#c1=c1[,-c(18,26)]

# B-splines e critérios de suavização
knots    <- seq(0.005,0.950,length=103)
norder   <- 6
nbasis   <- length(knots) + norder - 2
arbasis <- create.bspline.basis(range(knots), nbasis, norder, knots)
Lfdobj <- 4
lambda <- 1e-15
arfdPar <- fdPar(arbasis, Lfdobj, lambda)

# Curvas do momento de força
todasfd1 <- smooth.basis(as.vector(knots), as.matrix(c1[,1:36]), arfdPar)$fd

# Alinhamento contínuo
todasregfd1 <- register.fd(yfd=todasfd1)
todasfd1=todasregfd1$regfd
todaseval1=eval.fd(knots,todasfd1)
todaseval01=t(as.matrix(todaseval1))

# Criação do Modelo Logístico
Y=vector()
Y[1:18]=1
Y[19:36]=0
ar_fit<-pfr(Y~lf(todaseval01, argvals=knots, presmooth="bspline"),
```

```
family=binomial(link="logit"))

# Gráfico do coeficiente e outras estatísticas de interesse
plot(ar_fit, xlab="Tempo", ylab=expression(paste(beta(t))), xlim=c(0,1))
abline(h=0, lty=5, lwd=2, col="blue")

summary(ar_fit)
ar_fit$deviance
ar_fit$null.deviance
ar_fit$fitted.values

# Análise ROC
library(pROC)
analise_roc=roc(Y, ar_fit$fitted.values,
smoothed = TRUE, xlab="Especificidade", ylab="Sensibilidade",
ci=TRUE, ci.alpha=0.95, stratified=FALSE,
plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
print.auc=TRUE, show.thres=TRUE,
print.thres="best", print.thres.adj = c(1, -1))
```