MSc

2.º CICLO

FCUP
2017

U.PORTO

Novos Desenvolvimentos em Análise de Dados

Fábio José Nogueira Ferreira

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Novos Desenvolvimentos em Análise de Dados

Fábio José Nogueira Ferreira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Engenharia Matemática

2017

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

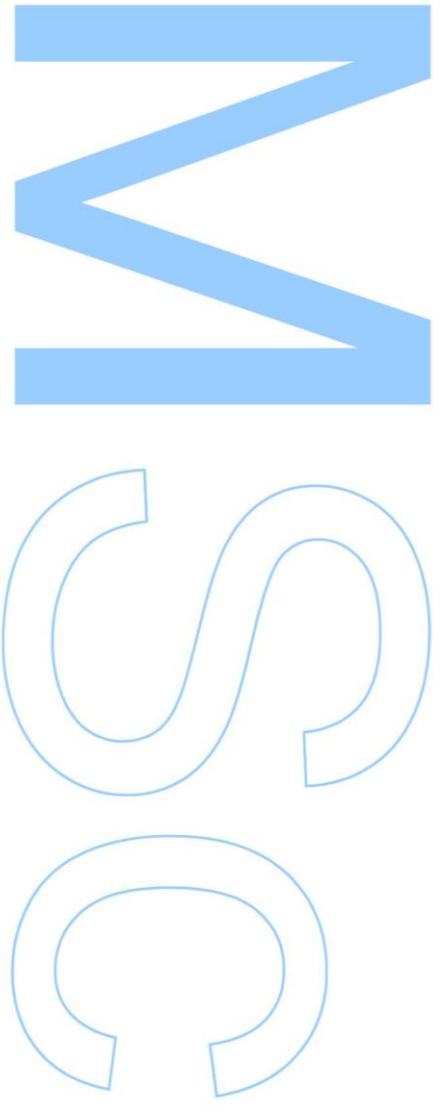
MSc

Novos Desenvolvimentos em Análise de Dados

Fábio José Nogueira Ferreira
Mestrado em Engenharia Matemática
Departamento de Matemática
2017

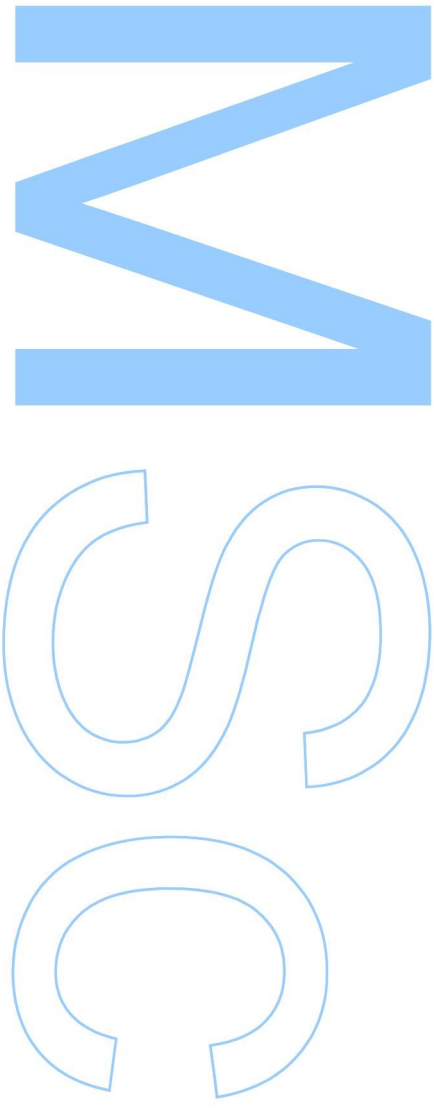
Orientadores
Joaquim Fernando Pinto da Costa, Professor / Investigador, Faculdade de Ciências da Universidade do Porto, CMUP

Ana Rita Pires Gaio, Professora / Investigadora, Faculdade de Ciências da Universidade do Porto, CMUP



Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,



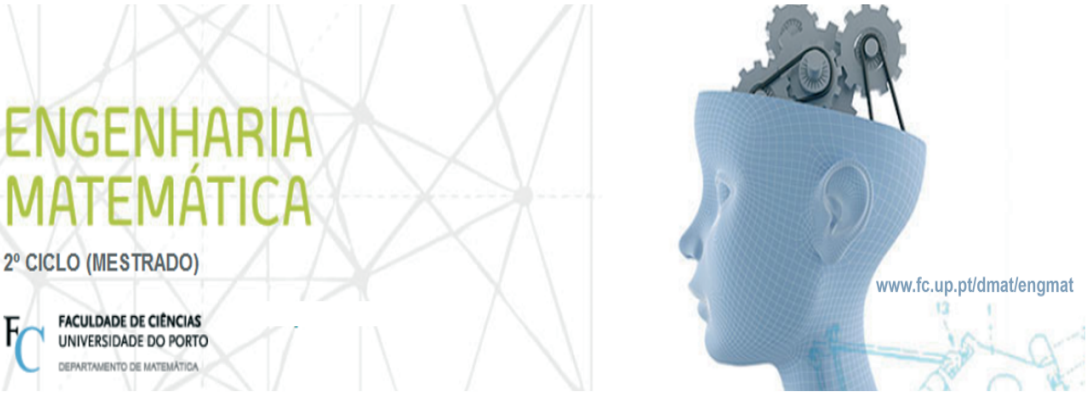
Agradecimentos

A realização desta dissertação marca o fim de uma das fases mais importantes da minha vida, portanto quero aqui deixar algumas palavras a pessoas que se mostraram essenciais nesta concretização.

Inicialmente, agradeço à minha família e amigos, com especial ênfase aos meus pais e avós, pelo apoio incondicional e incentivo para a chegada a este nível, são sem dúvida muito importantes na minha vida, toda a força e ensinamentos que me têm dado mostraram-se essenciais neste caminho.

De seguida, quero agradecer claramente à minha namorada, por toda a força, paciência e motivação que me tem dado, principalmente nos momentos de maior nervosismo. É muito bom ser apoiado naquilo que acredito, nos meus objetivos profissionais e pessoais e ainda mais ter a sorte de contermos tantos objetivos coincidentes.

Também deixo um agradecimento especial aos colegas Júlio Silva e Martina Mascarello pela disponibilidade e apoio prestados. E por fim, com a maior das considerações e admiraçções, agradeço aos meus professores orientadores, Prof. Joaquim Costa e Prof. Rita Gaio, por todo o conhecimento transmitido, disponibilidade, compreensão e apoio. Saio do mestrado com toda a certeza que não podia ter sido melhor guiado.



Tese realizada no âmbito do mestrado em Engenharia Matemática

Departamento de Matemática

Faculdade de Ciências da Universidade do Porto

http://www.fc.up.pt/dmat/engmat

Resumo

Esta dissertação propõe dois métodos inovadores de classificação. A primeira metodologia proposta enquadra-se no âmbito da classificação supervisionada e é motivada pelo facto de alguns problemas reais requererem uma classificação de objetos ou indivíduos em classes com uma ordem natural. Estes problemas são tradicionalmente resolvidos usando métodos convencionais destinados a uma classificação de classes nominais onde a noção de ordem é ignorada. A segunda metodologia proposta concerne a análise de clusters em dados longitudinais, caindo portanto no âmbito da classificação não supervisionada.

A primeira parte da dissertação é dedicada à adaptação do método tradicional de análise discriminante linear para classes ordinais. Sob o ponto de vista metodológico, utiliza o método da máxima verosimilhança e o método dos mínimos quadrados para estimação dos parâmetros. Restrições genéricas sobre as médias das classes e as probabilidades à priori são consideradas. Foram testadas três metodologias ordinais e uma parcialmente ordinal. Estas foram aplicadas em quatro conjuntos de dados reais e posteriormente comparadas com a metodologia tradicional através de quatro medidas de performance. Os resultados para uma das metodologias ordinais e para a metodologia parcialmente ordinal compararam-se favoravelmente com a análise discriminante linear usual.

O segundo método proposto nesta dissertação introduz uma metodologia de classificação não supervisionada para dados longitudinais. Um conjunto de dados é considerado longitudinal se cada indivíduo é medido repetidamente através do tempo, fazendo assim crescer um vetor de observações para cada indivíduo que tende a ter componentes substancialmente correlacionadas. Em estudos longitudinais com um grande número de trajetórias, o agrupamento destas e a obtenção de trajetórias médias podem ser de interesse. A metodologia proposta começa por modelar as correlações existentes entre observações de trajectórias individuais através de uma matriz pré-definida com parâmetros estimados dos dados. Depois é considerada uma distância do tipo Mahalanobis e o algoritmo das K-médias longitudinal é aplicado. Mostra-se depois que a metodologia desenvolvida é equivalente a usar o algoritmo longitudinal das K-médias em dados *adequadamente* transformados. Esta propriedade simplifica o processo para utilizadores gerais. Na circunstância particular de o interesse do estudo incidir sobre as trajectórias relativas (e não as absolutas), aconselhamos a utilização de um de dois perfis propostos antes da entrada do algoritmo. A metodologia foi testada em dados simulados e também em dados reais, usando trajectórias uni- e bi-dimensionais. A nova metodologia produziu em geral melhores resultados do que os obtidos por aplicação direta do algoritmo das K-médias longitudinal nos dados originais.

Palavras-chave: Classificação; Análise Discriminante Linear; Classes Ordenadas; K-Médias; Clustering Longitudinal.

Abstract

This dissertation intends to propose two innovative methods of classification. The first proposed methodology fits into the supervised classification and this is motivated by the fact that many real life problems require the classification of objects or individuals into naturally ordered classes. These problems are traditionally handled by conventional methods intended for the supervised classification of nominal classes where the order is ignored. The second proposed methodology concerns cluster analysis in longitudinal data, falling, therefore, in the scope of the unsupervised classification.

The first part of the dissertation is dedicated to the adaptation of the traditional method of linear discriminant analysis to ordinal classes. From the methodological point of view, it uses the maximum likelihood method and the least squares method to estimate the parameters. Generic restrictions on the means and prior probabilities of the classes are considered. Three ordinal and one partially ordinal methodologies were tested. The results for one of the ordinal methodologies and for the partially ordinal methodology were compared favorably with the usual Linear Discriminant Analysis.

The second method proposed in this dissertation introduces an unsupervised classification methodology for longitudinal data. The defining feature of a longitudinal data set is that individuals are measured repeatedly through time, giving rise to a vector of observations that tend to be intercorrelated. In longitudinal studies with a large number of subjects, clustering of the longitudinal trajectories and the definition of a much smaller number of mean trajectories is often of interest. The proposed methodology starts by modeling the existing correlations between observations of individual trajectories through a predefined matrix. Then a distance of the Mahalanobis type is considered and the longitudinal K-means algorithm is applied. It is later shown that the methodology developed is equivalent to using the longitudinal algorithm of the K-means in *appropriately* transformed data. This property simplifies the process for general users. Whenever the interest of the study concerns the relative (and not absolute) trajectories, we recommend using one of two proposed profiles before the algorithm enters. The methodology was tested in simulated data and also in real data, using uni- and bi-dimensional trajectories. The new methodology generally produced better results than those obtained by direct application of the longitudinal K-means algorithm in the original data.

Keywords: Classification; Linear discriminant analysis; Ordered classes; K-Means; Longitudinal clustering.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Esta dissertação propõe duas metodologias inovadoras, uma para classificação supervisionada e outra para classificação não supervisionada, sendo estas motivadas pela sua utilidade em panoramas reais ainda não muito explorados.

No capítulo 2 propomos uma metodologia específica para classificação supervisionada de classes ordinais. De facto, muitos problemas reais requerem uma classificação de objetos ou indivíduos em classes contendo uma ordem natural, porém estes problemas são tradicionalmente tratados com o uso de métodos convencionais onde a noção de ordem é ignorada. Um exemplo desses métodos é a Análise Discriminante Linear, que assume que cada classe pode ser modelada por uma função gaussiana.

Neste trabalho introduzímos três métodos de Análise Discriminante Linear específicos para classes ordinais e utilizámos o método de máxima verosimilhança, o método dos mínimos quadrados e ainda alguns métodos de otimização para estimar os parâmetros do modelo. Incluímos ainda um método que não é rigorosamente específico para dados ordinais pois a probabilidade de respeitar um comportamento ordinal é maior do que no método usual de Análise Discriminante Linear.

Para estas metodologias impusemos uma condição para que as probabilidades à posteriori seguissem um comportamento unimodal, obedecendo assim ao paradigma unimodal introduzido por J. Pinto da Costa (2005, 2008, 2010), e restrições genéricas sobre as médias das classes e as probabilidades à priori foram consideradas.

Por fim, foi criada uma função em linguagem R, na versão 3.3.2, onde testámos as metodologias propostas em quatro conjuntos de dados reais e comparámos com a metodologia tradicional através de quatro medidas de performance. Como função de otimização usámos o algoritmo *cobyla* da biblioteca *nloptr*, onde o ponto inicial a considerar é escolhido como o melhor dentro de uma grelha de pontos definida pelo utilizador. Os resultados para uma

das metodologias ordinais e para a metodologia parcialmente ordinal compararam-se favoravelmente com a Análise Discriminante Linear usual tornando-se assim uma boa alternativa, especialmente porque contêm um modelo de respostas mais apropriado a este tipo de classes.

No capítulo 3 apresentámos uma metodologia direcionada para a classificação não supervisionada em dados longitudinais. Numa primeira fase introduzimos uma nova metodologia não paramétrica de clustering de dados longitudinais. A motivação desta metodologia é que nos métodos tradicionais os agrupamentos das trajetórias apenas são feitos tomando em consideração a distância entre estas ignorando o seu comportamento ao longo do tempo, porém em muitos casos reais também é do interesse que nesse agrupamento o comportamento das trajetórias seja levado em conta.

Ao contrário dos métodos tradicionais também as correlações entre observações de trajetórias individuais são tomadas em conta com uma matriz de correlações pré-definida e parâmetros que são estimados dos dados. Assim sendo, como alternativa à distância Euclidiana usualmente utilizada no algoritmo das K-Médias longitudinal considerámos a aplicação de uma distância de Mahalanobis original que inclui esta matriz de correlações.

Para completude desta dissertação, incluimos uma breve apresentação sobre cálculo de distâncias longitudinais e ainda uma apresentação de alguns critérios para a escolha do número mais adequado de clusters, onde expomos os algoritmos e exemplos ilustrativos para uma melhor compreensão dos mesmos. A definição de alguns conceitos base necessários à compreensão destes algoritmos também foram apresentados.

Simplificámos o processo de clustering ao considerar uma transformação adequada dos dados que permite a utilização direta de um processo de clustering conhecido (e já implementado no R). De facto, mostramos que o nosso método com a nova distância Mahalanobis coincide com a aplicação do algoritmo K-Médias longitudinal usando a distância Euclidiana para certas trajetórias transformadas. Esta propriedade simplifica o processo para utilizadores gerais.

Posteriormente, sabendo que em certas circunstâncias pode ser importante o comportamento relativo em vez dos valores absolutos, aconselhamos a utilização, antes da entrada do algoritmo, de um de dois perfis propostos. Assim programámos a metodologia proposta com a utilização da linguagem R, versão 3.3.2, onde utilizámos as bibliotecas *kml* ou *kml3d* e testámos a metodologia em dados simulados com diferentes características e em dados reais. Considerámos trajetórias unidimensionais e bidimensionais para ilustrar a aplicação do método desenvolvido e os resultados foram comparados com os obtidos pela aplicação direta do algoritmo K-Médias longitudinal nos dados originais. Os resultados favoreceram a nova

metodologia.

Finalmente, gostariamos de referir que a presente tese não inclui um capítulo de considerações finais e trabalhos futuros de forma explícita. Optou-se por se fazer essa crítica no final do capítulo correspondente a cada metodologia, por se achar que isso facilitaria a leitura da tese e a tornaria mais coerente.

Capítulo 2

Análise discriminante linear para classes ordinais

2.1 Introdução

Muitos problemas de classificação supervisionada têm a particularidade das classes possuírem uma ordem natural, e neles é pretendido atribuir uma classe a uma dada observação. É frequente encontrarmos este tipo de situações em temas como a modelação econométrica, problemas de classificação biomédica, ciências sociais e comportamentais. Embora esta situação seja bastante frequente, quase nunca é considerada nos modelos tradicionais de classificação supervisionada.

Neste capítulo, utilizámos o método de máxima verosimilhança e o método dos mínimos quadrados para estimar os parâmetros da metodologia de análise discriminante linear proposta.

Considere-se o problema de classificação supervisionada com o objetivo de separar K classes ordenadas $\mathcal{C}_1 < \ldots < \mathcal{C}_K$, de um determinado espaço χ. Sejam $n_1, \ldots, n_K$ os tamanhos das classes, com $n_1 + \cdots + n_K = n$. A probabilidade à priori da classe k é π_k com $\sum_{k=1}^{K} \pi_k = 1$.

Para $k = 1, \ldots, K$ e $j = 1, \ldots, n_k$, denotamos por $x_{kj} \in \mathcal{R}^p$ o vetor da j-ésima observação da classe k. Como as classes são conhecidas desde o início, a realização de uma amostra aleatória de tamanho n consiste no par

$$\{(x_{kj}, c_k)\}_{k=1,\ldots,K;j=1,\ldots,n_k}.$$

A função (densidade) de probabilidade $X \in \chi$ na classe $\mathcal{C}_k$ é representada por $f_k(x)$.

A teoria de decisão de Bayes sugere que a classificação de uma nova observação x na

classe $\mathcal{C}_k$ maximize a probabilidade à posteriori $P(\mathcal{C}_k|x)$ dada por

$$P(\mathcal{C}_k|x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)} \tag{2.1.1}$$

O nosso método assume que num problema de classificação supervisionada com classes ordenadas, a variável aleatória classe $\mathcal{C}_x$ associada à nova observação x deve seguir uma distribuição de probabilidade unimodal, seguindo trabalhos anteriores de J. Pinto da Costa [9, 11, 12]. Isto significa que as probabilidades devem decrescer monotonamente para a esquerda e para a direita da classe onde a probabilidade máxima é obtida (ilustrado na figura 2.1). Por exemplo, suponhamos que pretendemos prever a temperatura para amanhã dentro de cinco classes: Muito frio, Frio, Ameno, Quente e Muito quente. Dada uma nova observação x, e depois de calculadas as probabilidades à posteriori $P(\mathcal{C}_1|x)$,...,$P(\mathcal{C}_5|x)$, se o valor mais alto for $P(\mathcal{C}_4|x)$, isto é, o mais provável é estar um dia Quente, então a segunda maior probabilidade deve ser um dia Muito quente ou um dia Ameno. A segunda maior probabilidade nunca devia ser, neste caso, um dia Muito frio. Por outras palavras, se a maior probabilidade é estar um dia Quente, não faz sentido que a segunda maior probabilidade seja um dia Muito frio. Porém, um facto como este nos métodos tradicionais de análise discriminante pode acontecer.

Desde o último meio século já alguns autores têm vindo a propor modelos de classificação supervisionada para classes ordinais. Em trabalhos específicos de modelos de Análise Discriminante Linear (Linear Discriminant Analysis - LDA[1]) podemos destacar B. Sun et al. [17] que propuseram em 2010 um novo método de regressão ordinal baseado na Análise Discriminante de Núcleo (Kernel Discriminant Analysis, KDA). Os autores impõem duas restrições ao modelo convencional que passa pela escolha de uma projeção que satisfaça os seguintes requisitos: 1) A projeção deve minimizar a distância dentro das classes e maximizar a distância entre classes simultaneamente. 2) A projeção deve garantir a informação ordinal das diferentes classes, ou seja, a projeção média das observações das classes mais altas devem ser maiores do que a projeção das classes de classificação mais baixas. W. Deng et al. [4] em 2014 extenderam a técnica de LDA à Análise de Classificação Linear (Linear Ranks Analysis - LRA) considerando a ordem de classificação dos centroides de cada classe num subspaço projetado. Sob uma restrição na ordem de classificação das classes, são propostos dois critérios: 1) minimização do erro de classificação com o pressuposto de que cada classe é homogénea com distribuição Gaussiana; 2) maximização da soma (média) das distâncias entre todos os pares de classes vizinhas. Ambos os critérios podem ser resolvidos de forma eficiente com otimização num subespaço unidimensional.

[1] Nesta dissertação, usamos siglas em inglês para uma mais rápida identificação da metodologia em causa.

Outros modelos têm sido sugeridos para a classificação de classes ordinais, principalmente modelos de Máquinas de Suporte Vetorial (SVMs), Redes Neuronais ou Árvores de Decisão. E. Frank e M. Hall [5] introduziram um simples processo de exploração de classes ordinais usando classificadores binários convencionais. Em 1980, P. McCullagh [15] propôs um modelo de regressão que incluiu informações ordinais dos dados, eliminando a necessidade de atribuição de rótulos às classes. Uma extensão deste trabalho é feita em 2001 por G. Tutz [18] através da generalização do modelo aditivo de Hastie e Tibshirani [6] incorporando termos não paramétricos. Shashua e Levin [16] introduziram uma formulação generalizada para o método de SVMs para classes ordinais. Mais recentemente, Kotsiantis [13] propôs uma técnica de classificação em cascata abrangendo um classificador de Árvores de Decisão e um algoritmo para um modelo de árvores. Herbrich et al. [7] aplicaram o Princípio da Minimização do Risco Estrutural natural para SVMs de Vapnik [19] para obter um novo esquema de aprendizagem baseado numa grande margem para a tarefa de regressão ordinal. Jianlin Cheng et al. [2] apresentaram uma adaptação às tradicionais Redes Neuronais para aprender categorias ordinais, através de uma generalização do método do percetrão. Em 2005, Joaquim Costa e Jaime Cardoso [11] introduziram um novo modelo de Redes Neuronais feedforward, para problemas de classificação de multiclasses, onde as classes são ordenadas. Os mesmos autores em 2007 [8] apresentaram um novo paradigma de aprendizagem de máquinas de vetores de suporte especificamente para classes ordinais. A técnica reduz o problema de classificar classes ordenadas para um problema padrão de duas classes. O método introduzido é mapeado em SVMs e Redes Reuronais. Em 2008, J. Costa et al. [9] no seguimento de trabalhos anteriores, introduziram duas aproximações para SVMs e Redes Neuronais feedforward, uma paramétrica, e uma não paramétrica e também uma nova medida de performance para classes ordinais, nomeadamente o coeficiente r_{int} utilizado mais à frente neste capítulo. Posteriormente, em 2010, J. Costa et al. [12] elaboraram uma nova metodologia para SVMs baseada no paradigma unimodal com o esquema All-at-Once para a classificação ordinal.

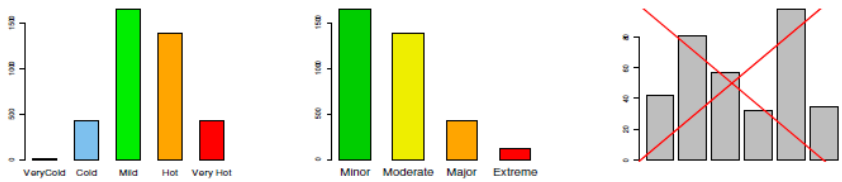


Figura 2.1: Paradigma unimodal

2.2 Modelo Tradicional

A Análise Discriminante Linear é uma técnica que se baseia em modelos para as funções de probabilidade.

Seja $f_k(x)$ a função densidade gaussiana multivariada da classe k com média $\mu(k) \in \mathcal{R}^p$ e matriz de variâncias-covariâncias $\Sigma_k \in \mathcal{R}^{p\times p}$, com a restrição $\Sigma_k = \Sigma, \forall k = 1, ..., K$.

É pretendido maximizar (2.1.1) ou de forma equivalente

$$log(\pi_k) + log(f_k(x)) = log(\pi(k)) - \frac{1}{2}(x - \mu(k))^t\Sigma^{-1}(x - \mu(k))$$

uma vez que o denominador é sempre positivo.

Então, para estimarmos os parâmetros do modelo podemos utilizar, por exemplo, o método de máxima verosimilhança e obtemos o seguinte resultado:

$$\hat{\pi}(k) = \frac{n_k}{n}$$

$$\hat{\mu}(k) = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}$$

$$\hat{\Sigma} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki} - \bar{\mu}(k))(x_{ki} - \bar{\mu}(k))^t$$

Assim, para compararmos duas classes, k e l, é suficiente considerar

$$\begin{aligned} log(\frac{P(C_k|x)}{P(C_l|x)}) &= log(\frac{f_k(x)}{f_l(x)}) + log(\frac{\pi(k)}{\pi(l)}) \\ &= \underbrace{log(\frac{\pi(k)}{\pi(l)}) - \frac{1}{2}(\mu(k) - \mu(l))^t\Sigma^{-1}(\mu(k) - \mu(l))}_{c_0} + x^t\Sigma^{-1}(\mu(k) - \mu(l)) \\ &= c_0 + x^t\Sigma^{-1}(\mu(k) - \mu(l)) \quad , c_0 \in \mathcal{R} \end{aligned}$$

que é, uma equação linear em x.

2.3 Modelo Inovador

Assumindo os pressupostos usuais da Análise Discriminante Linear, onde $f_k(x)$ é a função densidade gaussiana multivariada com média $\mu(k) \in \mathcal{R}^p$ e matriz de variâncias-covariâncias $\Sigma \in \mathcal{R}^{p\times p}$, independente das classes, uma condição suficiente para as probabilidades à posteriori seguirem uma distribuição unimodal é que a função $g = g(k)$ definida por

$$g(k) = \pi_k f_k(x)$$

tenha sempre a segunda derivada negativa, uma vez que o denominador em (2.1.1) é sempre positivo. Note-se que, embora k seja um inteiro, estamos a considerá-lo aqui como um real, para podermos efetuar a derivada. Nesse sentido, a condição da derivada ser negativa, satisfaz os nossos objetivos de unimodalidade, incluindo para os valores inteiros de k.

A função g é dada então por

$$g(k) = \pi(k)\frac{1}{|\Sigma|^{1/2}(2\pi)^{p/2}}e^{-\frac{1}{2}(x-\mu(k))^t\Sigma^{-1}(x-\mu(k))} \tag{2.3.1}$$

onde usamos a notação $\pi(k)$ e $\mu(k)$ para mostrar explicitamente que os parâmetros dependem da classe.

Agora impomos duas condições. A primeira é que a média $\mu(k)$ de cada classe k é uma função afim. Escrevemos

$$\mu(k) = (a_1 + b_1k, \ldots, a_p + b_pk) = a + bk, \qquad k = 1, \ldots, K$$

para vetores de valores reais $a = (a_1, \ldots, a_p)$ e $b = (b_1, \ldots, b_p)$ em R^p. Isto significa que estamos a assumir que as médias das K classes se encontram sobre uma reta. A segunda condição é que as probibilidades à priori sejam escritas de uma das seguintes formas:

1. Função Exponencial com um parâmetro:

   $$\pi(k) = ce^{-(k-\alpha)^2}, \qquad k = 1, \ldots, K$$

   onde a condição $\sum_{k=1}^{K}\pi(k) = 1$, implica que, $c = \dfrac{1}{\sum_{i=1}^{K}e^{-(i-\alpha)^2}}$.

2. Função Quadrática com dois parâmetros:

   $$\pi(k) = c_1 + c_2k + c_3k^2, \qquad k = 1, \ldots, K$$

   onde de forma análoga a condição $\sum_{k=1}^{K}\pi(k) = 1$, implica que, $c_1 = \frac{1}{K} - c_2\frac{K+1}{2} - c_3\frac{(K+1)(2K+1)}{6}$.

De forma a garantir a condição de unimodalidade começamos por aplicar a transformação logarítmica à função g (2.3.1)

$$(ln \circ g)(k) = \ln(\pi(k)) - \ln(|\Sigma|^{1/2}(2\pi)^{p/2}) - \frac{1}{2}(x-\mu(k))^t\Sigma^{-1}(x-\mu(k))$$

Usando regras de cálculo diferencial matricial [14], temos

$$(ln \circ g)'(k) = \frac{\pi'(k)}{\pi(k)} + (x-\mu(k))^t\Sigma^{-1}\mu'(k)$$

e portanto

$$(ln \circ g)''(k) = \underbrace{-\mu'(k)^t \Sigma^{-1} \mu'(k)}_{(a)} + \underbrace{(x - \mu(k))^t \Sigma^{-1} \mu''(k)}_{(b)} + \underbrace{\frac{\pi''(k)\pi(k) - \pi'(k)^2}{\pi(k)^2}}_{(c)}$$

Dado que Σ^{-1} é definida positiva[2], resulta que $(a) < 0$. Pela nossa primeira condição imposta $(b) = 0$. Resta explorarmos as duas hipóteses apresentadas na segunda condição para garantirmos a unimodalidade.

Para a função exponencial temos

$$\pi'(k) = (-2k + 2\alpha)ce^{-(k-\alpha)^2}$$

e

$$\pi''(k) = -2ce^{-(k-\alpha)^2} + (-2k + 2\alpha)^2 ce^{-(k-\alpha)^2}$$

Posto isto

$$(c) = \frac{\big[-2c^2e^{-2(k-\alpha)^2} + (-2k + 2\alpha)^2 c^2 e^{-2(k-\alpha)^2}\big] - (-2k + 2\alpha)^2 c^2 e^{-2(k-\alpha)^2}}{c^2 e^{-2(k-\alpha)^2}} = -2 < 0$$

E portanto ficam garantidas as condições de unimodalidade para este caso.

Para a função quadrática temos

$$\pi'(k) = c_2 + 2c_3k \; ; \; \pi''(k) = 2c_3$$

Portanto

$$(c) = \frac{2c_3(c1 + c_2k + c_3k^2) - (c_2 + 2c_3k)^2}{(c_1 + c_2k + c_3k^2)^2}$$

Ou seja, neste caso se garantirmos que

$$\begin{aligned} & 2c_3(c_1 + c_2k + c_3k^2) - (c_2 + 2c_3k)^2 \leq 0, \forall k \\ \Leftrightarrow\ & [3] - 2c_3(\frac{1}{K} - c_2\frac{K+1}{2} - c_3\frac{(K+1)(2K+1)}{6} - \frac{c_2^2}{2c_3} + \frac{c_2^2}{4c_3}) \geq 0 \\ \Leftrightarrow\ & -2\frac{c_3}{K} + c_2c_3(K+1) + c_3^2\frac{(K+1)(2K+1)}{3} + \frac{c_2^2}{2} \geqslant 0 \end{aligned}$$

é também satisfeita a condição $(\ln \circ g)''(k) < 0, \forall k$.

Queremos realçar que as condições impostas aqui, quer à posição das médias sobre uma reta, quer às expressões para as probabilidades à priori, são demasiado restritivas. Eventualmente existirão outras condições mais suaves que garantam a unimodalidade pretendida. Todavia, no âmbito desta tese de mestrado, não foram consideradas outras condições.

[2] **Demonstração:** Dado que Σ é definida positiva então Σ é invertível. Se definirmos $y = \Sigma x$ então $y^t\Sigma^{-1}y = x^t\Sigma^t\Sigma^{-1}\Sigma x = x^t\Sigma x > 0$

[3] O termo de maior grau da primeira inequação é sempre negativo portanto se a inequação é respeitada em $k = -\frac{c_2}{2c_3}$ também o será $\forall k$.

Garantidas as condições de unimodalidade, deduzimos a fórmula para os estimadores de máxima verosimilhança para os parâmetros desconhecidos no modelo discriminante unimodal com a função exponencial, nomeadamente α, a, b e Σ.

A função de máxima verosimilhança é dada por

$$\begin{aligned} L(\alpha, a, b, \Sigma; x_1, x_2, ..., x_n) &= f(x_1; \alpha, a, b, \Sigma) \times ... \times f(x_n; \alpha, a, b, \Sigma) \\ &= \prod_{k=1}^{K} \prod_{i=1}^{n_k} \pi_k \frac{1}{|\Sigma|^{1/2}(2\pi)^{p/2}} e^{-\frac{1}{2}(x_{ki}-\mu(k))^t \Sigma^{-1}(x_{ki}-\mu(k))} \end{aligned}$$

Se aplicarmos o logaritmo temos

$$\begin{aligned} v = \log(L) &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ln(\pi_k) - \frac{\ln(|\Sigma|)}{2} - \frac{p}{2}\ln(2\pi) - \frac{1}{2}(x_{ki} - a - bk)^t \Sigma^{-1}(x_{ki} - a - bk) \\ &= -\frac{n}{2}\ln(|\Sigma|) - n\frac{p}{2}\ln(2\pi) + \sum_{k=1}^{K} n_k \ln(\pi_k) - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki} - a - bk)^t \Sigma^{-1}(x_{ki} - a - bk) \end{aligned} \tag{2.3.2}$$

Novamente com o recurso ao cálculo diferencial matricial, para α obtemos

$$\frac{dv}{d\alpha} = \sum_{k=1}^{K} -2n_k(k - \alpha)$$

$$\begin{aligned} \frac{dv}{d\alpha} = 0 &\Leftrightarrow \sum_{k=1}^{K} -2n_k k + 2n_k \alpha = 0 \\ &\Leftrightarrow n\alpha = \sum_{k=1}^{K} n_k k \\ &\therefore \ \widehat{\alpha} = \frac{1}{n}\sum_{k=1}^{K} k n_k \end{aligned}$$

que é uma média pesada dos índices das classes $k = 1, \ldots, K$ da forma

$$\widehat{\alpha} = \sum_{k=1}^{K} \zeta_k k \qquad \text{com} \quad \sum_{k=1}^{K} \zeta_k = 1.$$

De igual forma, para obtermos o estimador de máxima verosimilhança de a e b começamos por

$$\begin{aligned} \frac{dv}{da} &= -\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki} - a - bk)^t[\Sigma^{-1} + (\Sigma^{-1})^t] \times (-1) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki} - a - bk)^t \Sigma^{-1} \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\Sigma^{-1}x_{ki} - n\Sigma^{-1}a - \sum_{k=1}^{K} n_k k \Sigma^{-1} b \end{aligned}$$

$$\frac{dv}{db} = -\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki} - a - bk)^t \times 2\Sigma^{-1}(-k)$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{n_k} k\Sigma^{-1}x_{ki} - \sum_{k=1}^{K} kn_k\Sigma^{-1}a - \sum_{k=1}^{K} k^2 n_k\Sigma^{-1}b$$

E seguidamente

$$\frac{dv}{da} = 0 \Leftrightarrow n\Sigma^{-1}a = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\Sigma^{-1}x_{ki} - \sum_{k=1}^{K} n_k k\Sigma^{-1}b$$
$$\Leftrightarrow a = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} x_{ki} - \frac{1}{n}\sum_{k=1}^{K} n_k kb$$

$$\frac{dv}{db} = 0 \Leftrightarrow \sum_{k=1}^{K} k^2 n_k\Sigma^{-1}b + \sum_{k=1}^{K} kn_k\Sigma^{-1}a = \sum_{k=1}^{K}\sum_{i=1}^{n_k} k\Sigma^{-1}x_{ki}$$
$$\Leftrightarrow \sum_{k=1}^{K} k^2 n_k\Sigma^{-1}b + \sum_{k=1}^{K} kn_k\Sigma^{-1}\left(\frac{1}{n}\sum_{s=1}^{K}\sum_{i=1}^{n_k} x_{si} - \frac{1}{n}\sum_{s=1}^{K} n_s sb\right) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} k\Sigma^{-1}x_{ki}$$
$$\Leftrightarrow \sum_{k=1}^{K} k^2 n_k b - \frac{1}{n}\sum_{k=1}^{K} kn_k\sum_{s=1}^{K} n_s sb = \sum_{k=1}^{K}\sum_{i=1}^{n_k} kx_{ki} - \frac{1}{n}\sum_{k=1}^{K} kn_k\sum_{s=1}^{K}\sum_{i=1}^{n_k} x_{si}$$
$$\Leftrightarrow \left[\sum_{k=1}^{K} k^2 n_k - \frac{1}{n}\left(\sum_{k=1}^{K} kn_k\right)^2\right] b = \sum_{k=1}^{K}(k - \frac{1}{n}\sum_{s=1}^{K} sn_s)\sum_{j=1}^{K} x_{kj}$$

$$\therefore \widehat{b} = \left(\sum_{s=1}^{K} s^2 n_s - \frac{1}{n}(\sum_{s=1}^{K} sn_s)^2\right)^{-1}\sum_{k=1}^{K}\left(k - \frac{1}{n}\sum_{s=1}^{K} sn_s\right)\sum_{j=1}^{n_k} x_{kj}. \quad (2.3.3)$$

Esta fórmula pode ser escrita na forma

$$\widehat{b} = \sum_{k=1}^{K}\gamma_k\overline{x}_k \quad \text{com} \quad \sum_{k=1}^{K}\gamma_k = 0$$

onde

$$\gamma_k = \left(\sum_{s=1}^{K} s^2 n_s - \frac{1}{n}\left(\sum_{s=1}^{K} sn_s\right)^2\right)^{-1} n_k\left(k - \frac{1}{n}\sum_{s=1}^{K} sn_s\right).$$

Daqui resulta que

$$a = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} x_{ki} - \frac{1}{n}\sum_{k=1}^{K} n_k k\frac{\sum_{m=1}^{K}\left(m - \frac{1}{n}\sum_{s=1}^{K} sn_s\right)}{\sum_{s=1}^{K} s^2 n_s - \frac{1}{n}(\sum_{s=1}^{K} sn_s)^2}\sum_{j=1}^{n_k} x_{kj}$$
$$\therefore \widehat{a} = \frac{1}{n}\sum_{k=1}^{K}\left(1 - \frac{nk(\sum_{s=1}^{K} sn_s) - (\sum_{s=1}^{K} sn_s)^2}{n\sum_{s=1}^{K} s^2 n_s - (\sum_{s=1}^{K} sn_s)^2}\right)\sum_{j=1}^{n_k} x_{kj}$$

que, por sua vez, pode ser expresso como uma combinação linear das médias amostrais das classes

$$\widehat{a} = \sum_{k=1}^{K} w_k\overline{x}_k \quad \text{com} \quad \sum_{k=1}^{K} w_k = 1.$$

onde

$$w_k = \frac{n_k}{n}(1 - \frac{nk(\sum_{s=1}^{K} sn_s) - (\sum_{s=1}^{K} sn_s)^2}{n\sum_{s=1}^{K} s^2 n_s - (\sum_{s=1}^{K} sn_s)^2}).$$

Ou seja, a reta que contém as médias no nosso modelo passa pelo ponto médio (ponderado) das médias amostrais. Uma espécie de "centro de gravidade".

O estimador de máxima verosimilhança para Σ é conseguido de forma análoga fazendo

$$\begin{aligned}
\frac{dv}{d\Sigma^{-1}} &= -\left(-\frac{n}{2}\ln(|\Sigma^{-1}|)\right)' - \frac{1}{2}\left[2\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}-a-bk)(x_{ki}-a-bk)^t\right. \\
&\quad \left. - \sum_{k=1}^{K}\sum_{i=1}^{n_k} diag\left((x_{ki}-a-bk)(x_{ki}-a-bk)^t\right)\right] \\
&= \frac{n}{2}(2\Sigma - diag[\Sigma]) - \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}-a-bk)(x_{ki}-a-bk)^t \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k} diag\left((x_{ki}-a-bk)(x_{ki}-a-bk)^t\right) \\
&= n\Sigma - \frac{n}{2}diag[\Sigma] - \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}-a-bk)(x_{ki}-a-bk)^t \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k} diag\left((x_{ki}-a-bk)(x_{ki}-a-bk)^t\right)
\end{aligned}$$

$$\begin{aligned}
\frac{dv}{d\Sigma^{-1}} = 0 \text{ [4]} &\Leftrightarrow n\Sigma - \frac{n}{2}diag[\Sigma] - \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}-a-bk)(x_{ki}-a-bk)^t \\
&\quad + \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k} diag\left((x_{ki}-a-bk)(x_{ki}-a-bk)^t\right) = 0 \\
&\therefore \widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{K}\sum_{\ell=1}^{n_k}(x_{k\ell}-\mu_k)(x_{k\ell}-\mu_k)^t. \qquad (2.3.4)
\end{aligned}$$

que é um valor nosso familiar.

Para a primeira função de probabilidades à priori também estimámos α pelo método dos mínimos quadrados, porém aqui apenas é possível obter o estimador com a utilização de métodos numéricos. Nós utilizámos o método de Newton, com função objectivo dada por

$$\min_{\alpha} \sum_{i=1}^{K}(\frac{n_i}{n} - \frac{e^{-(i-\alpha)^2}}{\sum_{l=1}^{K} e^{-(l-\alpha)^2}})^2. \qquad (2.3.5)$$

Isto é, pretendemos o valor de α que melhor aproxima as probabilidades à priori deste modelo das frequências das várias classes, no sentido dos mínimos quadrados.

[4] Pelo Lema 3.2.3 de [1], temos a garantia que é possível obter o estimador de Σ a partir da diferenciação em ordem a $Sigma^{-1}$.

Para a segunda função de probabilidades à priori estimámos c_2 e c_3 através de métodos de otimização e também pelo método dos mínimos quadrados, pois aqui existem restrições de desigualdades: uma para implicar que $(\ln \circ g)''(k) < 0$, e três para não gerarmos probabilidades negativas.

Portanto, neste caso pretendeu-se

$$\min_{c_2,c_3} \sum_{i=1}^{K} (\frac{n_i}{n} - \frac{1}{K} + c_2 \frac{K+1}{2} + c_3 \frac{(K+1)(2K+1)}{6} - c_2 i - c_3 i^2)^2 \tag{2.3.6}$$

$$, s.a. \begin{cases} \frac{1}{K} - c_2 \frac{K+1}{2} - c_3 \frac{(K+1)(2K+1)}{6} + c_2 + c_3 \geq 0 \\ -c_3 \geq 0 \\ \frac{1}{K} - c_2 \frac{K+1}{2} - c_3 \frac{(K+1)(2K+1)}{6} + c_2 K + c_3 K^2 \geq 0 \\ c_2 c_3 (K+1) + c_3^2 \frac{(K+1)(2K+1)}{3} + \frac{c_2^2}{2} - 2\frac{c_3}{K} \geq 0 \end{cases}$$

2.4 Resultados

2.4.1 Detalhes de implementação

Os nossos programas foram criados em linguagem R na versão 3.3.2. No método de otimização usámos o algoritmo *cobyla* da biblioteca *nloptr*, onde o ponto inicial é considerado num dos parâmetros da nossa função e escolhido posteriormente como o melhor dentro de uma grelha de pontos definida pelo utilizador.

Este problema contém uma função objetivo não linear e restrições também não lineares. Dada esta complexidade a solução que considerámos mais ajustada foi a utilização de um algorítmo de procura blindada[5] pois as alternativas testadas em bibliotecas do R já existentes mostraram-se incapazes de solucionar a questão. Nestas alternativas foram testados algoritmos como Particle Swarm Optimization (PSO), Differential Evolution Optimization, que se enquadram numa procura baseada na população (têm a vantagem do utilizador não necessitar indicar um ponto inicial). Porém, a nosso ver, a implementação existente no R não está ainda muito robusta.

Um exemplo que confirma esta consideração pertence ao método PSO e pode ser testado com o seguinte código:

```
library(mopsocd)
```

[5]Nomenclatura em conformidade com [3]

```
f<-function(x){
return(x[1]^2+x[2]^2)
}

gn <- function(x){
g1 <- -x[2]+10 <= 0.0
g2 <- x[1] <= 0.0
return(c(g1,g2))
}

coef=mopsocd(f,gn,varcnt=2,fncnt=1,opt=0,
lowerbound = c(-100,-100),upperbound = c(100,100))

#Outputs
coef[[1]][1] #x1=-2.989822e-12
coef[[1]][2] #x2=3.860841e-12
coef[[2]] #f(x1,x2)=2.384513e-23
```

Neste exemplo pretendia-se

$$\min_{x_1,x_2} x_1^2 + x_2^2, \quad s.a. \begin{cases} x_2 \geq 10 \\ x_1 \leq 0 \end{cases}$$

os parâmetros têm os seguintes significados; *varcnt*: o nº de restrições, *fncnt*: o nº de funções objetivo, *opt*: 0 minimização; 1 maximização e *lowerbound, upperbound* o domínio de procura. Porém as soluções obtidas pelo algorítmo foram aproximadamente 0 para as duas variáveis. Outros algorítmos foram testados e desconsiderados por razões semelhantes a esta.

Para avaliar os nossos métodos, separámos aleatoriamente os dados, em cada classe, em duas partes:

- 70% para treino;
- 30% para teste.

Este processo foi repetido 10 vezes e a média do erro e o desvio-padrão foram registadas. Neste esquema, foram comparados os resultados obtidos pelas seguintes metodologias:

- Análise discriminante linear conventional (cLDA);
- Análise discriminante linear unimodal com função de probabilidades à priori exponencial onde os parâmetros foram estimados por máxima verosimilhança (mleUDA);

- Análise discriminante linear unimodal com função de probabilidades à priori exponencial onde os parâmetros foram estimados pelos mínimos quadrados (lseUDA);
- Análise discriminante linear unimodal com função de probabilidades à priori quadrática onde os parâmetros foram estimados pelos mínimos quadrados (lsqUDA);
- O nosso modelo de análise discriminante linear unimodal com a função de probabilidades à priori estimada da forma usual, isto é, dada pelas frequências relativas dos dados (freqUDA). Com esta função de probabilidades não conseguimos assegurar a condição de unimodalidade para as probabilidades à posteriori, contudo é mais provável que aconteça do que no modelo de análise discriminante linear tradicional.

2.4.2 Medidas de Performance

Para avaliarmos e compararmos os métodos acima referidos, temos que tomar em conta que a variável que estamos a tentar prever é ordinal e portanto os erros não têm todos a mesma importância. Por exemplo, se prevermos a classe 2 para uma observação que pertence à classe 1 este erro não é tão mau como se tivessemos previsto a classe 5. Por esta razão temos que escolher medidas adequadas de erro. Por isso, para além da medida mais comum (MRE), que dá a mesma importância a todos os erros, usámos as outras que se seguem:

- **E**rro **M**édio **R**elativo (MRE): percentagem de elementos que foram mal classificados.

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n}\zeta_i, \qquad \text{onde } \zeta_i = \begin{cases} 1 & \text{se } \widehat{C}(i) \neq C(i) \text{ (classe prevista diferente da classe real)} \\ 0 & \text{outros casos} \end{cases}$$

- **E**rro **M**édio **A**bsoluto (MAE).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\,\widehat{C}(i) - C(i)|$$

- Coeficiente de correlação de Kendall entre a classe prevista $\widehat{C}$ e a classe real C. Esta na nossa opinião é uma boa medida de erro para classificadores de dados ordinais. Como complemento também incluímos uma outra medida, introduzida em [9, 10] que é semelhante a esta e que se apresenta no item seguinte.
- Coeficiente r_{int}. Dada uma tabela de contingencia do tipo

C / $\widehat{C}$	1	2	...	K	Total
1	n_{11}	n_{12}	...	n_{1K}	$n_{1\bullet}$
2	n_{21}	n_{22}	...	n_{2K}	$n_{2\bullet}$
...	...	...		...	...
K	n_{K1}	n_{K2}	...	n_{KK}	$n_{K\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet K}$	n

O cálculo desde coeficiente é dado por

$$r_{int} = -1 + 2\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{i'=i}^{K}\sum_{j'=j}^{K} n_{ij}n_{i'j'} - n}{\sqrt{(\sum_{i=1}^{K}\sum_{j=i}^{K} n_{i\bullet}n_{j\bullet} - n)(\sum_{i=1}^{K}\sum_{j=i}^{K} n_{\bullet i}n_{\bullet j} - n)}}$$

2.4.3 Aplicação da metodologia

Nesta subsecção iremos apresentar a performance dos nossos métodos, comparando-os com o método tradicional em quatro conjuntos de dados reais.

1. O primeiro conjunto de dados contém 1000 exemplos de avaliações anónimas, de alunos aos seus professores, realizadas no final de um MBA. Antes de receberem as notas finais os alunos foram convidados a avaliar os seus professores em quatro atributos tal como, apresentação oral e contribuição para o seu conhecimento profissional.

   A única variável de saída era a avaliação total do desempenho do professor.

   Variáveis: 4 variáveis de entrada ordinais e uma de saída também ordinal (5 classes).

Métodos / Medidas	MRE	MAE	Kendall	r_{int}
cLDA	0.401 (0.025)	0.433 (0.027)	0.629 (0.030)	0.679 (0.019)
freqUDA	0.391 (0.023)	0.424 (0.029)	0.640 (0.031)	0.685 (0.022)
mleUDA	0.424 (0.011)	0.455 (0.013)	0.615 (0.023)	0.69 (0.013)
lseUDA	0.431 (0.021)	0.464 (0.024)	0.605 (0.031)	0.685 (0.017)
lsqUDA	0.392 (0.024)	0.430 (0.027)	0.650 (0.023)	0.685 (0.016)

Tabela 2.1: Resultados para o conjunto de dados LEV.

Para este conjunto de dados os métodos freqUDA e lsqUDA superaram o método convencional (cUDA) em todos os critérios. O método freqUDA obteve um melhor resultado na medida MAE porém o método lsqUDA foi superior na medida de Kendall. No que diz respeito à variância o método lsqUDA consegiu ser um pouco melhor do que o método freqUDA.

2. Este conjunto de dados pretende associar as características das placas de circuito do processador com o desempenho do processamento das placas. Existem 209 observações.

   Variáveis: 6 variáveis de entrada e uma ordinal de saída (10 classes).

Métodos / Medidas	MRE	MAE	Kendall	r_{int}
cLDA	0.329 (0.032)	0.455 (0.040)	0.775 (0.035)	0.829 (0.021)
freqUDA	0.318 (0.033)	0.423 (0.063)	0.768 (0.034)	0.829 (0.024)
mleUDA	0.763 (0.040)	1.069 (0.070)	0.760 (0.031)	0.822 (0.020)
lseUDA	0.397 (0.037)	0.773 (0.051)	0.759 (0.071)	0.817 (0.046)
lsqUDA	0.356 (0.043)	0.442 (0.069)	0.802 (0.041)	0.836 (0.034)

Tabela 2.2: Resultados para o conjunto de dados CPU.

Neste conjunto de dados o método lsqUDA superou o método convencional em todos os critérios exceto no primeiro. Contudo, tal como indicado anteriormente, a primeira medida não é a mais indicada para este tipo de estudo e portanto o resultado é completamente positivo. O método freqUDA apenas foi inferior à metodologia usual na medida de kendall. A respeito da variância o método freqUDA consegiu ser um pouco melhor do que o método lsqUDA.

3. Esta base de dados contém informações adquiridas nos censos na área de Boston, EUA. Existem 506 observações.

   Variáveis: 13 variáveis de entrada e uma ordinal de saída (5 classes).

Métodos / Medidas	MRE	MAE	Kendall	r_{int}
cLDA	0.316 (0.038)	0.350 (0.040)	0.720 (0.035)	0.751 (0.027)
freqUDA	0.303 (0.031)	0.329 (0.036)	0.723 (0.037)	0.761 (0.024)
mleUDA	0.372 (0.037)	0.401 (0.041)	0.695 (0.042)	0.742 (0.029)
lseUDA	0.361 (0.026)	0.406 (0.027)	0.689 (0.025)	0.744 (0.018)
lsqUDA	0.357 (0.032)	0.381 (0.033)	0.741 (0.028)	0.751 (0.019)

Tabela 2.3: Resultados para o conjunto de dados Housing.

Aqui o método freqUDA superou o método convencional em todos os critérios. A respeito dos critérios mais adequados para o estudo o método lsqUDA apenas não superou o método usual na medida MAE, ainda assim teve um bom desempenho nas medidas mais importantes (kendall e r_{int}). Quanto à variância o método lsqUDA consegiu ser um pouco melhor do que o método freqUDA.

4. Esta base de dados contém 488 perfis de candidados a um emprego. Os resultados atribuidos por uma empresa de recrutamento a certos atributos determinam a variável de saída correspondente à avaliação final de cada candidato.

   Variáveis: 4 variáveis de entrada e uma ordinal de saída (9 classes).

Métodos / Medidas	MRE	MAE	Kendall	r_{int}
cLDA	0.301 (0.034)	0.317 (0.037)	0.867 (0.017)	0.842 (0.019)
freqUDA	0.297 (0.036)	0.316 (0.038)	0.869 (0.017)	0.843 (0.018)
mleUDA	0.410 (0.035)	0.449 (0.039)	0.840 (0.019)	0.814 (0.019)
lseUDA	0.386 (0.022)	0.431 (0.029)	0.838 (0.022)	0.816 (0.020)
lsqUDA	0.295 (0.036)	0.308 (0.034)	0.871 (0.014)	0.844 (0.016)

Tabela 2.4: Resultados para o conjunto de dados ESL.

Novamente os métodos freqUDA e lsqUDA superaram o método convencional em todos os critérios. Aqui também o método lsqUDA conseguiu ser um pouco melhor que o método freqUDA em todas as medidas, mesmo no que diz respeito à variância.

A análise dos resultados nestes quatro conjuntos de dados sugere que o nosso modelo

de análise discriminante unimodal é uma boa alternativa à tradicional análise discriminante linear. Não só o modelo das respostas é mais apropriado para este tipo de classes como também as medidas de erro obtidas são competitivas com o método tradicional. De facto, no que diz respeito às quatro medidas MER, MAE, kendall's tau-b and r_{int}, temos sempre um dos nossos modelos com melhores resultados que o modelo convencional. Particularmente nas medidas que são mais apropriadas a este tipo de estudo, podemos ver que os nossos modelos são sempre melhores e em particular o modelo lsqUDA é o que se destaca mais. Este é um aspeto importante porque neste modelo podemos assegurar que as respostas respeitam a condição de unimodalidade, sem perder qualidade no desempenho quando comparado com o da análise discriminante usual.

2.5 Discussão

Neste estudo apresentámos uma nova abordagem à análise discriminante linear em dados ordinais. A ideia principal é manter a ordinalidade das classes, impondo um modelo paramétrico para as probabilidades de saída.

Modelando a função de probabilidades à priori por uma função quadrática, produzimos um classificador mais simples e robusto, que se comparou favoravelmente com a metodologia convencional. O resultado comparativo não foi tão expressivo quando a função de probabidade à priori consistiu apenas das frequências relativas, mas mesmo nessa situação os resultados obtidos ainda foram razoáveis.

Para a função de probabilidades à priori exponencial os resultados não foram muito bons e uma causa disto poderá ser o facto destas funções só conterem um parâmetro de liberdade, α, e assim não serem suficientemente adaptáveis.

Outras direções para trabalhos futuros poderão incluir outras funções de probabilidades à priori com mais parâmetros de liberdade, como por exemplo, Pearson tipo III e Pearson tipo IV.

Função de distribuição Person tipo III:

$$f(x) = k\big(1 + \frac{x}{a}\big)^{\mu a} e^{-\mu x} \ , -a < x < \infty, \mu, a > -1$$

Função de distribuição Person tipo IV:

$$f(x) = k\big(1 + \frac{x^2}{a^2}\big)^{-m} e^{-\mu arctg(x/a)} \ , -\infty < x < \infty, \mu, a, m > 0$$

Parece-nos que estas funções podem melhorar ainda mais os resultados.

Capítulo 3

Clustering longitudinal

3.1 Introdução

Um conjunto de dados é considerado longitudinal se cada indivíduo é medido repetidamente ao longo do tempo, dando origem a um vetor de observações para cada indivíduo que tende a ter componentes correlacionadas. Isto apresenta um claro contraste com o estudo de dados transversais em que a informação conhecida para uma variável é única para cada indivíduo.

Dados longitudinais combinam elementos de dados multivariados e séries temporais. Contudo, diferem dos dados multivariados tradicionais ao incorporarem um padrão muito mais estruturado de interdependência entre as medições e diferem das clássicas séries temporais por consistirem de um grande número de séries curtas, em vez de séries longas.

Ao longo do último meio século, tem-se assistido a um progresso considerável no desenvolvimento de métodos estatísticos para análise de dados longitudinais [24, 32, 40, 49, 52]. Em particular, são bastantes os critérios que têm vindo a ser propostos para a escolha do ”melhor número de clusters”[22, 26, 34, 36, 44, 50] bem como métodos para o tratamento de dados em falta [34–36]. Aplicações desta teoria podem ser encontradas num largo domínio, desde a biomedicina até às ciências sociais e comportamentais [28, 30, 38, 45, 47, 48]. Também novas metodologias têm sido propostas [27, 42]. A. Ciampi et al. [27] consideraram o problema de clustering em dados dependentes no tempo. O modelo consiste de uma mistura de extensões de modelos lineares mistos (Extended Linear Mixed Models) e a estimação é feita por máxima verosimilhança, usando o método Expectation-Maximization (EM). Os dados são modelados por uma mistura de distribuições normais multivariadas e é considerada uma parametrização nas matrizes de variância-covariância através da decomposição espectral das matrizes. Antonello Maruotti et al. [42] proposeram um método que se baseia numa extensão

do algoritmo K-médias clássico, onde um modelo auto-regressivo de vetor multivariado (multivariate vector autoregressive model - MVAR) é adicionalmente assumido, para modelar a evolução dos centros dos clusters ao longo do tempo. Este método agrupa as observações longitudinais multivariadas com foco na evolução das partições ao longo do tempo e simultaneamente tem em consideração a heterogeneidade latente e a dependência do tempo. Através de uma configuração longitudinal univariada, o método tenta descrever a dinâmica dos processos modelando o valor atual do resultado como uma soma linear ponderada dos valores anteriores. A inferência do modelo é baseada no método do gradiente descendente.

Mesmo para a situação de uma única variável resposta, a visualização de todas as trajetórias pode tornar-se insuficiente para uma identificação eficaz dos padrões importantes, portanto em estudos longitudinais, o agrupamento das trajetórias e a obtenção de trajetórias médias podem ser facilmente de interesse.

Vários métodos foram construídos para extender a análise de agrupamento a dados longitudinais. A maioria deles pode ser agrupada em métodos não paramétricos ou métodos baseados em modelos [34, 39].

Os métodos de clustering não paramétricos integram a abordagem tradicional. Não existem suposições sobre como os dados foram gerados e trajetórias próximas são agrupadas de acordo com alguma medida de dissemelhança. Os principais ingredientes são a medida de dissemelhança, o algoritmo de agrupamento e o número de clusters a considerar [39]. Enquanto a medida de dissemelhança quantifica o carácter distintivo das trajetórias, o algoritmo de agrupamento visa optimizar um critério baseado na medida de dissemelhança. Um dos algoritmos de agrupamento mais simples e amplamente utilizado é o K-médias [37] e exemplos desta metodologia podem ser encontrados em [25, 41, 51].

Os métodos baseados em modelos assumem que os dados provêm de um número finito de populações e que, dentro de cada população, os dados podem ser modelados usando um modelo estatístico padrão. Por outras palavras, assume-se que os dados provêm de uma mistura de distribuições [33, 43, 46, 53]. Esta abordagem baseada em modelos é conhecida como clustering em modelos de misturas.

Neste estudo, introduzimos uma nova metodologia não paramétrica para agrupamento de dados longitudinais. As correlações entre as observações das trajetórias individuais são levadas em conta por matrizes de correlação pré-definidas com parâmetros que são estimados a partir dos dados. É considerada uma distância do tipo Mahalanobis com uma matriz de correlação pré-especificada e posteriormente o algoritmo das K-médias é aplicado. Um resultado útil é que o processo coincide com a aplicação do algoritmo K-Médias longitudinal,

acessível por exemplo na biblioteca *kml* e *kml3d* do R, usando a distância euclidiana em certas trajetórias transformadas, simplificando assim o processo para o utilizador.

3.2 Cálculo de distâncias

Num capítulo em que temos como objetivo agrupar trajectórias em classes homogéneas torna-se essencial ter bem presente a noção de distância. Estando as nossas trajetórias indexadas no tempo, explicamos agora como podemos calcular distâncias nos casos unidimensional e multidimensional.

- Caso unidimensional

  Considere-se um conjunto S de n observações onde cada observação é medida em t tempos diferentes. Para a observação i, denomina-se $x_i = (x_{i1}, x_{i2}, ..., x_{it})$ como uma trajetória e para calcular a distância entre duas trajetórias x_i e x_l o método mais simples passa por utilizar as fórmulas já existentes para distâncias entre dois individuos com t variáveis.

  Por exemplo, para o cálculo da distância Euclidiana entre duas trajetórias temos:

  $$d(x_i, x_l) = \sqrt{\sum_{j=1}^{t}(x_{ij} - x_{lj})^2}$$

  Este é o método que vigora na biblioteca *kml*.

- Caso multidimensional

  Considere-se para a trajetória i, no tempo j e variável p denotado como x_{ijp}. Então

  $$x_{i..} = \begin{pmatrix} x_{i.1} \\ x_{i.2} \\ ... \\ x_{i.p} \end{pmatrix} = \begin{pmatrix} x_{i11} & x_{i21} & ... & x_{it1} \\ x_{i12} & x_{i22} & ... & x_{it2} \\ ... & ... & & ... \\ x_{i1p} & x_{i2p} & ... & x_{itp} \end{pmatrix}$$

  Para definir a distância entre duas trajetórias conjuntas muitos métodos são possíveis. A biblioteca *kml3d* foca-se nos dois métodos seguintes.

  Seja $Dist$ uma função distância e $||.||$ uma norma. No primeiro método, de forma a calcular a distância d entre $x_{1..}$ e $x_{2..}$, para cada tempo j fixado definimos

  $$d_{j.}(x_{1j.}, x_{2j.}) = Dist(x_{1j.}, x_{2j.}).$$

O resultado dá um vetor

$$\left(\ d_{1.}(x_{11.}, x_{21.}), d_{2.}(x_{12.}, x_{22.}), ..., d_{t.}(x_{1t.}, x_{2t.}) \ \right)$$

Posteriormente combinamos estas t distâncias usando a função que algebricamente corresponde à norma $||.||$ do vetor de distâncias. Ou seja, a distância entre $x_{1..}$ e $x_{2..}$ é dada por

$$d(x_{1..}, x_{2..}) = || \left(\ d_{1.}(x_{11.}, x_{21.}), d_{2.}(x_{12.}, x_{22.}), ..., d_{t.}(x_{1t.}, x_{2t.}) \ \right) ||$$

O segundo método consiste em calcular a distância d' entre $x_{1..}$ e $x_{2..}$, para cada variável X. Definimos a distância como

$$d'_{.X}(x_{1.X}, x_{2.X}) = Dist(x_{1.X}, x_{2.X})$$

e de forma análoga, a distância é dada por

$$d'(x_{1..}, x_{2..}) = || \left(\ d_{.1}(x_{1.1}, x_{2.1}), d_{.2}(x_{1.2}, x_{2.2}), ..., d_{.p}(x_{1.p}, x_{2.p}) \ \right) ||$$

Escolhas diferentes para a norma $||.||$ e da distância $Dist$ podem conduzir à definição de um grande número de distâncias.

Refira-se que, no caso em que $||.||$ corresponde à norma-p usual e $Dist$ à distância de Minkowski com parâmetro q, escolhendo o método d ou d' obtemos o mesmo resultado. Assim, caso se utilize a distância Euclidiana ($q = 2$), ou a de Manhattan ($q = 1$) ou o Máximo ($q = +\infty$) a escolha do método (1 ou 2) é indiferente. De facto tem-se

$$\begin{aligned} d(x_{1..}, x_{2..}) &= \sqrt[q]{\sum_j (dj.(x_{1j.}, x_{2j.}))^q} = \sqrt[q]{\sum_j \left(\ \sqrt[q]{\textstyle\sum_X |x_{1jX} - x_{2jX}|^q} \ \right)^q} \\ &= \sqrt[q]{\sum_j \sum_X |x_{1jX} - x_{2jX}|^q} = \sqrt[q]{\sum_X \left(\ \sqrt[q]{\textstyle\sum_j |x_{1jX} - x_{2jX}|^q} \ \right)^q} \\ &= \sqrt[q]{\sum_X (d.X(x_{1.X}, x_{2.X}))^q} = d'(x_{1..}, x_{2..}) \end{aligned}$$

3.3 Método das K-Médias

Ao processo de particionamento de um grupo de dados num número menor de classes chamamos de Clustering. No nosso caso, o objetivo é atribuir a cada trajetória um grupo.

K-Médias é um dos vários métodos de agrupamento que visa encontrar os melhores agrupamentos que minimizem a distância das trajetórias, dentro do seu grupo e maximizam essas distâncias entre grupos diferentes. Um grupo fica definido pela sua trajetória média.

Primeiro é necessário indicarmos o número de grupos a atribuir aos nossos dados, normalmente denominado por k. Posteriormente devemos escolher os centros iniciais dos clusters, $\bar{x}_i, i \in 1, .., k$. É usual atribuitem-se trajetórias aleatórias como centros iniciais. O passo seguinte consiste em percorrer todas as trajetórias e atribuir-lhes o centro mais próximo, devendo este ser atualizado cada vez que uma trajetória é analizada. Este passo deve ser percorrido iterativamente até que nunhuma alteração se verifique.

Por outras palavras e em suma, o método das K-Médias pretende encontrar

$$\underset{c}{arg\,min} \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \bar{x}_i)$$

onde c_i é o conjunto de trajetórias que pertencem ao cluster i, ou seja,

$$c_i = \{j : d(x_j, \bar{x}_i) \leq d(x_j, \bar{x}_l), l \neq i, j = 1, ..., n\}$$

e os centros de cada cluster são calculados através da média usual

$$\bar{x}_i = \frac{1}{|c_i|} \sum_{j \in c_i} x_j, \forall i$$

3.4 Critérios para a escolha do nº de grupos

O início desta secção consiste da introdução de alguns conceitos base para a posterior definição dos critérios.

O **índice de Gower** [20] é um coeficiente genérico de dissemelhança que toma em consideração os vários tipos de natureza das variáveis. Suponha-se que os dados contêm p variáveis de natureza mista. Então a dissimilhança $d(i, j)$ entre os objetos i e j é definida como

$$d_{ij} = d(i, j) = \frac{\sum_{l=1}^{p} \delta_{ij;l} d_{ij,l}}{\sum_{l=1}^{p} \delta_{ij;l}},$$

sendo que $\delta_{ij;l}$ é uma variável dicotómica e é 0 quando pelo menos uma das medidas $x_{i;l}$ e $x_{j;l}$ da variável l estão em falta. Se a natureza da variável é quantitativa toma-se $d_{ij;l} = \frac{|x_{il} - x_{jl}|}{R_l}$, onde R_l é a amplitude da variável l. No caso em que l é uma variável nominal ou binária $d_{ij;l}$ é sempre 1 e $\delta_{ij;l}$ é 1 quando $x_{i;l} \neq x_{j;l}$ e 0 caso contrário.

Refira-se também que se $\delta_{ij;l}$ for sempre 0, $d(i, j)$ não pode ser calculado. Neste caso deve-se atribuir um valor convencional ou remover os objetos i e j.

Em suma, a **matriz de distâncias** Q **de Gower** corresponde a uma matriz da forma

$$Q = \begin{pmatrix} d_{11} & d_{21} & ... & d_{n1} \\ d_{21} & d_{22} & ... & d_{n2} \\ ... & ... & & ... \\ d_{n1} & d_{n2} & ... & d_{np} \end{pmatrix}$$

A distância de Gower é a distância utilizada em dados longitudinais sempre que existem dados em falta ou variáveis qualitativas.

A **matriz de dispersão** R [21] $\in \mathbb{R}^{p\times p}$ é dada por

$$R = \sum_{k=1}^{g}\sum_{l=1}^{n_k}(x_{kl} - \bar{x})(x_{kl} - \bar{x})'$$

onde $x_{kl} \in \mathbb{R}^p$ representa um vetor de observações da l-ésima trajetória do grupo k e $\bar{x} \in \mathbb{R}^p$ representa a trajetória média de todas as observações.

Em espaços euclidianos, a matriz R decompõe-se na soma $R = B + W$ [21], sendo que a **matriz de dispersão dentro dos grupos** W é dada por

$$W = \sum_{k=1}^{g}\sum_{l=1}^{n_k}(x_{kl} - \bar{x}_k)(x_{kl} - \bar{x}_k)'$$

onde $\bar{x}_k \in \mathbb{R}^p$ representa a média das observações dentro do grupo k, e a **matriz de dispersão entre grupos** B corresponde a

$$B = \sum_{k=1}^{g} n_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})'$$

Os critérios a serem apresentados nas secções que se seguem são os encorporados nas bibliotecas utilizadas, sendo estes os mais usados e com melhores desempenhos segundo a literatura [50].

3.4.1 Critério de Calinski Harabasz

Se representarmos o nº de indivíduos por n e o nº de variáveis por p, podemos representar os pontos $P_1, P_2, ..., P_n$ no espaço Euclidiano pela matriz:

$$X = \begin{pmatrix} P_{11} & P_{21} & ... & P_{n1} \\ P_{12} & P_{22} & ... & P_{n2} \\ ... & ... & & ... \\ P_{1p} & P_{2p} & ... & P_{np} \end{pmatrix}$$

onde linhas correspondem a variáveis e colunas a indivíduos. Depois disto, é possível construir a matriz de distâncias Q de Gower, essencial para o início do método, e a matriz de dispersão R. A distância ao quardrado de d_{ij} entre P_i e P_j será definida pela função

$$d_{ij}^2 = (X_{.i} - X_{.j})'(X_{.i} - X_{.j}) \qquad i, j = 1, 2, ..., n.$$

Um elemento importante na análise de clusters é a dispersão de um grupo, e para um grupo de n indivíduos ela é dada pela soma dos quadrados das distâncias entre cada ponto e o seu centroide. Esta soma é igual ao $Tr(R)$ e pode ser obtida da seguinte forma [23]:

$$Tr(R) = n^{-1}(d_{12}^2 + d_{13}^2 + ... + d_{n-1,n}^2) \tag{3.4.1}$$

Esta fórmula é muito útil pois evita o cálculo da matriz R e deve-se referir que é válida no espaço Euclidiano mesmo quando os eixos não são ortogonais.

Pretendem-se separar n indivíduos em k clusters, então temos $n-1$ intervalos entre os indivíduos e precisamos de escolher apenas $k-1$ desses intervalos. Isto perfaz ${}^{n-1}C_{k-1}$ possibilidades de separação. Se representarmos o nº de indivíduos em cada cluster por $n_1, n_2, ..., n_k$, podemos calcular a soma dos quadrados dentro de cada cluster (*Within Group Sum of Squares*), por aplicação de (3.4.1), e no final considerar a soma para os todos os clusters. A matriz R pode ser decomposta em $R = B + W$, sendo que

$$WGSS = Tr(W) = Tr(R_1) + Tr(R_2) + ... + Tr(R_k)$$

onde

$$Tr(R_g) = n_g^{-1}(d_{12}^2(g) + d_{13}^2(g) + ... + d_{n-1,n}^2(g))$$

com $d_{ij}(g)$ denotando a distância entre P_i e P_j do cluster g ($g = 1, 2, ..., k$).

Se k não é conhecido então este critério pode ser utilizado da seguinte maneira: primeiro utilizamos $k = 2$, depois $k = 3$, e assim sucessivamente. Em cada caso pretendemos não só encontrar o mínimo $WGSS$ mas também o máximo $BGSS = Tr(B)$ (soma dos quadrados entre os grupos - Between-Group Sum of Squares). O critério é dado por

$$CH = \frac{BGSS}{k-1} \div \frac{WGSS}{n-k} \tag{3.4.2}$$

Sendo este critério um informal indicador para o melhor nº de grupos. De facto este quociente é análogo à estatística F, associada a modelos de regressão.

Apesar de efetivamente não existir teoria estatística satisfatória para justificar o uso de CH (3.4.2), este tem algumas características matemáticas desejáveis e encorajadoras.

Seja $\bar{d}^2$ média das distâncias quadradas (existem $\frac{n(n-1)}{2}$ valores distintos de distâncias) e $\bar{d}^2(g)$ a média análoga para o cluster g; então, de (3.4.1)

$$TSS = n^{-1}\frac{n(n-1)}{2}\bar{d}^2 = \frac{1}{2}(n-1)\bar{d}^2$$

$$WGSS = Tr(R_1) + Tr(R_2) + ... + Tr(R_k) = \frac{1}{2}[(n_1-1)\bar{d}^2(1) + (n_2-1)\bar{d}^2(2) + ... + (n_k-1)\bar{d}^2(k)]$$

sendo que uma outra interessante forma de escrita é

$$\begin{aligned}
2 \times WGSS &= \bar{d}^2 \times (n - n_1 - n_2 - ... - n_k - k + k) + 2 \times WGSS \\
&= \bar{d}^2 \times (n - k - n_1 + 1 - n_2 + 1 - ... - n_k + 1) + 2 \times WGSS \\
&= \bar{d}^2 \times (n-k) - (n_1-1)\bar{d}^2 - ... - (n_k-1)\bar{d}^2 + 2 \times WGSS \\
&= \bar{d}^2 \times (n-k) - [(n_1-1)(\bar{d}^2 - \bar{d}^2(1)) - ... - (n_k-1)(\bar{d}^2 - \bar{d}^2(k))] \\
&= (n-k)(\bar{d}^2 - A_k)
\end{aligned}$$

onde

$$A_k = \frac{1}{n-k}[(n_1-1)(\bar{d}^2 - \bar{d}^2(1)) + (n_2-1)(\bar{d}^2 - \bar{d}^2(2)) + ... + (n_k-1)(\bar{d}^2 - \bar{d}^2(k))]$$

de onde resulta que

$$BGSS = TSS - WGSS = \frac{1}{2}[(k-1)\bar{d}^2 + (n-k)A_k]$$

Esta dedução permite reescrever CH da seguinte maneira:

$$CH = (3.4.2) = (\bar{d}^2 + \frac{n-k}{k-1}A_k) \div (\bar{d}^2 - A_k)$$

e aqui o caminho passa apenas pela maximização de A_k.

Algumas modificações a este critétio têm vindo a ser sugeridas [36], nomeadamente:

- Critério Calinski Harabasz 2, sugerido por Kryszczuk [36]

$$CH2 = \frac{BGSS}{n-k} \div \frac{WGSS}{n-1}$$

- Critério Calinski Harabasz 3, sugerido por Genolini [36]

$$CH3 = \frac{BGSS}{\sqrt{k-1}} \div \frac{WGSS}{n-k}$$

Para uma melhor compreensão do critério, incluímos de seguida dois exemplos elucidativos.

- Caso unidimensional

Exemplo 3.4.1.1. *Seja S o conjunto das 3 trajectórias y_1, y_2, y_3 seguintes, indexadas em 2 tempos:*

$$y_1 = (5, 2);\ y_2 = (-1, 0);\ y_3 = (9, 4)$$

Estas trajetórias estão claramente divididas em dois grupos $g_1 = \{y_1, y_3\}$ e $g_2 = \{y_2\}$.

Começamos por determinar as distâncias entre cada uma das trajetórias:

$$d_{12}^2 = 6^2 + 2^2 = 40;$$
$$d_{13}^2 = 4^2 + 2^2 = 20;$$
$$d_{23}^2 = 10^2 + 4^2 = 116$$

Seguindo o cálculo dos traços das matrizes:

$$Tr(R) = \frac{1}{3}(d_{12}^2 + d_{13}^2 + d_{23}^2) = \frac{176}{3};$$
$$Tr(W) = Tr(R_1) + Tr(R_2) = \frac{1}{2}d_{13}^2 + 0 = 10;$$
$$Tr(B) = \frac{176}{3} - 10 = \frac{146}{3}$$
$$\therefore CH = \frac{\frac{146}{3}}{2-1} \div \frac{10}{3-2} \simeq 4,867$$

- Caso multidimensional

Exemplo 3.4.1.2. *Seja S o conjunto das 3 trajectórias bivariadas y_1, y_2, y_3 seguintes, indexadas em 2 tempos:*

$$y_{1..} = \begin{pmatrix} 5 & 2 \\ 3 & -2 \end{pmatrix};\ y_{2..} = \begin{pmatrix} -1 & 0 \\ 9 & 0 \end{pmatrix};\ y_{3..} = \begin{pmatrix} 9 & 4 \\ 1 & 0 \end{pmatrix}$$

Estas trajetórias estão novamente divididas em dois grupos $g_1 = \{y_1, y_3\}$ e $g_2 = \{y_2\}$.

Para o cálculo das distâncias entre cada uma das trajetórias utilizou-se o primeiro método exposto anteriormente, com a norma Euclidiana:

$$d_{12}^2 = ||(\sqrt{6^2 + 6^2}, \sqrt{2^2 + 2^2})||^2 = 6^2 + 6^2 + 2^2 + 2^2 = 80;$$
$$d_{13}^2 = 4^2 + 2^2 + 2^2 + 2^2 = 28;$$
$$d_{23}^2 = 10^2 + 8^2 + 4^4 = 180$$

Seguindo o cálculo dos traços das matrizes:

$$Tr(R) = \frac{1}{3}(d_{12}^2 + d_{13}^2 + d_{23}^2) = 96;$$
$$Tr(W) = Tr(R_1) + Tr(R_2) = \frac{1}{2}d_{28}^2 + 0 = 14;$$
$$Tr(B) = 96 - 14 = 82$$
$$\therefore CH = \frac{82}{2-1} \div \frac{14}{3-2} \simeq 5,857$$

Estes resultados podem ser obtidos através do seguinte código (incluimos um alerta para um comando desatualizado):

```
v1 <- matrix(c(5,-1,9,2,0,4),ncol=2);v1
v2 <- matrix(c(3,9,1,-2,0,0),ncol=2);v2

raw <- clusterLongData3d(array(cbind(data.matrix(v1),
data.matrix(v2)), dim=c(nrow(v1),ncol(v1),2)))
kml3d(raw, nbClusters=2)

#comando correto
ob1 <- partition(getClusters(raw,2),raw)
ob1["Calinski.Harabatz"] #5.857143

#comando desatualizado
ListPartition_show(raw) #5.779127
```

3.4.2 Outros critérios

Para além dos critérios Calinski-Harabasz já descritos ainda iremos considerar outros dois (todos incluídos na biblioteca kml), nomeadamente

- **Critério Davies - Bouldin:** [36]

$$DB = mean(Prox(cluster_i, cluster_j))$$

com a proximidade *Prox* definida por

$$Prox(i,j) = \frac{DistInt(i) + DistInt(j)}{DistExt(i,j)}$$

onde $DistInt(i)$ representa a distância máxima entre duas trajetórias do cluster i e $DistExt(i,j)$ representa a distância entre os centros dos respetivos clusters.

- **Critério Ray - Turi:** [1]

$$RT = \frac{1}{n \times t} \frac{WGSS}{min_{k<k'} \Delta_{kk'}^2}$$

onde denotamos $\Delta_{kk'}$ como a distância entre os centros dos clusters k e k', ou seja

$$\Delta_{kk'} = ||C^{\{k'\}} - C^{\{k\}}||$$

Tal como atrás, calculámos os valores dos critérios em dada um dos exemplos introduzidos anteriormente.

- Caso unidimensional

Exemplo 3.4.2.1. *Recorde-se que as trajetórias são*

$$y_1 = (5, 2);\ y_2 = (-1, 0);\ y_3 = (9, 4)$$

e estão divididas em dois grupos $g_1 = \{y_1, y_3\}$ *e* $g_2 = \{y_2\}$*. Sendo assim*

$$C^{\{1\}} = (7, 3)\ ;\ C^{\{2\}} = (-1, 0)\ \therefore DistExt(1, 2) = \sqrt{73} = \Delta_{12}$$

Portanto

$$DB = Prox(1, 2) = \frac{d_{13}}{\sqrt{73}} = \frac{\sqrt{20}}{\sqrt{73}} \simeq 0,523$$

e

$$RT = \frac{1}{6} \times \frac{10}{73} \simeq 0,0228$$

- Caso multidimensional

Exemplo 3.4.2.2. *Novamente recordando as trajetórias temos*

$$y_{1..} = \begin{pmatrix} 5 & 2 \\ 3 & -2 \end{pmatrix};\ y_{2..} = \begin{pmatrix} -1 & 0 \\ 9 & 0 \end{pmatrix};\ y_{3..} = \begin{pmatrix} 9 & 4 \\ 1 & 0 \end{pmatrix}$$

estas estão também divididas em $g_1 = \{y_1, y_3\}$ *e* $g_2 = \{y_2\}$*. Assim*

$$C^{\{1\}} = \begin{pmatrix} 7 & 3 \\ 2 & -1 \end{pmatrix}\ ;\ C^{\{2\}} = \begin{pmatrix} -1 & 0 \\ 9 & 0 \end{pmatrix}$$

$$\therefore DistExt(1, 2) = ||(\sqrt{8^2 + 7^2}, \sqrt{3^2 + 1^2})|| = \sqrt{123} = \Delta_{12}$$

Portanto

$$DB = Prox(1, 2) = \frac{d_{13}}{\sqrt{123}} = \frac{\sqrt{28}}{\sqrt{123}} \simeq 0,477$$

e

$$RT = \frac{1}{12} \times \frac{10}{73} \simeq 9,49 \times 10^{-3}$$

[1] Esta é a formula utilizada no pacote do R que é contudo diferente da originalmente proposta em [36]

É importante referir que na biblioteca kml do R, todos estes critérios são uniformizados com o objetivo do utilizador escolher sempre o número de clusters que corresponda ao valor máximo do critério. Na prática os critérios desta subsecção deveriam ser minimizados e portanto na biblioteca estes são multiplicados por -1.

3.5 Metodologia proposta

O nosso principal objetivo é agrupar trajetórias tal que curvas *próximas* e *semelhantes* fiquem juntas. O primeiro passo consiste em definir a função distância entre duas trajetórias, digamos $X_i, X_\ell \in \mathbb{R}^{\mathrm{p}\times\mathrm{T}}$. Em vez da distância Euclidiana usual, que é regularmente utilizada em algoritmos de clustering longitudinal como *kml* e *kml3d*, consideramos a correlação entre os tempos na distância de Mahalanobis, que foi introduzida recentemente por[29] e é dada por

$$d_M^2(X_i, X_\ell) = (X_i - X_\ell)^t \Sigma^{-1} (X_i - X_\ell)$$

onde $\Sigma \in \mathbb{R}^{(p\times T)\times(p\times T)}$ é a matriz diagonal por blocos

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \Sigma_p \end{bmatrix}$$

assumindo que a matriz de variâncias-covariâncias entre diferentes tempos para a variável k é Σ_k, $k = 1, \dots, p$. Esta distância apresenta várias vantagens sobre a distância usual Euclidiana: a) permite que as variâncias em cada direcção possam ser diferentes; b) toma em conta as covariâncias entre diferentes tempos; c) reduz-se à distância Euclidiana para dados não correlacionados com variância unitária.

Considere-se agora o caso de uma única variável (p=1) e iguais variâncias para diferentes tempos. A menos da multiplicação por uma constante, a matriz Σ reduz-se à matriz de correlação entre tempos diferentes. Então, como em séries temporais, podemos modelar esta matriz por uma matriz de correlações auto-regressiva $\Sigma \in \mathbb{R}^{\mathrm{TxT}}$ de ordem 1, AR(1):

$$\Sigma = \begin{bmatrix} 1 & r & r^2 & r^3 & \dots & r^{T-1} \\ r & 1 & r & r^2 & \dots & r^{T-2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ r^{T-1} & r^{T-2} & r^{T-3} & r^{T-4} & \dots & 1 \end{bmatrix}$$

Um estimador de r pode ser obtido do coeficiente de correlação amostral entre o vetor de observações dos tempos 1, 2, ..., $T-1$ e o correspondente de atraso[2] 1. Como exemplo, se existirem 4 tempos diferentes, $\hat{r} = Cov(c(X_{j=1}, X_{j=2}, X_{j=3}), c(X_{j=2}, X_{j=3}, X_{j=4}))$.

Ainda sob a suposição de uma única variável e iguais variâncias através dos instantes de tempo, outra possibilidade é

$$\Sigma = \begin{bmatrix} 1 & r & r & \dots & r \\ r & 1 & r & \dots & r \\ \vdots & \vdots & \vdots & & \vdots \\ r & r & r & \dots & 1 \end{bmatrix}$$

que reflete uma estrutura simétrica composta correspondendo a uma correlação uniforme entre diferentes instantes de tempos.

De facto, qualquer estrutura de uma matriz de correlação poderia ser usada; selecionamos as duas acima por serem as mais comuns em modelação de dados longitudinais.

A situação em que se consideram diferentes variâncias para uma única variável segue da decomposição usual da matriz de variâncias-covariâncias dada por

$$\Sigma = \Gamma R \Gamma$$

onde Γ é matriz diagonal única e R é a matriz de correlação.

Ao longo do estudo, iremos apenas considerar a matriz R para a definição da distância, ignorando as variâncias.

O passo seguinte no processo de clustering consiste na aplicação do algoritmo k-médias com a distância Mahalanobis descrita acima. Consideramos três tipos diferentes de variáveis de entrada: dados brutos e duas versões de dados perfil. De facto, por vezes o interesse em clustering longitudinal depende do comportamento relativo de curvas individuais em vez do seu valor absoluto. Por exemplo, se temos o preço de dois stocks ao longo de T tempos diferentes, podemos estar interessados em agrupar stocks com comportamentos semelhantes (sobem e descem ao mesmo tempo mesmo tomando valores muito diferentes) em vez de stocks que têm preços semelhantes mas comportamentos diferentes. Para estas situações propomos o uso dos perfis, isto é, utilizar uma das seguintes operações iniciais antes de proceder ao clustering:

$$W_{ij} \leftarrow \frac{X_{ij}}{\sum_{l=1}^{n} X_{il}}, \qquad (W_{i1}, \dots, W_{iT}) \leftarrow \frac{1}{\|X_i\|}(X_{i1}, \dots, X_{iT}).$$

Ao longo do estudo estes serão denotados como perfil I e II respetivamente. Enquanto no primeiro a soma dos valores de coordenadas de cada curva é 1, no segundo todas as trajetórias tomam valores na superfície da esférica unitária.

[2] *lag*, em inglês

Agora vamos tratar da aplicação do algoritmo. A aplicação da metodologia anterior requer a definição da distância de Mahalanobis referida, bem como a determinação do estimador correspondente. O processo de aplicação pode ser simplificado, bastando para isso considerar a transformação

$$Y = \Sigma^{-1/2} X.$$

A existência de $\Sigma^{-1/2}$ segue da decomposição de Cholesky de Σ. De facto, para a distância Euclidiana usual d, tem-se

$$\begin{aligned} d^2(Y_i, Y_\ell) &= (Y_i - Y_\ell)^t (Y_i - Y_\ell) \\ &= (\Sigma^{-1/2} X_i - \Sigma^{-1/2} X_\ell)^t (\Sigma^{-1/2} X_i - \Sigma^{-1/2} X_\ell) \\ &= (X_i - X_\ell)^t (\Sigma^{-1/2})^t \Sigma^{-1/2} (X_i - X_\ell) \\ &= d_M^2(X_i, X_\ell) \end{aligned}$$

i.e., a distância Euclidiana entre os dados transformados coincide com a distância de Mahalanobis nos dados originais. Nos cálculos acima é usada a simetria de $\Sigma^{-1/2}$, que é consequência da simetria de Σ.

Em conclusão, pela aplicação da transformação sugerida podemos usar o algoritmo bem conhecido kml ou kml3d [36], para agrupar os dados de acordo com a nova metodologia introduzida em[29].

3.6 Resultados em dados simulados

Nesta secção, a metodologia proposta para o clustering é aplicada em diferentes conjuntos de dados simulados e os resultados são apresentados.

Para a aplicação do clustering K-Médias longitudinal com a distância Euclidiana usámos a biblioteca *kml* do software R [36], construída especificamente para o agrupamento de trajetórias longitudinais. Os conjuntos de dados simulados foram criados a partir da função *gald* (*generateArtificialLongData*), que cria um conjunto de dados longitudinais simulados (unidimensionais) e transforma-os num objeto da classe ClusterLongData, pronto para ser usado pela função *kml*.

Nós gerámos dois conjuntos de dados que estão apresentados na figura 3.1. O primeiro contém 3 clusters, 4 unidades de tempo e 150 trajetórias (50 em cada grupo) podendo todas

estas serem escritas pela notação de uma função afim: 1) $y = -t + 10$, 2) $y = t$ e 3) $y = 15$. O segundo contém 2 clusters, 8 unidades de tempo e 200 trajetórias (100 em cada grupo), obtidas, de um modo semelhante ao anterior, das funções trigonométricas: 1) $y = \sin(t)$ e 2) $y = -\sin(t)$. As trajectórias médias são definidas pelo parâmetro meanTrajectories. As outras trajetórias são determinadas pelos parâmetros *personalVariation* e *residualVariation*; o primeiro define a variação individual entre uma trajetória e a sua trajetória média e o segundo atribui um ruído a todas as trajetórias. Estes dois últimos parâmetros foram modelados com uma distribuição uniforme, respectivamente $\mathcal{U}$(1,4) e $\mathcal{U}$(1,4) para a primeira situação e $\mathcal{U}$(1,10) e $\mathcal{U}$(1,2) na segunda situação.

A escolha destas funções não foi aleatória. À medida que fomos trabalhando o tema percebemos que o método tradicional, que usa a distância Euclidiana, faz sempre uma divisão de grupos por conterem trajetórias próximas e não pelo comportamento destas. Assim, nos nossos exemplos, incluímos trajetórias simples, com alguma sobreposição e comparámos os métodos. Nós aplicámos sempre para o nosso método a transformação Y_A, que usa a matriz AR(1) como matriz de variâncias-covariâncias.

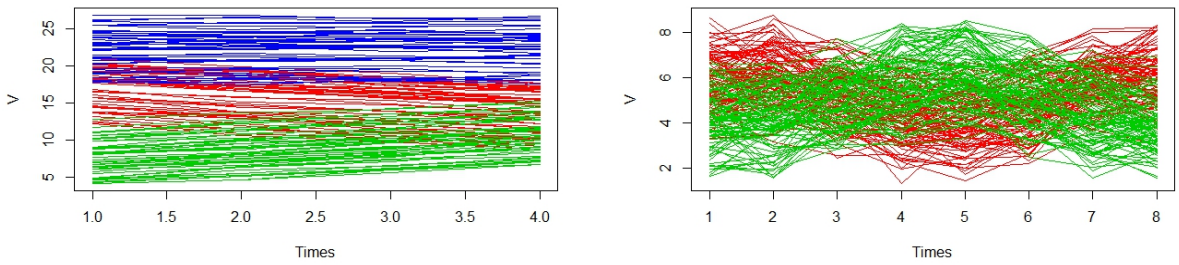


Figura 3.1: Trajetórias simuladas. Conjunto de dados 1 (lado esquerdo), Conjunto de dados 2 (lado direito).

1. Análise do primeiro conjunto de dados:

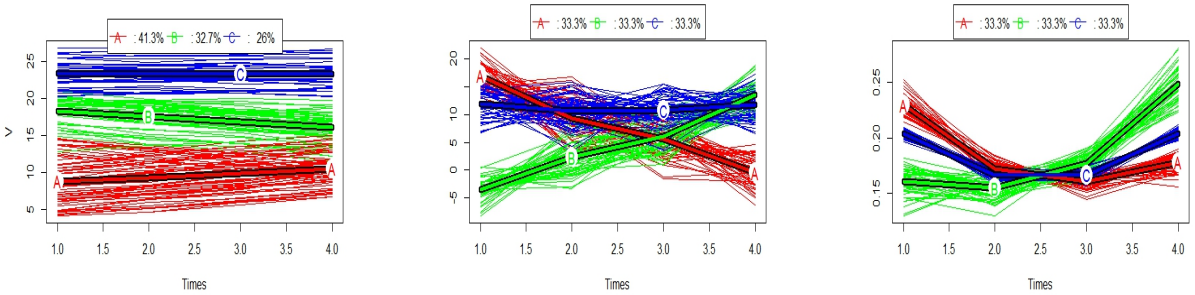


Figura 3.2: Trajetórias afins. Da esquerda para a direita: Dados originais X, Dados transformados Y_A s/ perfis, Dados transformados Y_A c/ perfil I.

Numa primeira análise deparámo-nos rápidamente com uma vantagem do nosso método. Para o método tradicional os critérios sugeriram que o número de grupos mais adequado era 6 e os nossos métodos (utilizando apenas a transformação Y_A ou a transformação com o perfil I) sugeriram 3 grupos, o número de grupos real. Apesar de conhecido o verdadeiro número de grupos, avaliámos ambos os métodos com a verdadeira divisão, tendo o objetivo de analisar a performance dos mesmos e não dos critérios de escolha do número de grupos.

A utilização de perfis num caso como este é sempre indicada dado que esta técnica vai aproximar as trajetórias com o mesmo comportamento e assim fica mais fácil identificar os grupos existentes. Os resultados obtidos para as diferentes metodologias estão apresentados na tabela que se segue.

$C/\hat{C}$	A	B	C
A	37	13	0
B	1	49	0
C	11	0	39

$C/\hat{C}$	A	B	C
A	50	0	0
B	0	50	0
C	0	0	50

Tabela 3.1: Dados originais X (lado esquerdo); Dados transformados Y_A s/ perfis e Y_A c/ perfil I (lado direito, resultados coincidentes).

Neste exemplo ambos os nossos métodos conseguiram ter uma precisão de 100% enquanto o método tradicional obteve uma precisão de $83,33\%$. De facto, mesmo sem perfis o nosso método obteve um resultado excelente. Ainda assim, gráficamente, é mais clara a atribuição dos grupos na metodologia com o perfil.

Os resultados deste exemplo corroboraram as razões intuitivas que mencionámos atrás relativamente às falhas do método tradicional. Este último faz, de facto, uma divisão por camadas e não por comportamentos o que o levou a alguns desacertos principalmente nas zonas de sobreposição entre grupos. Por outro lado, os nossos métodos conseguiram identificar perfeitamente todos os comportamentos, mesmo em zonas de sobreposição, sendo estes os claros fatores de vantagem na nossa metodologia.

2. Análise do segundo conjunto de dados:

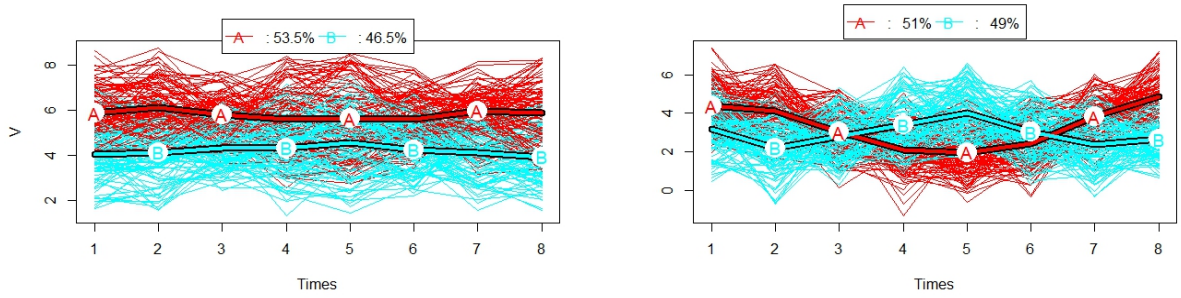


Figura 3.3: Trajetórias trigonométricas. Dados originais X, Dados transformados Y_A s/ perfis.

Uma vez mais a sugestão do número de grupos adequado foi desajustada no método tradicional. Este sugeriu 4 grupos e o nosso método sugeriu os dois grupos reais. Neste exemplo, a utilização dos perfis não foi considerada pois todas as trajetórias (e consequentemente, todos os comportamentos) estão maioritariamente sobrepostas pelo que em casos como este não há necessidade de se utilizar esta técnica.

Tal como no exemplo anterior, avaliámos a performance de ambos os métodos com a divisão correta e apresentámos os resultados na tabela que se segue.

$C/\hat{C}$	A	B
A	66	34
B	41	59

$C/\hat{C}$	A	B
A	100	0
B	2	98

Tabela 3.2: Dados originais X (lado esquerdo); Dados transformados Y_A s/ perfis (lado direito).

O nosso método obteve agora uma precisão de 99% enquanto que o método tradicional conseguiu apenas $62,5\%$. Mais uma vez observamos que quando as trajetórias estão sobrepostas e têm comportamentos diferentes, o método tradicional não tem resultados muito favoráveis. Por outro lado o nosso método apresenta resultados bastante bons.

Para complementar o estudo, também criámos um código, em R, que apresenta a divisão conseguida pela nova metodologia sobre as trajectórias originais. Os resultados obtidos foram os seguintes:

1. Comparação para o primeiro conjunto de dados:

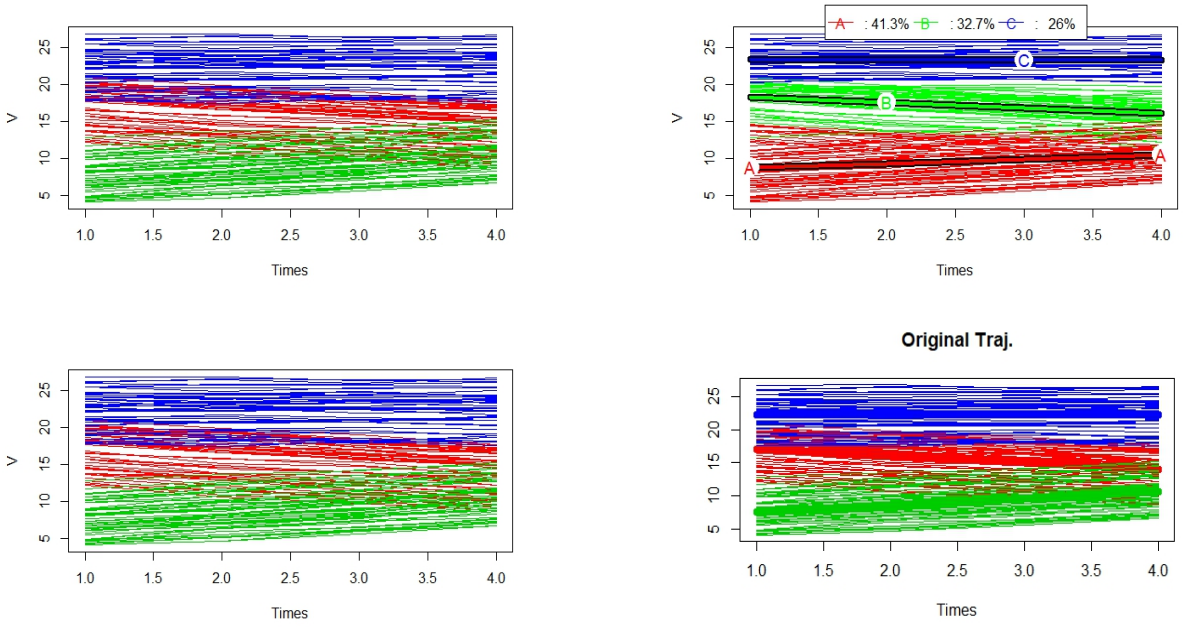


Figura 3.4: Trajetórias afins - Trajetórias com a verdadeira atribuição vs Metodologia tradicional.
Em baixo - Trajetórias com a verdadeira atribuição vs Metodologia inovadora (com e sem perfis, uma vez que os resultados foram coincidentes).

De facto, na metododologia usual este conjunto de simulações apresenta falhas evidentes em regiões onde existem sobreposições de trajetórias. Por exemplo, é possível observar que na passagem do grupo C para o grupo B, o método tradicional não apresentou sobreposições de diferentes comportamentos sendo estas notórias nas trajetórias que apresentam as verdadeiras atribuições.

Por outro lado, a nova metodologia, até mesmo sem perfis, conseguiu identificar todos os comportamentos na perfeição.

2. Agora a comparação para o segundo conjunto de dados:

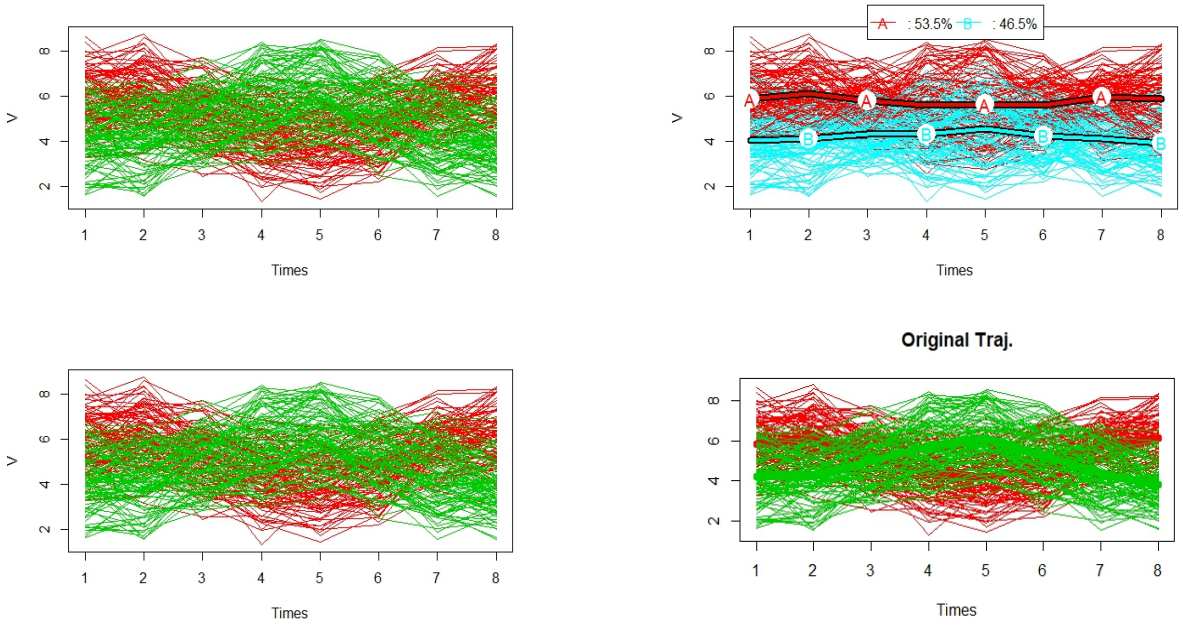


Figura 3.5: Trajetórias trigonométricas - Trajetórias com a verdadeira atribuição vs Metodologia tradicional. Em baixo: Trajetórias com a verdadeira atribuição vs Metodologia inovadora.

No último conjunto de simulações as falhas da metodologia usual são mais evidentes. De facto até as trajetórias médias estão bastante desfasadas do comportamento real. Novamente as sobreposições voltaram a ser essencias para equivocar a metodologia tradicional que desconsidera o comportamento destas.

A nossa metodologia, mais uma vez, conseguiu identificar os diferentes comportamentos de uma forma muito positiva.

3.7 Resultados em dados reais

Nesta subsecção, iremos apresentar os resultados do nosso método para três conjuntos de dados reais, no qual um inclui duas variáveis e servirá para ilustrar a aplicação da nossa metodologia no caso onde temos trajetórias multidimensionais. Nós aplicámos sempre o método tradicional kml e o nosso método (considerando a transformação Y_A, que usa a matriz AR(1) como matriz de variâncias-covariâncias, seguida pelo *kml* ou *kml3d*) bem como nosso método com o perfil I ou II nos casos onde estes se justificam.

1. O conjunto de dados WeightLoss [3] contém perdas de pesos, ao longo de três meses, para 34 indivíduos que estavam divididos em: Controlo, Dieta e Dieta + Exercício.

[3] https://vincentarelbundock.github.io/Rdatasets/datasets.html

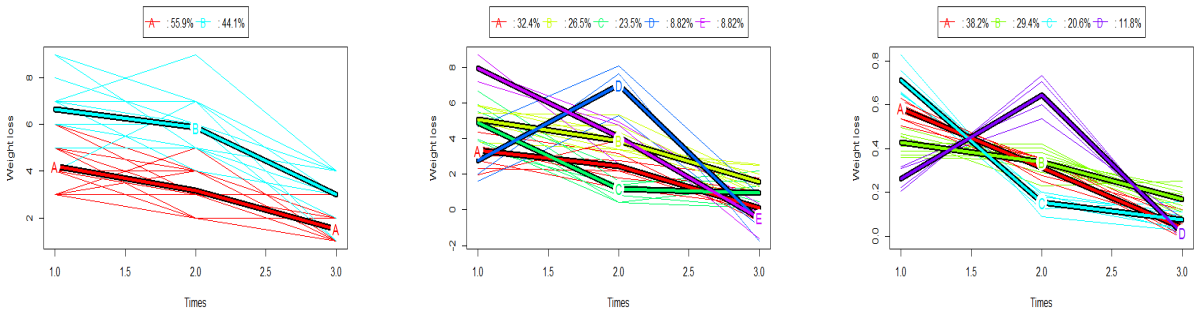


Figura 3.6: Conjunto de dados WeightLoss - Dados originais X, Dados transformados Y_A e Y_A com perfil II

O método tradicional (kml nos dados originais) escolheu dois grupos, com perfis de comportamento semelhantes. É notório que os grupos apenas diferem no seu valor inicial e os padrões médios são semelhantes através do tempo. O nosso método (sem perfis) escolheu cinco grupos diferentes que diferem por vezes no seu valor inicial, por vezes no comportamento ao longo do tempo ou em ambos. Por exemplo, alguns indivíduos têm uma perda de peso muito maior no meio do período em estudo, e o nosso método, contrariamente ao tradicional, deteta esse grupo, que é muito diferente de todos os outros. Um comportamento semelhante é observado nos resultados do nosso método com dados de perfil. Aqui estamos ignorando as diferenças iniciais no peso dos 34 indivíduos e apenas estamos focados nos padrões de perda de peso ao longo do tempo. De acordo com os resultados, existem três (talvez quatro) padrões muito diferentes. No geral, parece-nos que os resultados dos nossos métodos neste exemplo são mais ricos, pois eles detetam diferenças claras nos padrões de perda de peso.

2. O conjunto de dados Sleepstudy [3] contém reações médias de tempo, por dia, de indivíduos que estavam a ser sujeitos à privação do sono. No dia 1, os 18 indivíduos tinham a sua quantidade de sono normal; depois dessa noite eles foram restritos a 3 horas de sono por noite. As observações representam a duração média (em segundos) da reação dos indivíduos a uma série de testes, durante 10 dias.

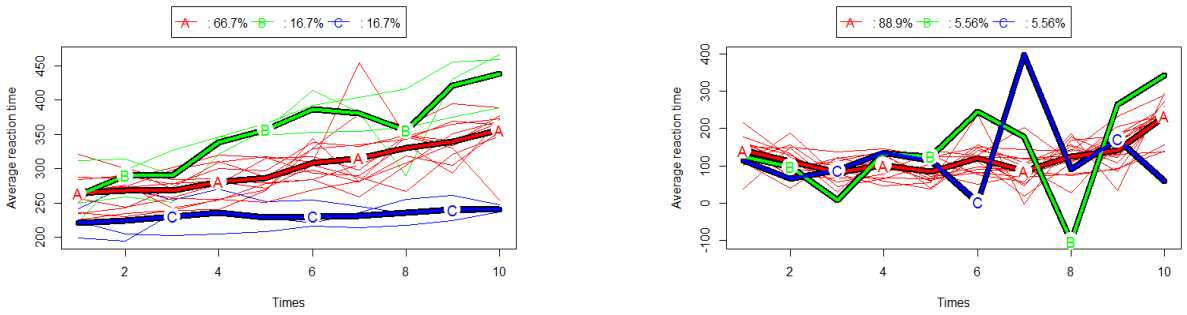


Figura 3.7: Conjunto de dados Sleepstudy - Dados originais X, Dados transformados Y_A s/ perfis

Tanto o método tradicional (kml nos dados originais) como o nosso método (sem perfis) concordaram quanto ao melhor número de grupos, três. Os clusters encontrados pelo método tradicional diferem apenas na velocidade na qual os tempos de reação aumentam ao longo do período de 10 dias. O nosso método encontra um ”grande”cluster correspondente a tempos de reação que oscilam, com tendência a aumentar, ao longo do tempo, e outros dois pequenos clusters, contendo cada um apenas uma observação, correspondendo a dois indivíduos com um comportamento muito diferente dos restantes. Um deles, no dia sete, teve um tempo de reação muito grande e depois retornou ao ”normal”; O outro, no dia 8, teve o comportamento oposto. Esses clusters ”raros”não foram capturados pelo método tradicional. Neste conjunto de dados, o uso de perfis não nos pareceu necessário, pois não houve diferenças de escala nas observações.

3. O conjunto de dados Longair [4] contém duas variáveis, tarifa média e nº médio de passageiros semanais, medidos durante 26 trimestres para 4177 mercados. O 9/11/2001 ocorreu durante o trimestre 21.

   Inicialmente vamos apresentar os resultados para cada variável separadamente e seguidamente apresentaremos os resultados para as trajetórias conjuntas.

   Começamos então por apresentar os resultados para a variável que contém as tarifas médias:

[4] http://www.stat.ufl.edu/ winner/datasets.html

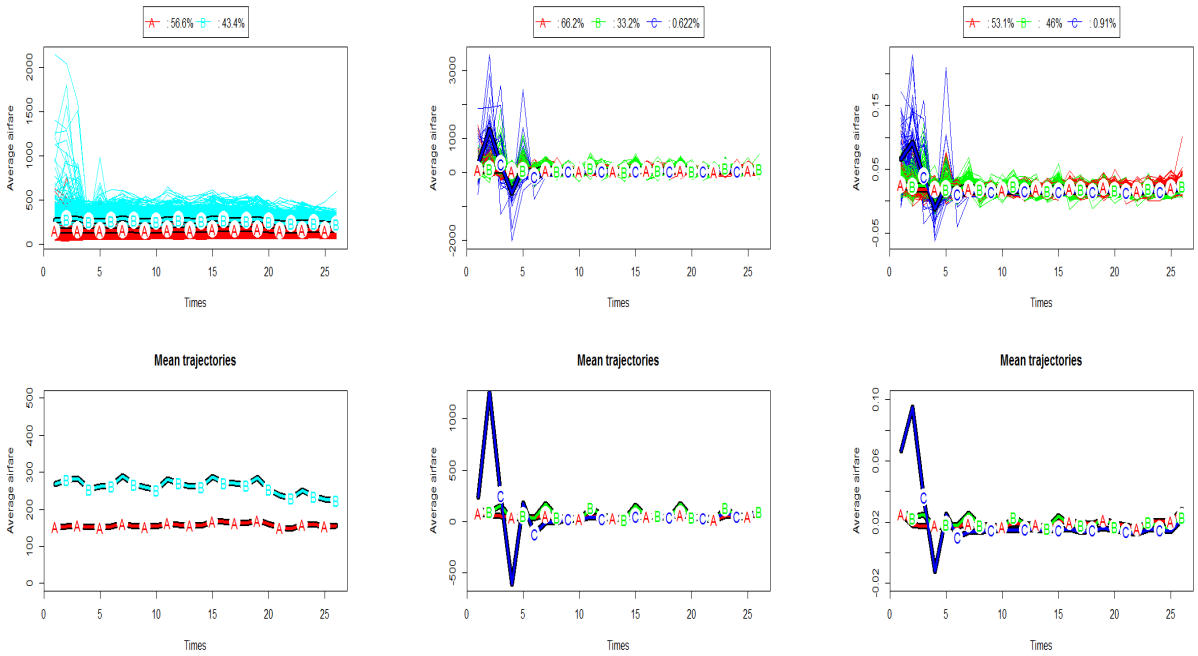


Figura 3.8: Conjunto de dados Longair, primeira variável - Dados originais X, Dados transformados Y_A s/ perfis, Y_A c/ perfil I
Em baixo: Trajetórias médias: Dados originais X, Dados transformados Y_A s/ perfis e Y_A c/ perfil I

O método tradicional criou dois grupos, um representando mercados com tarifas médias altas e outro tarifas médias baixas. Estas representações são, em nossa opinião, uma maneira um tanto elementar de dividir os dados em clusters. O nosso método sem perfis considera a melhor divisão em três grupos. Um grupo com alta variação no início e depois estabiliza e outros dois correspondentes a mercados ”contraditórios”, isto é, mercados que se movem em direções opostas; Quando um aumenta, o outro diminui. Um comportamento semelhante foi observado para a aplicação do nosso método com perfis.

Agora, a análise para a variável com o nº médio semanal de passageiros:

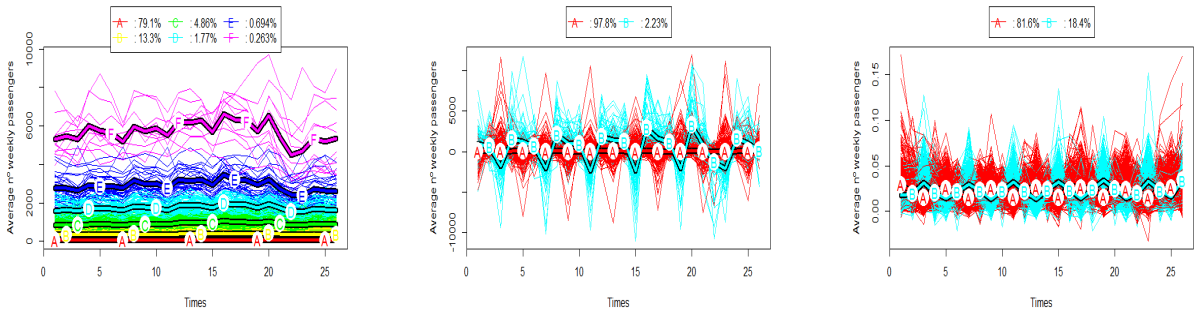


Figura 3.9: Conjunto de dados Longair, segunda variável - Dados originais X, Dados transformados Y_A s/ perfis e Y_A c/ perfil I

Aqui, o método tradicional considera seis grupos, correspondendo a dividir o número de passageiros em diferentes intervalos, de um número reduzido de passageiros a um número elevado de passageiros. O comportamento relativo dos diferentes mercados ao longo do tempo nunca é capturado. Os nossos métodos encontraram dois clusters, novamente como acima correspondendo a mercados que se movem em direções opostas: quando um sobe, o outro desce.

Finalmente, a análise conjunta para as 4177 trajetórias (mercados) para as duas variáveis consideradas:

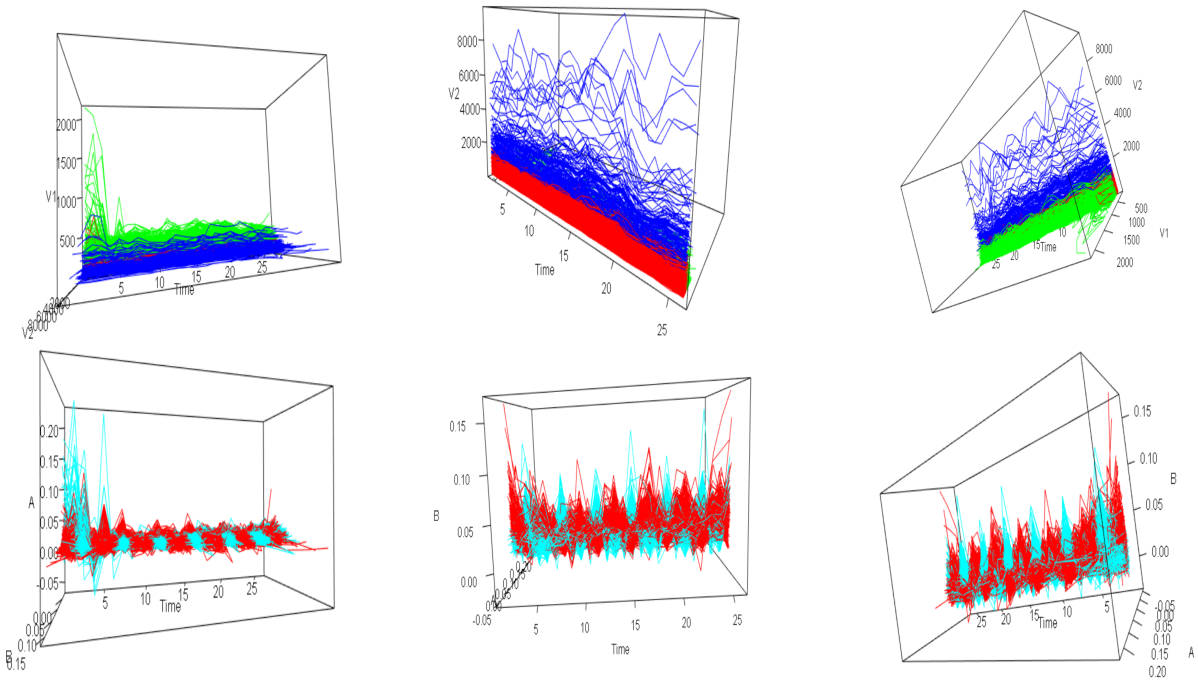


Figura 3.10: Conjunto de dados Longair - Dados originais X
Em baixo: Dados transformados Y_A c/ perfil I

Para o comportamento conjunto, o método tradicional de kml3d considerou três grupos

de mercados, correspondentes a valores altos / baixos em ambas as variáveis, de forma semelhante ao que ocorreu com cada variável separadamente. O nosso método considerou dois grupos que, como no caso unidimensional, têm um comportamento contraditório.

Tal como na secção anterior, para complementar o estudo apresentamos nas trajetórias originais a divisão feita pela nova metodologia. Os resultados obtidos foram os seguintes:

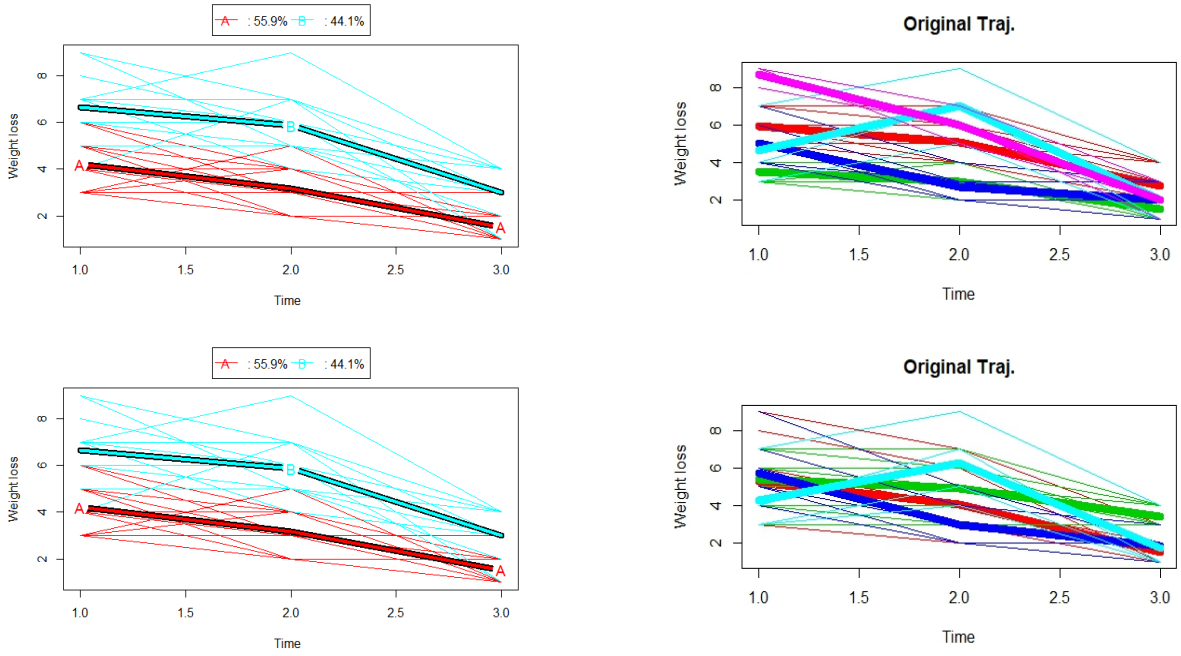


Figura 3.11: Conjunto de dados WeightLoss - Metodologia tradicional, Metodologia inovadora s/ perfis
Em baixo: Metodologia tradicional, Metodologia inovadora c/ perfil II

Para o primeiro conjunto de dados a discussão apresentada acima continua em conformidade. A conjetura de que o método tradicional faz a divisão em grupos ignorando o comportamento das trajetórias e apenas considera a altura destas é bastante evidente nesta base de dados. Contrariamente ao método tradicional, cuja divisão fornecida contém duas trajetórias médias com comportamentos idênticos mas desfasadas em altitude a nossa metodologia sugeriu divisões com comportamentos distintos.

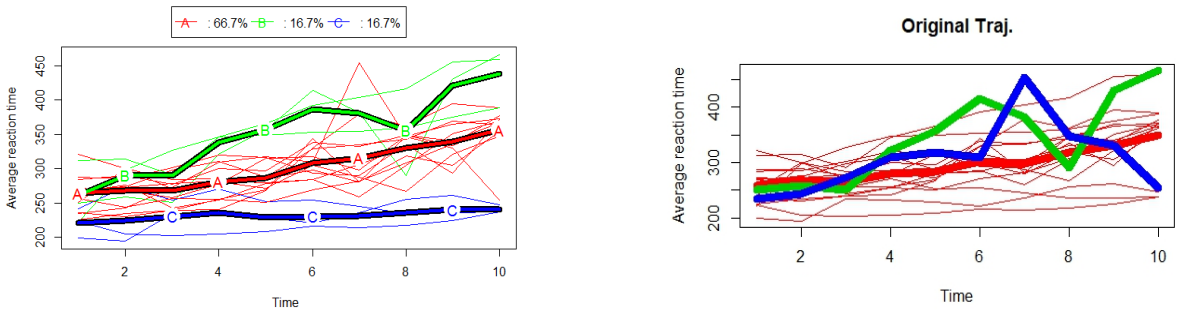


Figura 3.12: Conjunto de dados Sleepstudy - Metodologia tradicional, Metodologia inovadora s/ perfis

Novamente a discussão acima continua em conformidade. A nossa metodologia identificou duas trajetórias com comportamentos completamente diferentes das trajetórias vermelhas com os comportamentos mais comuns.

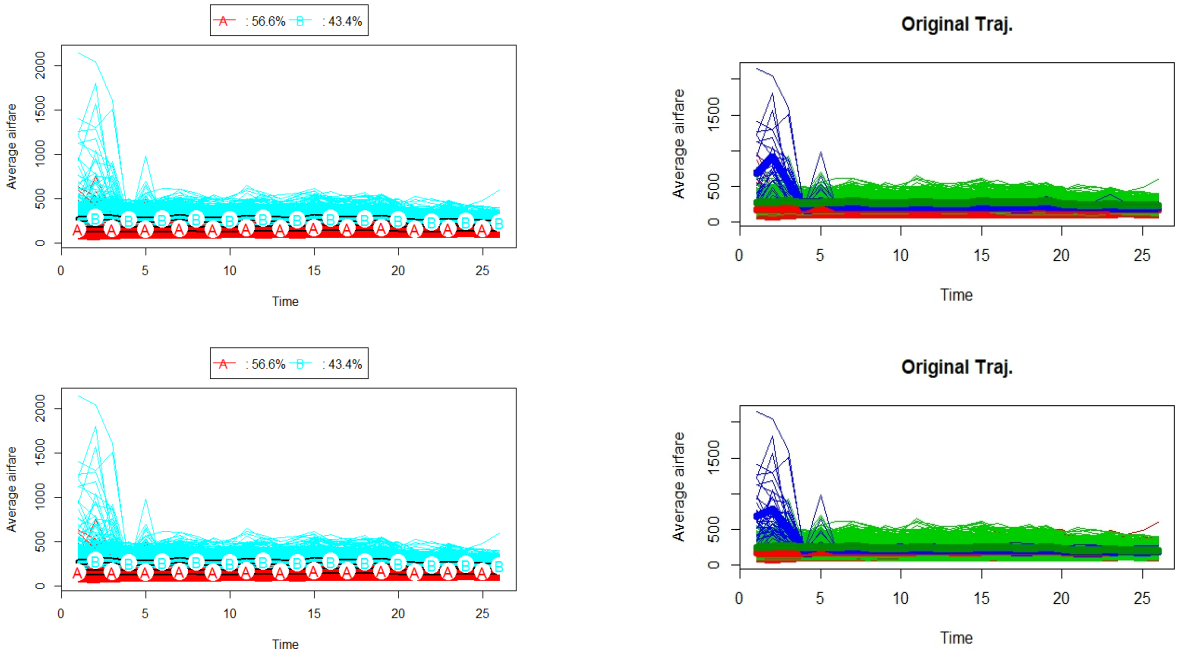


Figura 3.13: Conjunto de dados Longair V1 - Metodologia tradicional, Metodologia inovadora s/ perfis
Em baixo: Metodologia tradicional, Metodologia inovadora c/ perfil I

Neste conjunto de dados não se tornam tão notórios os comportamentos contraditórios, muito provavelmente porque as inclinações não são tão acentuadas nas trajetórias originais, contudo toda a restante discussão continua coerente.

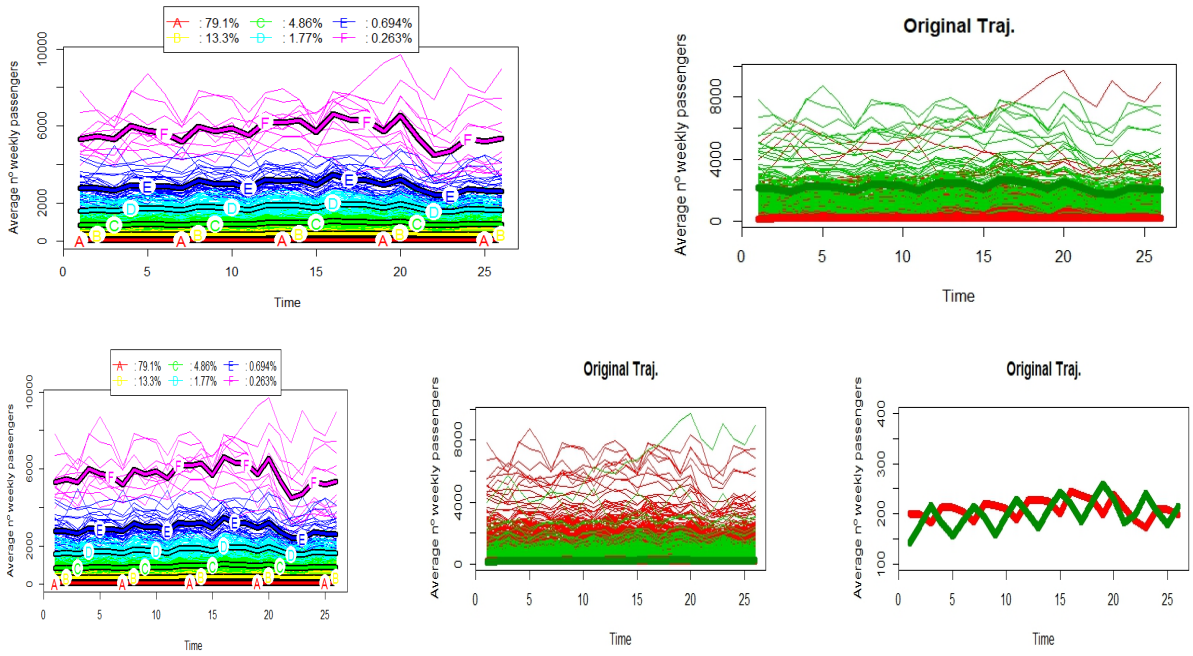


Figura 3.14: Conjunto de dados Longair V2 - Metodologia tradicional, Metodologia inovadora s/ perfis
Em baixo: Metodologia tradicional, Metodologia inovadora c/ perfil I, Trajetórias médias da metodologia inovadora c/ perfil I

Aqui também apresentámos um gráfico só com as trajétorias médias na metodologia com perfis, pois apesar de estas quando colocadas em conjunto com todas as trajetórias se apresentarem sobrepostas, com esta análise mais pormenorizada podemos evidenciar os comportamentos contraditórios descritos anteriormente. De facto, a conjetura volta aqui a ser reafirmada. O método tradicional criou bastantes grupos ”em camadas”ignorando por completo o comportamento das trajetórias.

Para finalizar, consideramos que a programação mais básica e fácil do ponto de vista do utilizador é completamente suficiente para a análise dos conjuntos de dados, tornando esta metodologia de fácil é rápida adaptação a um utilizador comum.

3.8 Discussão

Neste capítudo apresentámos uma nova metodologia de clustering para dados longitudinais e a sua aplicação a dados simulados e também a dados reais. Os resultados da nova metodologia foram bastante positivos e em geral melhores do que os obtidos por aplicação direta do algoritmo kml aos dados originais.

O novo processo é de fácil aplicação; deduzimos uma transformação linear que permite uma aplicação direta do algoritmo tradicional de K-Médias longitudinal, acessível livremente

numa biblioteca do R (*kml* e *kml3d*).

Neste trabalho considerámos trajetórias unidimensionais e bidimensionais para ilustrar a aplicação do nosso método. No caso das trajetórias bidimensional (multidimensional), podemos também, se necessário, considerar não só a correlação entre diferentes tempos dentro de cada variável mas também a correlação entre diferentes variáveis e tempos simultaneamente. Para fazer isso, uma outra estrutura para a matriz Σ precisava ser considerada. Aqui consideramos uma matriz diagonal por blocos.

Dados longitudinais incluem elementos de dados de séries temporais. Embora a Análise Classificatória no contexto das séries temporais tenha sido explorada, houve poucos desenvolvimentos no caso de dados longitudinais, especialmente no que diz respeito a métodos não paramétricos.

Bibliografia

[1] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, 3rd Edition (1984)

[2] J. Cheng, Z. Wang, and G. Pollastri, A Neural Network Approach to Ordinal Regression, Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, (2008)

[3] P. Cortez, Modern Optimization with R, Springer, 1st Edition, pp. 31-117 (2014)

[4] W. Deng, J. Hu, J. Guo, Linear Ranking Analysis, Beijing University of Posts and Telecommunications (2014)

[5] E. Frank and M. Hall, A simple approach to ordinal classification, in EMCL '01: Proceedings of the 12th European Conference on Machine Learning. London, UK: Springer-Verlag, pp. 145–156 (2001)

[6] T. J. Hastie and R. J. Tibshirani, Generalized additive models, Monographs on Statisics and Applied Probability, vol. 43, pp. 297–318, (1990)

[7] R. Herbrich, T. Graepel, and K. Obermayer, Regression models for ordinal data: A machine learning approach, Technical University of Berlin, Tech. Rep., (1999).

[8] J. S. Cardoso and J. F. P. da Costa, Learning to classify ordinal data: the data replication method, Journal of Machine Learning Research, vol. 8, (2007).

[9] J. F. P. da Costa, H. Alonso and J. S. Cardoso, The unimodal model for the classification of ordinal data, Neural Networks, vol. 21, (2008).

[10] J. F. P. da Costa, H. Alonso, and J. S. Cardoso, Corrigendum to: The unimodal model for the classification of ordinal data, Neural Networks, (2014)

[11] J. F. P. da Costa and J. S. Cardoso, Classification of ordinal data using neural networks, Lecture Notes in Artificial Intelligence, vol. 3720 (2005).

[12] J. F. P. da Costa, R. Sousa e J. S. Cardoso. An all-at-once Unimodal SVM Approach for Ordinal Classification. In Proceedings of the Ninth International Conference on Machine Learning and Applications (ICMLA 2010), Washington DC, USA, 12-14 Dec. (2010)

[13] S. Kotsiantis, Cascade generalisation for ordinal problems, International Journal of Artificial Intelligence and Soft Computing, vol. 2, pp. 46–57(12) (2010).

[14] J. R. Magnus and H. Neudecker, Matrix Differential Calculus with applications in Statistics and Econometrics, John Wiley & Sons, Third Edition (2007)

[15] P. McCullagh, Regression models for ordinal data, Journal Royal Statistical Society, Series B, vol. 42, pp. 109–142, (1980)

[16] A. Shashua and A. Levin, Ranking with large margin principle: Two approaches, in Advances in Neural Information Processing Systems 15, Thrun and K. Obermayer, Eds. Cambridge, MA: MIT Press, pp. 937–944 (2003)

[17] B. Sun et al., Kernel Discriminant Learning for Ordinal Regression, IEEE transactions on knowledge and data engineering, vol. 22, (2010)

[18] G. Tutz, Generalized semiparametrically structured ordinal models, Biometrics, vol. 59, pp. 263–273, (2003)

[19] V. N. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1st Edition (1998).

[20] Leonard Kaufman, Peter J. Rousseeuw; Finding groups in data - An introduction to cluster analysis; Wiley; 1-37 (1990).

[21] Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl; Cluster Analysis; Wiley - 5th edition; 111-142 (2010).

[22] Bernard Desgraupes; Clustering Indices; Package clusterCrit for R (2013)

[23] Pierre Legendre, Louis Legendre; Numerical Ecology; Elsevier - 3rd edition; 247-264 (2012).

[24] C. Abraham, P. Cornillon, E. Matzner-Lober, N. Molinari, Unsupervised curve clustering using B-splines, Scandinavian Journal of Statistics 30, pages 581–595 (2003)

[25] Beauchaine TP, Beauchaine RJ, A Comparison of Maximum Covariance and K-Means Cluster Analysis in Classifying Cases Into Known Taxon Groups. Psychological Methods 7(2): 245–261 (2002)

[26] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, Azam Niknafs. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, Journal of Statistical Software, Volume 63, Issue 6 (2014)

[27] A. Ciampi et al., Model-Based Clustering of Longitudinal Data: Application to Modeling Disease Course and Gene Expression Trajectories, Communications in Statistics - Simulation and Computation (2012)

[28] Chi-hung Clarence Ng, Examining the self-congruent engagement hypothesis: the link between academic self-schemas, motivational goals, learning approaches and achievement within an academic year, Educational Psychology (2013)

[29] J. Pinto Costa et al. Some Developments in the Clustering of Longitudinal Trajectories, Journal of Statistical Computation and Simulation, (submitted, 2017)

[30] Elizabeth C. Delmelle, Mapping the DNA of Urban Neighborhoods: Clustering Longitudinal Sequences of Neighborhood Socioeconomic Change, Annals of the American Association of Geographers (2015)

[31] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger, Analysis of Longitudinal Data. Oxford University Press Inc., New York (2002).

[32] Garrett Fitzmaurice, Nan Laird, James Ware, Applied Longitudinal Analysis. John Wiley & Sons, Inc., New Jersey (2004).

[33] Fraley, C. and Raftery, A.E., Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97, 611-631 (2002)

[34] C. Genolini et al., KmL: k-means for longitudinal data, Springer, Volume 25, Issue 2, pp. 317-328 (2010)

[35] Christophe Genolini, René Écochard, Hélène Jacqmin-Gadda. Copy Mean: A New Method to Impute Intermittent Missing Values in Longitudinal Studies, Open Journal of Statistics (2013)

[36] C. Genolini et al., kml and kml3d: Packages to Cluster Longitudinal Data, Journal of Statistical Software, Volume 65, Issue 4, (2015)

[37] T. Hastie et al, The Elements of Statistical Learning. Data Mining Inference and Predictions. Springer (2009)

[38] Donald Hedeker, Robert D. Gibbons, Longitudinal Data Analysis. Wiley Series in Probability and Statistics, (2006)

[39] B.C. Heggeseth, Longitudinal Cluster Analysis with applications to growth trajectories, University of California, Berkeley (2013)

[40] G. James, C. Sugar, Clustering for sparsely sampled functional data, Journal of the American Statistical Association 98, pages 397–408 (2003)

[41] Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ, Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC Bioinformatics 5:172 (2004)

[42] Antonello Maruotti et al., Time-varying clustering of multivariate longitudinal observations, Communications in Statistics - Theory and Methods (2016)

[43] Melnykov, Volodymyr and Maitra, Ranjan. Finite mixture models and model-based clustering. Statist. Surv. 4 (2010)

[44] Milligan GW, Cooper MC, An Examination of Procedures for Determining the Number of Clusters in a Data Set, Psychometrika, pages 159-179 (1985)

[45] Robin Morris et al, Developmental classification of reading-disabled children. Journal of Clinical and Experimental Neuropsychology (1986)

[46] Oh, M.-S. and Raftery, A.E. Model-based Clustering with Dissimilarities: A Bayesian Approach. Journal of Computational and Graphical Statistics, 16, 559-585. (2007)

[47] Céline Le Pichon et al., Using a continuous riverscape survey to examine the effects of the spatial structure of functional habitats on fish distribution, Journal of Freshwater Ecology, Volume 31, (2015)

[48] Sijun Qin et al, Forage crops alter soil bacterial and fungal communities in an apple orchard. Acta Agriculturae Scandinavica, Section B — Soil & Plant Science (2015)

[49] F. Rossi, B. Conan-Guez, A.E. Golli, Clustering functional data with the SOM algorithm, in: Proceedings of ESANN, pages 305–312 (2004)

[50] Yosung Shim, Jiwon Chung, In-Chan Choi. A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm, IEEE Computer Society, (2005)

[51] P. Sousa, A. Oliveira, M. Gomes, A.R. Gaio, R. Duarte, Longitudinal clustering of tuberculosis incidence and predictors for the time profiles: the impact of HIV. Int J Tuberc Lung Dis., Volume 20(8), pages 1027-32 (2016)

[52] T. Tarpey, K. Kinateder, Clustering functional data, Journal of Classification 20, pages 93–114 (2003)

[53] Duy Q. Vu, David R. Hunter and Michael Schweinberger, Model-based Clustering of Large Networks. The Annals of applied Statistics. (2013)