Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Agrupamento baseado em um *Kernel* de Mahalanobis adaptativo

Fernanda Florencio Costa

Abril/2020

Fernanda Florencio Costa

Agrupamento baseado em um *Kernel* de Mahalanobis adaptativo

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba como requisito fundamental para obtenção do Grau de Bacharel em Estatística.

Orientador: Prof. Dr. Marcelo Rodrigo Portela Ferreira

João Pessoa
Abril de 2020

Catalogação na publicação
Seção de Catalogação e Classificação

C838a Costa, Fernanda Florencio.
Agrupamento baseado em um Kernel de Mahalanobis adaptativo / Fernanda Florencio Costa. - João Pessoa, 2020.
49 p. : il.

Orientação: Marcelo Rodrigo Portela Ferreira.
TCC (Graduação/Bacharelado em Estatística) - UFPB/CCEN.

1. Funções Kernel. 2. Método de agrupamento - Distância de Mahalanobis. I. Ferreira, Marcelo Rodrigo Portela. II. Título.

UFPB/CCEN CDU 519.6(043.2)

Elaborado por Josélia Maria Oliveira da Silva - CRB-15/113

Agrupamento baseado em um *Kernel* de Mahalanobis adaptativo

Fernanda Florencio Costa

Banca Examinadora

Profº Dr. Marcelo Rodrigo Portela Ferreira
Orientador - UFPB

Profº Dr. Eufrásio de Andrade Lima Neto
Examinador - UFPB

Profº Dr. Telmo de Menezes e Silva Filho
Examinador - UFPB

CONCEITO FINAL: _____

Com amor, dedico esse trabalho aos meus pais, Ivonete e Antônio. Aos meus familiares e amigos que, com carinho e apoio, contribuíram para que eu chegasse até esta etapa da minha vida.

Agradecimentos

Em primeiro lugar agradeço à Deus que me deu forças para perseverar e nunca me deixou faltar nada. Grata pelo Senhor sempre iluminar meus passos e me guiar pelos melhores caminhos.

Aos meus pais, Ivonete e Antônio, vocês são minha base. Obrigada por acreditarem na minha capacidade e investirem na minha pessoa. Agradeço pelo apoio incondicional em todos os momentos difíceis da minha vida acadêmica e pessoal e por todos os valores ensinados até hoje.

Aos meus irmãos Thiago e Thiego por sempre encontrarem uma forma de me tirar um sorriso. Em especial a Thiego, por todos os conselhos quando pensei em desistir e todo cuidado comigo.

A todos da minha família que se fazem presente em minha vida, por todo carinho e palavras de incentivo. Em especial minha Vó, Maria, que mesmo sem compreender direito meu percurso, com suas doces e simples palavras sempre me motivou a seguir em frente.

A todos os professores do Departamento de Estatística pelos ensinamentos. Em especial ao professor Marcelo, por toda orientação nessa monografia e nos projetos de PIBIC. Agradeço por sua confiança, seus ensinamentos, paciência e pelos momentos de descontração compartilhados.

Aos meus amigos(as) Priscila, Matheus, Lary, Bruna e Lucas por todo companheirismo, apoio, brincadeiras, palavras de incentivo, festinhas, risadas e lágrimas enxugadas.

Aos meus amigos da graduação para vida, Milleny e Netto, por tudo que vivenciamos juntos, por todo carinho, incentivo, aprendizado, conselhos, noites em claro, boas risadas e por aguentarem meus pequenos surtos durante todo esse período de graduação.

Sem citar nomes, para não ser injusta, agradeço a todos os amigos e colegas conquistados durante a gradução por todo aprendizado compartilhado e bagunça.

Aos professores Eufrásio e Telmo, pela disposição em participar da banca avaliadora e por todas as sugestões e incentivos a esse trabalho.

"A força não provém da capacidade física. Provém de uma vontade indomável".

— Mahatma Gandhi

Resumo

Métodos de agrupamento são ferramentas úteis para explorar estruturas em conjuntos de dados e têm sido largamente utilizados para reconhecimento não supervisionado de observações. Agrupar significa organizar um conjunto de observações (objetos, indivíduos, pixels, etc.) em grupos de forma que observações pertencentes a um mesmo grupo têm alta similaridade, enquanto que observações pertencentes a grupos diferentes têm alta dissimilaridade. A medida de dissimilaridade é um componente importante de qualquer método de agrupamento e a distância Euclidiana é a mais comumente utilizada. Contudo, só apresenta bom desempenho em dados cujos grupos são aproximadamente hiperesféricos e/ou aproximadamente linearmente separáveis. Diversos métodos capazes de lidar com dados cuja estrutura é complexa têm sido propostos, dentre os quais, métodos de agrupamento baseados em funções *kernel*. A essência desses métodos envolve a realização de um mapeamento não-linear Φ do espaço original p-dimensional $X \subset \mathbb{R}^p$ para um espaço de dimensão mais alta (possivelmente infinita), chamado de espaço de características, $\mathcal{F}$. A função *kernel* mais comumente utilizada é a Gaussiana, que, apesar de suas boas características, é baseada na distância Euclidiana, que como dito anteriormente, apesenta limitações. Assim, a distância de Mahalanobis, que leva em consideração as correlações entre variáveis e é invariante em escala, é uma melhor escolha para lidar com dados cuja estrutura é complexa. O objetivo desse trabalho é propor métodos de agrupamento baseados em um *kernel* de Mahalanobis adaptativo, no qual a matriz de covariância pode ser diferente para cada grupo ou comum a todos os grupos, podendo ser ainda, em cada caso, diagonal, quando não considera as covariâncias das variáveis ou completa, caso em que as covariâncias são consideradas. Essa abordagem permite lidar com uma classe muito maior de problemas de agrupamento. Experimentos com dados simulados e reais mostram a eficiência dos métodos propostos.

Palavras-Chave: Agrupamento, Funções *Kernel*, Distância de Mahalanobis.

Abstract

Clustering methods are useful tools for exploring structures in data sets and have been widely used for unsupervised pattern recognition. Grouping means organizing a set of observations (objects, individuals, pixels, etc.) into groups in such a way that observations belonging to the same group have high similarity, while observations belonging to different groups are highly dissimilar. The dissimilarity is an important component of any grouping method and the Euclidean distance is the most commonly used. However, it only performs well on data whose groups are approximately hyperspherical and/or approximately linearly separable. Several methods capable of dealing with data whose structure is complex have been proposed, among which, clustering methods based on kernels functions. The essence of these methods involves performing a nonlinear mapping Φ from the original p-dimensional space $X \subset \mathbb{R}^p$ to a higher (possibly infinite) space, called feature space, $\mathcal{F}$. The most commonly used kernel function is the Gaussian, and, although its good characteristics, it is based on the Euclidean distance, which, as stated earlier, has limitations. Thus, the Mahalanobis distance, which takes into account the correlations between variables and is scale-invariant, is a better choice for dealing with data whose structure is complex. The aim of this work is to propose clustering methods based on an adaptive Mahalanobis, in which the covariance matrix can be different for each group or the same for all groups. These matrices can be also diagonal, when covariances are not considered, or full, when covariances are taken into account. This approach allows to deal with a much larger class of grouping problems. Experiments with simulated and real data show the efficiency of the proposed method.

Keywords: Clustering, *Kernel* Functions, Mahalanobis distance.

Sumário

Lista de Algoritmos

Lista de Tabelas

Lista de Figuras

CAPÍTULO 1

Introdução

1.1 Introdução

É comum muitas aplicações fornecerem ou guardarem grandes quantidades de dados para análise. Essas informações, quando analisadas individualmente, podem não ser úteis, sendo necessária uma análise conjunta de todos os dados. Um dos principais meios encontrados para essa análise é a categorização dos dados em grupos.

Reunir objetos similares em grupos para produzir determinada divisão é uma das habilidades básica dos seres humanos. Isso vem sendo feito desde os primórdios do seu surgimento quando, por exemplo, eles identificaram que muitos objetos possuiam certas propriedades, como a comestibilidade de alimentos, a usabilidade de ferramentas, entre outros. Desta forma, surge a ideia de agrupamento.

Métodos de agrupamento são ferramentas usadas na mineração de dados e tem sido amplamente utilizados para o reconhecimento não supervisionado de observações. A mineração de dados compreende os principais algoritmos que permitem obter ideias e conhecimentos fundamentais a partir de dados massivos. É um campo interdisciplinar que mescla conceitos de áreas afins, como sistemas de banco de dados, estatística, aprendizado de máquina e reconhecimento de observações. (ZAKI; JR; MEIRA, 2014)

Agrupar é uma tarefa natural para os seres humanos e consiste em organizar um conjunto de observações (objetos, indivíduos, genes, pixels, etc.), com base em critérios de similaridade (ou dissimilaridade), em grupos, de tal forma que observações pertencentes a

um mesmo grupo têm um alto grau de similaridade, enquanto que observações pertencentes a grupos diferentes têm um alto grau de dissimilaridade. Os métodos de agrupamento são abordados em muitos contextos e disciplinas, por exemplo, recuperação de documentos, segmentação de imagens, segmentação de mercado, agrupar páginas de internet com base no seu conteúdo, agrupar pessoas em redes sociais com base em suas preferências e/ou características e como dito anteriormente, mineração de dados. A finalidade da análise de agrupamento é que observações em um mesmo grupo sejam homogêneas e observações em grupos diferentes sejam heterogêneas. As técnicas de agrupamento mais populares podem ser divididas em métodos hierárquicos e métodos não hierárquicos (particionais).

Os métodos de agrupamento hierárquicos são capazes de encontrar estruturas que podem ser divididas em subestruturas recursivamente. Tais métodos produzem uma resposta representada por uma estrutura completa de hierarquia, isto é, uma sequência aninhada de partições do conjunto de observações de entrada. O resultado é uma estrutura hierárquica de grupos conhecida como dendrograma. Por exemplo, em taxonomia, um objeto pode pertencer sucessivamente a uma espécie, um gênero, uma família, uma ordem, etc. O nó raiz do dendograma representa todo o conjunto de dados e cada nó folha é considerado como um objeto de dados. Os nós intermediários descrevem a extensão em que os objetos são próximos um do outro e a altura do dendograma, geralmente, expressa a distância entre cada par de objetos ou aglomerados ou um objeto e um aglomerado.

Essa representação fornece descrições e visualizações muito informativas para as estruturas de agrupamento de dados em potencial, especialmente quando existem relações hierárquicas reais nos dados, como os dados de pesquisas evolutivas sobre diferentes espécies de organizações. Esses algoritmos são classificados em métodos aglomerativos e métodos divisivos. Os métodos aglomerativos, partindo de n grupos unitários, onde n representa o número de observações, reduzem os dados em um único grupo contendo todos os elementos através de uma sequência de junções. Os divisivos, por sua vez, partem de um grupo único com n elementos e através de sucessivas divisões resultam em n grupos unitários (XU; WUNSCH, 2005).

Os métodos de agrupamento particionais produzem uma partição única do conjunto de observações em um número fixo de grupos, através da otimização (geralmente local) de uma função objetivo. Não apresenta nenhuma outra subpartição, como no caso dos algoritmos hierárquicos. O resultado é a criação de hipersuperfícies de separação entre os grupos. Esses métodos são mais flexíveis que os hierárquicos e permitem que elementos amostrais mudem de grupo em cada passo do algoritmo, se essa mudança permitir uma melhor solução em termos de variabilidade da partição resultante. Os métodos k-médias e k-medóides são os principais algoritmos de agrupamento não-hierárquicos. A idéia do algoritmo k-médias é alternar entre a atualização dos centróides dos grupos e a atualização

da partição até que nenhuma observação mude de grupo. O k-medóides possui estrutura e funcionamento similares ao k-médias, o que os diferencia é que enquanto no k-médias o centróide não precisa pertencer ao conjunto de dados, no k-medóides o centro é um dos pontos do conjunto, denominado medóide (DONI, 2004). Métodos de agrupamento particionais são desenvolvidos sob dois paradigmas: agrupamento *hard* (rígido) e agrupamento *fuzzy* (difuso). Nos métodos de agrupamento do tipo rígido os grupos são naturalmente disjuntos, isto é, cada observação pode pertencer a um, e somente um, grupo. No caso dos métodos de agrupamento difuso uma mesma observação pode pertencer a todos os grupos com um certo grau de pertinência.

Um princípio básico para construção de algoritmos de agrupamento são as medidas de dissimilaridade (ou similaridade). Exemplos básicos de medidas de dissimilaridade são as medidas de distância, dentre as quais, as mais comuns são a Euclidiana, a de Mahalanobis e o Coeficiente de Correlação de Pearson. Dentre essas, a distância Euclidiana é a mais comumente utilizada. Segundo Ferreira e Carvalho (2014) os métodos de agrupamento baseados na distância euclidiana apresentam bom desempenho quando as estruturas dos grupos são aproximadamente hiperesféricas e/ou aproximadamente linearmente separáveis. No entanto, quando a estrutura dos dados apresenta formas não hiperésfericas e/ou observações não lineramente separáveis o desempenho pode não ser satisfatório. Devido a essa limitação, diversos métodos vêm sendo propostos para lidar com dados cuja estrutura é complexa, dentre os quais, métodos de agrupamento baseados em funções *kernel*.

Diversos estudos demonstram que os métodos de agrupamento baseados em *kernel*, por produzirem hipersurperfícies não lineares de separação entre os grupos, apresentam desempenho superior aos métodos de agrupamento convencionais quando a estrutura dos dados é complexa. Algoritmos baseados em *kernel* estão sendo desenvolvidos sob duas abordagens: kernelização da métrica, onde os centróides dos grupos são obtidos no espaço original das observações e as distâncias das observações aos centróides são calculadas através de funções *kernel*, e agrupamento no espaço de características, em que os centróides dos grupos são obtidos no espaço de características (FERREIRA; CARVALHO, 2014).

A base dos algoritmos de agrupamento baseados em *kernel* é a função *kernel*, que mede a similaridade entre dois exemplos em um espaço p-dimensional. A função *kernel* mais comumente utilizada é a Gaussiana. Apesar de suas boas características, esse *kernel* é baseado na distância Euclidiana, isto é, assume que as observações são mais prováveis de serem distribuídas em uma região hiperesférica (variâncias iguais e covariâncias nulas). Entretanto, exemplos em dois grupos diferentes são mais prováveis de estarem distribuídos dentro de duas regiões hiperelipsoidais diferentes.

A distância de Mahalanobis, que leva em consideração as correlações entre as variá-

veis e é invariante em escala, é uma melhor escolha para lidar com regiões hiperelipsoidais (WANG; YEUNG; TSANG, 2007). Ferreira e Carvalho (2014) propuseram, sob o enfoque da kernelização da métrica, um algoritmo do tipo *kernel* k-médias baseado em um *kernel* adaptativo de Mahalanobis. Esse algoritmo foi construído a partir de uma distância quadrática adaptativa definida por uma matriz simétrica positiva definida que é modificada a cada iteração do algoritmo. Contudo o método proposto considera apenas o caso em que a matriz de covariâncias é a mesma para todos os grupos, tendo a matriz diagonal (caso em que não há covariância entre as variáveis) como caso particular, isto é, os grupos podem assumir formas hiperelipsoidais, mas existe a imposição de que essas regiões sejam iguais.

Nesse trabalho de conclusão de curso, serão apresentada extensões do método proposto por Ferreira e Carvalho (2014) para o caso em que a matriz de covariâncias pode ser diferente para cada grupo, isto é, os grupos podem assumir formas hiper-elipsoidais e essas regiões podem ser diferentes entre os grupos. Dessa forma, uma classe muito maior de problemas de agrupamento pode ser abordada.

1.2 Organização do trabalho

Além deste capítulo de introdução, este trabalho é composto por mais quatro capítulos, como segue:

- No Capítulo 2 é apresentada uma revisão bibliográfica a respeito da teoria básica dos métodos de agrupamento rígido (*hard*), bem como a ideia dos métodos baseados em *kernel*;
- O Capítulo 3 introduz a principal contribuição deste trabalho: métodos de agrupamento baseados no *kernel* de Mahalanobis com distâncias quadráticas adaptativas. Aqui foi considerado o caso sob kernelização da métrica e demonstradas as expressões matemáticas obtidas, bem como os algoritmos construídos para fins computacionais;
- No Capítulo 4 é apresentado os índices de avaliação utilizados para comparação dos métodos, além dos conjuntos de experimentos numéricos aplicados tanto a dados simulados quanto a dados reais. Os resultados obtidos mostram a superioridade do método proposto neste trabalho, em relação aos métodos de agrupamento convencionais;
- Por fim, no Capítulo 5, são apresentadas as conclusões deste trabalho.

CAPÍTULO 2

Métodos de agrupamento *hard* para dados convencionais

A Análise de Agrupamentos é uma técnica de análise de dados multivariados bastante útil em muitas situações diferentes. Pode ser utilizada para reduzir a dimensão de um conjunto de dados, resumindo uma ampla gama de objetos à informação do centro do seu conjunto. O objetivo é formar grupos de observações onde os elementos pertencentes ao mesmo grupo sejam mais similares entre si do que em relação aos elementos de grupos distintos. Além disso, como é uma técnica de aprendizado não supervisionado, também pode ser útil para extrair características escondidas dos dados e desenvolver hipóteses a respeito de sua natureza.

Segundo Xu e Tian (2015), os algoritmos tradicionais de agrupamento podem ser divididos em 9 categorias. A Tabela 2.1 contém essas categorias e seus respectivos algoritmos.

Tabela 2.1: Algoritmos Tradicionais

Categoria	Algoritmos típicos
Métodos Particionais	k-médias, k-medóides, PAM, CLARA, CLARANS
Métodos Hierárquicos	BIRCH, CURE, ROCK, Chameleon
Métodos baseados em teoria fuzzy	FCM, FCS, MM
Métodos baseados em distribuições	DBCLASD, GMM
Métodos baseados em densidade	DBSCAN, OPTICS, Mean-shift
Métodos baseados na teoria dos grafos	CLICK, MST
Métodos baseados em grade	STING, CLIQUE
Métodos baseados na teoria fractal	FC
Métodos baseados em modelos	COBWEB, GMM, SOM, ART

Dentre os métodos citados na Tabela 2.1, duas classes de métodos prevalecem, são os métodos hierárquicos e os não-hierárquicos (ou particionais). Os Hierárquicos organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos. Sua saída é um conjunto aninhado de partições conhecido como dendograma. Esses métodos se dividem em aglomerativos (consideram n grupos iniciais com uma observação cada e sucessivamente chega a um único grupo contendo todas as observações) e divisivos (consideram um único grupo inicial com todas as observações e sucessivamente chega a n grupos cada um com uma observação). Os Particionais são baseados na minimização de uma função objetivo, onde as observações são agrupados em um número k de grupos definidos a priori. Esses métodos contém duas categorias, rígida (*hard*) e difusa (*fuzzy*). Os métodos *hard* consideram que cada observação só pode pertencer a um grupo por vez, ou seja, não permite interseção entre os grupos. Já os métodos *fuzzy*, cada observação pertence a um grupo com um certo grau de pertinência. Uma exposição detalhada dos principais métodos de agrupamento difuso pode ser encontrada (HÖPPNER et al., 1999).

Os métodos hierárquicos são fundamentados em funções de ligação que são usadas em cada passo do algoritmo para comparação dos grupos disponíveis para agrupamentos. Existem várias funções de ligação e cada uma gera um método diferente. A Tabela 2.2 apresenta algumas das funções de ligação mais tradicionais.

Tabela 2.2: Principais funções de ligação dos métodos hierárquicos

Método	Fórmula
Ligação Simples	$d(i,j) = \min d(i,j)$, em que $i \in C_1, j \in C_2$
Ligação Completa	$d(i,j) = \max d(i,j)$, em que $i \in C_1, j \in C_2$
Ligação Média	$d(i,j) = \frac{1}{n_i,n_j} \sum_{i,j \in C_1,C_2} d(i,j)$, em que $i \in C_1, j \in C_2$
Centróides	$d(i,j) = d(\mu_i, \mu_j)$, em que $\mu_i \in C_1, \mu_j \in C_2$

No método de ligação simples a métrica é a distância mínima de cada par de objetos formados por um objeto pertencente a C_i e outro pertencente a C_k. No caso do método de ligação completa a dissimilaridade entre dois agrupamentos, C_i e C_k, é definida como a maior das dissimilaridades dentro de cada par de objetos formados por um objeto pertencente a C_i e outro pertencente a C_k, isto é, para todos os objetos $j \neq i, k$ (COSTA, 2005). O método de ligação média considera uma média de todos os membros dos grupos cuja distância está sendo calculada. Como consequência, ele é menos influenciado por valores extremos, como é o caso dos métodos de ligação simples e ligação completa. O método de ligação por centroídes considera a distância entre grupos como sendo o quadrado da distância euclidiana entre os centróides dos grupos.

Técnicas hierárquicas são boas ferramentas para produzir uma partição inicial de um conjunto de observações. Contudo, a propriedade de hierarquia causa problema de não dissociação de grupos que foram unidos em determinado passo do algoritmo, pois só é permitido a entrada de elementos nos grupos já formados. Já as técnicas não-hierárquicas (particionais) são mais flexíveis e possibilitam que elementos amostrais mudem de grupo em cada passo do algoritmo, se essa mudança permitir uma melhor solução em termos de variabilidade da partição resultante. O objetivo básico das técnicas particionais é buscar diretamente uma partição dos dados em k grupos de modo a preservar a coesão interna e isolamento entre elementos. Esse trabalho é baseado em técnicas particionais e no método k-médias. A seguir é apresentado uma revisão básica acerca do método k-médias, e sua versão sob o enfoque de kernelização da métrica.

2.1 Método k-médias

É um dos métodos particionais mais conhecidos e utilizados devido ao fato de ser de fácil implementação e ter baixo custo computacional, já que sua complexidade é de ordem $O(nKl)$. O algoritmo k-médias, aqui rotulado por KMeans, busca minimizar a distância dos elementos a um conjunto de k centros dado por $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k\}$ de forma iterativa.

Seja $\Omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ um conjunto de n observações, onde cada $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ é descrito por p variáveis e seja $\boldsymbol{P} = \{P_1, \ldots, P_K\}$ uma partição de Ω em K grupos disjuntos, onde cada grupo k $(k = 1, \ldots, K)$ é caracterizado por um ponto central $\boldsymbol{y}_k$, conhecido como protótipo ou centróide. Sendo assim, o algoritmo k-médias, baseado na distância euclidiana, busca obter os grupos através da minimização da seguinte função objetivo:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} ||\boldsymbol{x}_i - \boldsymbol{y}_k||^2 \tag{2.1}$$

O centróide $\boldsymbol{y}_k$ que minimiza J é a média aritmética das observações $\boldsymbol{x}_i$ petencentes ao grupo k, ou seja:

$$\boldsymbol{y}_k = \frac{1}{n_k} \sum_{i \in P_k} \boldsymbol{x}_i \ , \tag{2.2}$$

onde n_k é o número de observações pertencentes ao grupo k.

O algoritmo a seguir resume o passo a passo iterativo para obter o agrupamento pelo método k-médias.

Algoritmo 1: MÉTODO k-MÉDIAS

(1) Inicialização

Fixe K (o número de grupos), $2 \leq K < n$; Escolha K observações distintos $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ pertencentes a Ω como centróides iniciais e aloque cada observação i de acordo com o centróide mais próximo $\boldsymbol{y}_h$ ($h = \arg\min_{1 \leq k \leq K} ||\boldsymbol{x}_i - \boldsymbol{y}_k||^2$) para obter a partição inicial $\boldsymbol{P} = \{P_1, \ldots, P_K\}$

(2) Atualização dos centróides

Atualize os centróides dos grupos $\boldsymbol{y}_k$ ($k = 1, \ldots, K$) segundo a Equação (2.2).

(3) Atualização da partição

$test \leftarrow 0$

para $i = 1$ até n faça

defina o grupo vencedor P_h tal que

$h = \arg\min_{1 \leq k \leq K} ||\boldsymbol{x}_i - \boldsymbol{y}_k||^2$

se $i \in P_k$ e $h \neq k$

$test \leftarrow 1$

$P_h \leftarrow P_h \cup \{i\}$

$P_k \leftarrow P_k \setminus \{i\}$

(4) Critério de parada

Se $test = 0$, então pare, caso contrário, volte ao passo (2).

2.2 Conceitos básicos de *kernel*

A partir do desenvolvimento do algoritmo *kernel* k-médias (GIROLAMI, 2002) diversos métodos de agrupamento foram modificados de modo a incorporarem funções *kernels*. Além disso, uma grande variedade de métodos de agrupamento baseados em *kernel* foram propostos (FILIPPONE et al., 2008).

A essência dos métodos baseados em *kernel* envolve a realização de um mapeamento não-linear Φ do espaço original p-dimensional $X \subset \mathbb{R}^p$ para um espaço de dimensão mais alta (possivelmente infinita), chamado espaço de características, $\mathcal{F}$. A razão para passar a

dimensões mais altas é que em tais dimensões pode ser possível obter grupos bem definidos e linearmente separáveis.

Seja $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ um conjunto não vazio, onde $\boldsymbol{x}_i \in \mathbb{R}^p, \forall_i$. Uma função $\mathcal{K} : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ é dita um *kernel* positivo definido (ou *kernel* de Mercer) se, e somente se, $\mathcal{K}$ é simétrica, ou seja, $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_k) = \mathcal{K}(\boldsymbol{x}_k, \boldsymbol{x}_i)$ e a seguinte desigualdade é válida (MERCER, 1909):

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_i c_k \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_k) \geq 0 \quad \forall_n \geq 2, \tag{2.3}$$

onde $c_r \in \mathbb{R}\ \forall r = 1, \cdots, n$.

Seja $\Phi : X \to \mathcal{F}$ um mapeamento não-linear do espaço original X para um espaço de características de alta dimensão $\mathcal{F}$. Aplicando o mapeamento não-linear Φ, o produto interno $\boldsymbol{x}_i^\top \boldsymbol{x}_k$ no espaço original é mapeado para $\Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_k)$ no espaço de características. A essência dos métodos baseados em *kernel* é que o mapeamento não-linear Φ não precisa ser explicitamente especificado porque todo *kernel* de Mercer pode ser expresso pela expressão (2.4), que é usualmente referida como *kernel trick* (MULLER et al., 2001; SCHÖLKOPF; SMOLA; MÜLLER, 1998):

$$\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_k) = \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_k). \tag{2.4}$$

Usualmente, os métodos de agrupamento baseados em *kernel* trabalham com uma matriz $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, chamada de matriz *kernel*, onde cada elemento $k_{il} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_l)$, $i = 1, \cdots, n$, $l = 1, \cdots, n$.

Em consequência da Eq. (2.4), é possível calcular distâncias Euclidianas em $\mathcal{F}$ da seguinte maneira Muller et al. (2001), Schölkopf, Smola e Müller (1998):

$$\begin{aligned} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_k)||^2 &= (\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_k)^\top)(\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_k)) \\ &= \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_i) - 2\Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_k) + \Phi(\boldsymbol{x}_k)^\top \Phi(\boldsymbol{x}_k) \\ &= \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_k) + \mathcal{K}(\boldsymbol{x}_k, \boldsymbol{x}_k). \end{aligned} \tag{2.5}$$

Diversos métodos de agrupamentos têm sido modificados de modo a incorporarem funções *kernel*. Essas modificações estão sendo desenvolvidas sob duas abordagens principais:

1. Kernelização da métrica: os centróides dos grupos são obtidos no espaço original das observações e as distâncias das observações aos centróides dos grupos são calculadas

através de funções *kernel*:

$$||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{y}_k) + \mathcal{K}(\boldsymbol{y}_k, \boldsymbol{y}_k) \tag{2.6}$$

2. Agrupamento no espaço característica: é realizado um mapeamento de cada observação por meio de uma função não-linear Φ e então obtêm-se os centróides dos grupos no espaço de características. Isto é, seja $\boldsymbol{y}_k^{\Phi}$ o centróide do k-ésimo grupo no espaço de características é possível calcular $||\Phi(\boldsymbol{x}_i) - \boldsymbol{y}_k^{\Phi}||^2$ sem a necessidade de se obter $\boldsymbol{y}_k^{\Phi}$ através do uso do *kernel trick* (Eq. (2.4)).

A tabela 2.3 apresenta os principais exemplos de funções *kernel* presentes na literatura.

Tabela 2.3: Exemplos de funções *kernel*.

Tipo da Função *kernel*	$\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_k)$
Linear	$\boldsymbol{x}_i^\top \boldsymbol{x}_k$
Polinomial de grau d	$(\gamma \boldsymbol{x}_i^\top \boldsymbol{x}_k + \theta)^d$, $\gamma > 0$, $\theta \geq 0$, $d \in \mathbb{N}$
Sigmóide	$\tanh(\gamma \boldsymbol{x}_i^\top \boldsymbol{x}_k + \theta)$, $\gamma > 0$, $\theta \geq 0$
Gaussiana	$\exp\left\{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2}{2\sigma^2}\right\}, \sigma > 0$
Laplaciana	$\exp\left\{-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_k\|\right\}, \gamma > 0$

Como mencionado anteriormente, a principal vantagem do uso do *kernel* é que, ao passarmos para um espaço de mais alta dimensão, um conjunto de observações de entrada não linearmente separável pode tornar-se linearmente separável, e, embora não conheçamos o mapeamento não linear Φ e não possamos obter os centróides dos grupos, as distâncias entre as observações e os centróides dos grupos podem ser calculadas através de funções *kernel*.

2.3 Método *Kernel* k-médias sob kernelização da métrica

Seja $\Omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ um conjunto de n observações, onde cada $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ é descrito por p variáveis e seja $\boldsymbol{P} = \{P_1, \ldots, P_K\}$ uma partição de Ω em K grupos disjuntos. A idéia básica do algoritmo *kernel* k-médias baseado na kernelização da métrica, aqui

rotulado por KKMeans, é minimizar a seguinte função objetivo:

$$\begin{aligned} J &= \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 \\ &= \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{y}_k) + \mathcal{K}(\boldsymbol{y}_k, \boldsymbol{y}_k) \end{aligned} \tag{2.7}$$

onde $\boldsymbol{y}_k$ é o centróide do k-ésimo grupo ($k = 1, \cdots, K$).

A derivação dos centróides dos grupos depende da escolha da função *kernel*. Considerando o *kernel* gaussiano, que é o mais comumente utilizado (FERREIRA, 2013), então $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i) = 1$, $\forall i$, e a função objetivo dada em (2.7) pode ser expressa da seguinte forma:

$$J = 2 \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} (1 - \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{y}_k)) \tag{2.8}$$

Derivando a Eq. (2.8) em relação ao centróide do k-ésimo grupo $\boldsymbol{y}_k$ e igualando ao vetor nulo obtem-se a seguinte equação de atualização:

$$\boldsymbol{y}_k = \frac{\sum_{i \in P_k} \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{y}_k) \boldsymbol{x}_i}{\sum_{i \in P_k} \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{y}_k)}, \quad k = 1, \cdots, K. \tag{2.9}$$

Para definir a melhor partição do conjunto de observações de entrada Ω, os centróides dos grupos $\boldsymbol{y}_k$ ($k = 1, \cdots, K$) são mantidos fixos. Os grupos P_k ($k = 1, \cdots, K$), que minimizam a função objetivo dada pela equação (2.8) são atualizados de acordo com a seguinte regra de alocação:

$$\begin{aligned} P_k = \big\{ & i \in \Omega : ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 \leq ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{v}_h)||^2, \\ & \forall h \neq k, h = 1, \cdots, K \big\}. \end{aligned} \tag{2.10}$$

O algoritmo a seguir resume o processo iterativo para obter o agrupamento pelo método *kernel* k-médias baseado em kernelização da métrica.

Algoritmo 2: MÉTODO *kernel* k-MÉDIAS BASEADO NA KERNELIZAÇÃO DA MÉTRICA

(1) Inicialização

Fixe K (o número de grupos), $2 \leq K < n$; escolha aleatoriamente uma partição $\boldsymbol{P}$ de Ω em K observações distintos $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ pertencentes a Ω como centróides iniciais e aloque cada observação i de acordo com o centróide mais próximo $\boldsymbol{y}_h$ ($h = \arg\min_{1 \leq k \leq K} ||\boldsymbol{x}_i - \boldsymbol{y}_k||^2$) para obter a partição inicial $\boldsymbol{P} = \{P_1, \ldots, P_K\}$

(2) Atualização dos centróides

Atualize os centróides dos grupos $\boldsymbol{y}_k$ ($k = 1, \ldots, K$) segundo a Equação ((2.9)) ;

(3) Atualização da partição

$test \leftarrow 0$

para $i = 1$ até n faça

defina o grupo vencedor P_h tal que

$h = \arg\min_{1 \leq k \leq K} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2$

se $i \in P_k$ e $h \neq k$

$test \leftarrow 1$

$P_h \leftarrow P_h \cup \{i\}$

$P_k \leftarrow P_k \setminus \{i\}$

(4) Critério de parada

Se $test = 0$, então pare, caso contrário, volte ao passo (2).

CAPÍTULO 3

Método Proposto

3.1 Introdução

Nesse capítulo, será apresentada a principal contribuição desse trabalho. Encontram-se introduzidos os métodos de agrupamento *hard* baseados no *Kernel* de Mahalanobis com distâncias quadráticas adaptativas. As distâncias quadráticas são definidas por uma matriz simétrica positiva definida $\boldsymbol{M}$ (chamada matriz de covariância), que apresenta dimensão $p \times p$ associada com o k-ésimo grupo ($k = 1, \cdots, K$), sob a restrição de que $\det(\boldsymbol{M}) = 1$. Baseada nessa definição foram definidas matrizes de covarâncias que podem ser única ou diferente para cada grupo. Além disso, as matrizes podem ser completa (distâncias adaptativas locais) ou diagonais (distâncias adaptativas globais).

A ideia principal do método de agrupamento *hard* baseado em uma única distância adaptativa consiste em comparar grupos e seus centróides por meio de uma única distância que muda a cada iteração do algoritmo, mas é a mesma para todos os grupos. No caso do método de agrupamento *hard* baseado em uma distância adaptativa diferente para cada grupo, compara-se os grupos e seus centróides usando uma distância diferente associada a cada grupo que muda em cada iteração do algoritmo, isto é, a distância não é determinada de forma absoluta e é diferente de um grupo para o outro.

Esses métodos adapatativos procuram uma partição de um conjunto de observações em k grupos que correspondem a um conjunto de k centróides $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}$ e uma distância quadrática adaptativa única ou diferente para cada grupo, de modo que um critério que mede o ajuste entre os grupos e seus representantes (centróides) seja minimizado localmente.

Na literatura, o *kernel* Gaussiano é o mais comumente utilizado. De modo geral, esse *kernel* fornece bons resultados e requer somente o ajuste de um parâmetro. Contudo, ele é baseado na distância Euclidiana entre dois parâmetros $\boldsymbol{x}_i$ e $\boldsymbol{x}_k$ em $\mathbf{R}^p$. Como dito anteriormente, a distância Euclidiana fornece bons resultados quando os grupos são, aproximadamente, hiperesféricos e linearmente separáveis, o que significa que cada variável tem a mesma variância e que não existe covariância entre elas. Contudo, sabe-se que observações em dois grupos diferentes são mais prováveis de estarem distribuídas dentro de duas regiões hiperelipsoidais diferentes. Diante da limitação da distância Euclidiana em lidar com grupos de formas hiperelipsoidais, os quais surgem em diversas situações práticas e experimentais, uma melhor escolha é considerar a distância de Mahalanobis.

A distância de Mahalanobis é uma melhor escolha para dados com estrutura complexa, pois leva em consideração as covariâncias entre as variáveis e é invariante em escala. A distância de Mahalanobis entre uma observação $\boldsymbol{x}_i$ e o centróide global $\boldsymbol{y}$ em $\mathbb{R}^p$ é dada por:

$$d^2_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{x}_i, \boldsymbol{y}) = (\boldsymbol{x}_i - \boldsymbol{y})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{y}) \tag{3.1}$$

em que

$$\boldsymbol{y} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \text{ e } \boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{y})(\boldsymbol{x}_i - \boldsymbol{y})^\top. \tag{3.2}$$

Inspirados pela definição de distância de Mahalanobis, é possível definir um *kernel* adaptativo de Mahalanobis substituindo a distância euclidiana no *kernel* Gaussiano por uma distância do tipo Mahalanobis (CAMPS-VALLS et al., 2007).

$$\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{x}_k) = \exp\left\{-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_k)}{2\sigma^2}\right\} \tag{3.3}$$

onde

$$d^2_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_k) = (\boldsymbol{x}_i - \boldsymbol{x}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_k) \tag{3.4}$$

é uma distância quadrática definida por uma matriz simétrica positiva definida $\boldsymbol{M}$. Aqui,

a distância de Mahalanobis é calculada entre duas observações $\boldsymbol{x}_i$ e $\boldsymbol{x}_k$ em $\mathbb{R}^p$ ao invés de uma observação $\boldsymbol{x}_i$ e o centróide global $\boldsymbol{y}$. É possível verificar que o *kernel* de Mahalanobis é uma extensão do gaussiano. Note que, fazendo $\boldsymbol{M} = \boldsymbol{I}$, onde $\boldsymbol{I}$ é a matriz identidade $p \times p$, temos o *kernel* Gaussiano.

Baseados no conceito de distância de Mahalanobis, serão consideradas quatro distâncias quadráticas adaptativas entre uma observação $\boldsymbol{x}_i$ e o centróide do k-ésimo grupo $\boldsymbol{y}_k$, definidas como:

i) Distância quadrática adaptativa definida por uma matriz de covariâncias completa e única para todos os grupos:

$$\boldsymbol{M}_k = \boldsymbol{M} : d^2_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{y}_k) = (\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k), \quad k = 1, \cdots, K. \tag{3.5}$$

ii) Distância quadrática adaptativa definida por uma matriz de covariâncias diagonal e única para todos os grupos:

$$\begin{aligned} \boldsymbol{M}_k = \boldsymbol{M} = diag\{\lambda_1, \cdots, \lambda_p\} : d^2_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{y}_k) &= (\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k) \\ = \Sigma_{j=1}^p \lambda_j (\boldsymbol{x}_{ij} - \boldsymbol{y}_{kj})^2, &\quad k = 1, \cdots, K. \end{aligned} \tag{3.6}$$

iii) Distância quadrática adaptativa definida por uma matriz de covariâncias completa e diferente para cada grupo:

$$\boldsymbol{M}_k : d^2_{\boldsymbol{M}_k}(\boldsymbol{x}_i, \boldsymbol{y}_k) = (\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}_k(\boldsymbol{x}_i - \boldsymbol{y}_k), \quad k = 1, \cdots, K. \tag{3.7}$$

iv) Distância quadrática adaptativa definida por uma matriz de covariâncias diagonal e diferente para cada grupo:

$$\begin{aligned} \boldsymbol{M}_k = diag\{\lambda_1^k, \cdots, \lambda_p^k\} : d^2_{\boldsymbol{M}_k}(\boldsymbol{x}_i, \boldsymbol{y}_k) &= (\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}_k(\boldsymbol{x}_i - \boldsymbol{y}_k) \\ = \Sigma_{j=1}^p \lambda_j^k (\boldsymbol{x}_{ij} - \boldsymbol{y}_{kj})^2, &\quad k = 1, \cdots, K. \end{aligned} \tag{3.8}$$

3.2 Método de agrupamento *hard* baseado no *kernel* de Mahalanobis sob o enfoque da kernelização da métrica

Nessa seção são apresentados os métodos de agrupamentos *hard* baseados no *kernel* de Mahalanobis com distâncias quadráticas adaptativas sob o enfoque da kernelização da métrica. Os algoritmos (aqui definidos como AMKKMeansGF-K, AMKKMeansGD-K, AMKKMeansLF-K e AMKKMeansLD-K), respectivamente, de acordo com as distâncias citadas na seção anterior, otimizam uma função objetivo J que mede o ajuste entre os grupos e os seus centróides, a qual pode ser definida da seguinte forma:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \varphi^2(\boldsymbol{x}_i, \boldsymbol{y}_k) \tag{3.9}$$

onde $\varphi^2(\boldsymbol{x}_i, \boldsymbol{y}_k)$ é uma medida de distância adequada (distâncias definidas anteriormente pelas equações (3.5), (3.6), (3.7) e (3.8)) entre uma observação $\boldsymbol{x}_i$ e o centróide do k-ésimo grupo $\boldsymbol{y}_k$ calculada por meio de uma função *kernel* de Mahalanobis. Considerando as medidas de distâncias propostas neste capítulo, temos:

i) Método de agrupamento *hard* baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas definida por uma matriz de covariância completa e única para todos os grupos (AMKKMeansGF-K):

$$\begin{aligned} J1 &= \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 \\ &= 2\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left(1 - \exp\left\{-\frac{(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k)}{2\sigma^2}\right\}\right) \\ &= 2\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left(1 - \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)\right) \end{aligned} \tag{3.10}$$

ii) Método de agrupamento *hard* baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas definida por uma matriz de covariância diagonal e única para todos os grupos (AMKKMeansGD-K):

$$
\begin{aligned}
J2 &= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k} ||\Phi(\boldsymbol{x}_i)-\Phi(\boldsymbol{y}_k)||^2 \\
&= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k}\left(1-\exp\left\{-\frac{(\boldsymbol{x}_i-\boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{y}_k)}{2\sigma^2}\right\}\right) \\
&= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k}\left(1-\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k)\right)
\end{aligned} \tag{3.11}
$$

em que $\boldsymbol{M}$ é definida por:

$$
\boldsymbol{M} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}
$$

iii) Método de agrupamento *hard* baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas definida por uma matriz de covariância completa e distinta para os grupos (AMKKMeansLF-K):

$$
\begin{aligned}
J3 &= \sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k} ||\Phi(\boldsymbol{x}_i)-\Phi(\boldsymbol{y}_k)||^2 \\
&= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k}\left(1-\exp\left\{-\frac{(\boldsymbol{x}_i-\boldsymbol{y}_k)^\top \boldsymbol{M_k}(\boldsymbol{x}_i-\boldsymbol{y}_k)}{2\sigma^2}\right\}\right) \\
&= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k}\left(1-\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k)\right)
\end{aligned} \tag{3.12}
$$

iv) Método de agrupamento *hard* baseado no *kernel* de Mahalanobis com distâncias quadráticas adaptativas definida por uma matriz de covariância diagonal e diferente para os grupos (AMKKMeansLD-K):

$$
\begin{aligned}
J4 &= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k} ||\Phi(\boldsymbol{x}_i)-\Phi(\boldsymbol{y}_k)||^2 \\
&= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k} \left(1-\exp\left\{-\frac{(\boldsymbol{x}_i-\boldsymbol{y}_k)^\top \boldsymbol{M_k}(\boldsymbol{x}_i-\boldsymbol{y}_k)}{2\sigma^2}\right\}\right) \\
&= 2\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in P_k} \left(1-\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k)\right)
\end{aligned} \tag{3.13}
$$

em que $\boldsymbol{M_k}$ é definida por:

$$
\boldsymbol{M_k} = \begin{pmatrix} \lambda_{1k} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{kK} \end{pmatrix}
$$

Para o desenvolvimento do método de agrupamento *hard* baseado no *kernel* de Mahalanobis o algoritmo foi formulado em três etapas A primeira consiste em calcular os centróides dos grupos. Nesse passo a partição e a matriz $\boldsymbol{M}$ são mantidas fixas.

Proposição 3.1. *Qualquer que seja a medida de distância quadrática adaptativa dada pelas equações (3.5),(3.6),(3.7) e (3.8), e considerando o* kernel *de Mahalanobis ($\mathcal{K}^{(m)}$), o centróide $\boldsymbol{y}_k$ que minimiza a função objetivo J dada na equação* (3.9) *é calculado da seguinte forma:*

$$
\boldsymbol{y}_k = \frac{\sum_{\boldsymbol{x}_i\in P_k}\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k)\boldsymbol{x}_i}{\sum_{\boldsymbol{x}_i\in P_k}\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k)}, \quad k=1,\ldots,K \tag{3.14}
$$

Demonstração. Considerando o *kernel* de Mahalanobis, temos $\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{x}_i) = 1, (i = 1,\cdots,k)$ e $||\Phi(\boldsymbol{x}_i)-\Phi(\boldsymbol{y}_k)||^2 = 2\left\{1-\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k)\right\}$. Sendo assim, para o k-ésimo grupo o problema se torna encontrar o componente $\boldsymbol{y}_k$ que maximiza o seguinte termo:

$$
\sum_{\boldsymbol{x}_i\in P_k}\mathcal{K}^{(m)}(\boldsymbol{x}_i,\boldsymbol{y}_k) \tag{3.15}
$$

Derivando o termo (3.15) em relação a $\boldsymbol{y}_k$ temos que:

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{y}_k} &= \sum_{\boldsymbol{x}_i \in P_k} \frac{\partial \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)}{\partial \boldsymbol{y}_k} \\ &= \sum_{\boldsymbol{x}_i \in P_k} \frac{\partial \mathcal{K}^{(m)} \exp\left\{-\frac{(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k)}{2\sigma^2}\right\}}{\partial \boldsymbol{y}_k} \\ &= \left(-\frac{1}{2\sigma^2}\right) \sum_{\boldsymbol{x}_i \in P_k} \exp\left\{-\frac{(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k)}{2\sigma^2}\right\} (-2\boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k))\end{aligned}$$

Reescrevendo a expressão acima e igualando ao vetor nulo, temos, para $(k = 1, \cdots, K)$ o seguinte resultado:

$$\begin{aligned}&\left(-\frac{1}{2\sigma^2}\right) \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k) \, (-2\boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k)) = 0 \\ &\left(\frac{1}{\sigma^2}\right) \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k) \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k) = 0 \\ &\left(\frac{1}{\sigma^2}\right) \boldsymbol{M} \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k) \boldsymbol{x}_i = \left(\frac{1}{\sigma^2}\right) \boldsymbol{M} \boldsymbol{y}_k \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)\end{aligned}$$

Multiplicando ambos os lados por $\boldsymbol{M}^{-1}$ obtemos a expressão para atualização dos centróides:

$$\boldsymbol{y}_k = \frac{\sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k) \boldsymbol{x}_i}{\sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)}$$

□

A segunda etapa é referente ao cálculo das matrizes de covariância $\boldsymbol{M}$, correspondente a cada distância adaptativa proposta nessa monografia. Nessa etapa, a partição e os centróides $\boldsymbol{y}_k$ dos grupos correspondentes $P_k (k = 1, \cdots, K)$ são mantidos fixos. Para o cálculo da matriz é necessário encontrar as expressões que minimizam os seguintes critérios de agrupamento das equações (3.10), (3.11), (3.12) e (3.13).

Proposição 3.2. *As matrizes simétricas positivas definidas* $\boldsymbol{M}$*, que minimizam a função objetivo* J *dada pela equação* (3.9) *com a restrição de* $\det(\boldsymbol{M}) = 1$*, são atualizadas de acordo com as seguintes expressões:*

i) Caso onde a matriz é a mesma para todos os grupos e não diagonal, definida pelo algoritmo AMKKMeansGF-K:

$$\begin{aligned} \boldsymbol{M} &= [\det(\boldsymbol{Q})]^{\frac{1}{p}}\boldsymbol{Q}^{-1}, \text{ em que } \boldsymbol{Q} = \sum_{k=1}^{K}\boldsymbol{Q}_k \text{ e} \\ \boldsymbol{Q}_k &= \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \end{aligned} \tag{3.16}$$

ii) Caso onde a matriz é a mesma para todos os grupos e diagonal, definida pelo algoritmo AMKKMeansGD-K:

$$\begin{aligned} \boldsymbol{M} &= [\det(\boldsymbol{Q})]^{\frac{1}{p}}\boldsymbol{Q}^{-1}, \text{ em que } \boldsymbol{Q} = \sum_{k=1}^{K} diag(\boldsymbol{Q}_k) \text{ e} \\ \boldsymbol{Q}_k &= \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \end{aligned} \tag{3.17}$$

iii) Caso onde a matriz é diferente para cada grupo e não diagonal, definida pelo algoritmo AMKKMeansLF-K:

$$\begin{aligned} \boldsymbol{M}_k &= [\det(\boldsymbol{Q_k})]^{\frac{1}{p}}\boldsymbol{Q_k}^{-1}, \text{ em que} \\ \boldsymbol{Q}_k &= \sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \end{aligned} \tag{3.18}$$

iv) Caso onde a matriz é diferente para cada grupo e diagonal, definida pelo algoritmo AMKKMeansLD-K:

$$\begin{aligned} \boldsymbol{M}_k &= [\det(\boldsymbol{Q}_k)]^{\frac{1}{p}}\boldsymbol{Q}_k^{-1}, \text{ em que} \\ \boldsymbol{Q}_k &= diag\left(\sum_{\boldsymbol{x}_i \in P_k} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top\right) \end{aligned} \tag{3.19}$$

Demonstração. Serão demonstradas as equações para o primeiro caso **i)**, referente ao algoritmo AMKKMeansGF, em que as matrizes de covariância são as mesmas para todos os grupos e completas.

A função objetivo a ser minimizada é dada por:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 \tag{3.20}$$

Expandindo $||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2$, temos:

$$\begin{aligned} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 =& (\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k))^\top (\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)) \\ =& \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{y}_k) - \Phi(\boldsymbol{y}_k)^\top \Phi(\boldsymbol{x}_i) + \Phi(\boldsymbol{y}_k)^\top \Phi(\boldsymbol{y}_k) \\ =& \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_i) - 2\Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{v}_k) + \Phi(\boldsymbol{y}_k)^\top \Phi(\boldsymbol{y}_k) \end{aligned}$$

Segundo a definição de funções *kernels*, o seguinte resultado é válido:

$$\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_i) \tag{3.21}$$

Dessa forma, podemos reescrever a função objetivo dada na equação (3.20) da seguinte forma:

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \{\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{y}_k) + \mathcal{K}(\boldsymbol{y}_k, \boldsymbol{y}_i)\} \tag{3.22}$$

Seja $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_i)$ o *kernel* de Mahalanobis. Então:

- $\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{x}_i) = 1$
- $\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k) = \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{y}_k)\right\}$

Substituindo na equação (3.22), temos que a função objetivo a ser minimizada, considerando o *kernel* de Mahalanobis é:

$$
\begin{aligned}
J &= \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left\{1 - 2\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k) + 1\right\} \\
&= \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} (2 - 2\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)) \\
&= \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} 2(1 - \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)) \\
&= 2\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} (1 - \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)) \\
&= 2\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left(1 - \exp\left\{-\frac{(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{y}_k)}{2\sigma^2}\right\}\right)
\end{aligned}
\tag{3.23}
$$

Para encontrar a matriz de covariância $\boldsymbol{M}$ foi considerada a restrição de que $det(\boldsymbol{M}) = 1$ e aplicado multiplicadores de Lagrange.

Seja $g(\boldsymbol{M}) = 1 - det(\boldsymbol{M}) \Rightarrow g(\boldsymbol{M}) = 0$

$$
\begin{aligned}
L(\boldsymbol{M}, \alpha) =& J(\boldsymbol{M}) + \alpha g(\boldsymbol{M}) \\
=& \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} (2 - 2\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)) + \alpha(1 - \det(\boldsymbol{M}))
\end{aligned}
\tag{3.24}
$$

Derivando a Eq. (3.24) em relação a $\boldsymbol{M}$, obtemos:

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{M}} =& \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \frac{\partial(2 - 2\mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)}{\partial \boldsymbol{M}} + \alpha \frac{\partial(1 - \det(\boldsymbol{M}))}{\partial \boldsymbol{M}} \\
=& \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left\{\frac{1}{\sigma^2} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top (\boldsymbol{x}_i - \boldsymbol{y}_k)\right\} - \alpha \det(\boldsymbol{M}) \boldsymbol{M}^{-1}
\end{aligned}
\tag{3.25}
$$

Fazendo $\frac{\partial L}{\partial \boldsymbol{M}} = 0$, temos:

$$\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left\{ \frac{1}{\sigma^2} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top (\boldsymbol{x}_i - \boldsymbol{y}_k) \right\} = \alpha \det(\boldsymbol{M}) \boldsymbol{M}^{-1} \tag{3.26}$$

Considerando a restrição de que o $\det(\boldsymbol{M}) = 1$, seguem os resultados:

$$\begin{aligned}
\alpha \boldsymbol{M}^{-1} =& \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left\{ \frac{1}{\sigma^2} \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top (\boldsymbol{x}_i - \boldsymbol{y}_k) \right\} \\
\boldsymbol{M}^{-1} =& \frac{1}{\sigma^2 \alpha} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in P_k} \left\{ \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top (\boldsymbol{x}_i - \boldsymbol{y}_k) \right\} \\
\boldsymbol{M}^{-1} =& \frac{1}{\sigma^2 \alpha} \sum_{k=1}^{K} \boldsymbol{Q}_K \text{ em que} \\
\boldsymbol{Q}_k =& \sum_{\boldsymbol{x}_i \in P_k} \left\{ \mathcal{K}^{(m)}(\boldsymbol{x}_i, \boldsymbol{y}_k)(\boldsymbol{x}_i - \boldsymbol{y}_k)^\top (\boldsymbol{x}_i - \boldsymbol{y}_k) \right\}
\end{aligned} \tag{3.27}$$

Dessa forma,

$$\boldsymbol{M}^{-1} = \frac{1}{\sigma^2 \alpha} \boldsymbol{Q} \quad \text{em que} \quad \boldsymbol{Q} = \sum_{k=1}^{K} \boldsymbol{Q}_k \tag{3.28}$$

Portanto,

$$\det(\boldsymbol{M}^{-1}) = \frac{\det(\boldsymbol{Q})}{\sigma^{2p} \alpha^p} = 1, \quad \text{pois} \quad \det(\boldsymbol{M}^{-1}) = \frac{1}{det(\boldsymbol{M})} \tag{3.29}$$

$$\begin{aligned}
\Rightarrow \alpha^p =& \frac{\det(\boldsymbol{Q})}{\sigma^{2p}} \\
\alpha =& \sqrt[p]{\frac{\det((Q))}{\sigma^{2p}}} \\
=& \frac{1}{\sigma^2} [\det(\boldsymbol{Q})]^{1/p}
\end{aligned} \tag{3.30}$$

Substituindo o α encontrado ((3.30)) na equação (3.28) encontramos a seguinte ex-

pressão para o cálculo da matriz $\boldsymbol{M}$:

$$\begin{aligned} \boldsymbol{M}^{-1} =& \frac{\boldsymbol{Q}}{[\det(\boldsymbol{Q})]^{1/p}} \\ \Rightarrow \boldsymbol{M} =& \frac{1}{\boldsymbol{M}^{-1}} \\ \boldsymbol{M} =& \frac{[\det(\boldsymbol{Q})]^{1/p}}{\boldsymbol{Q}} \\ \boldsymbol{M} =& [\det(\boldsymbol{Q})]^{1/p}\boldsymbol{Q}^{-1} \end{aligned} \tag{3.31}$$

□

Para os casos **ii)**, **iii)** e **iv)** os resultados seguem de forma análoga a demonstração para **i)**. Para os casos **iii)** e **iv)**, em que as matrizes são diferentes para cada grupo, basta considerar a restrinção de que $\det(\boldsymbol{M}_k) = 1$.

A terceira e última etapa consiste na alocação das n observações aos k grupos, nesse passo os centróides dos grupos $\boldsymbol{y}_k$ $(k = 1, \cdots, K)$ e a matriz $\boldsymbol{M}$ são mantidos fixos. Os grupos $P_k(k = 1, \cdots, K)$, que minimizam o critério J de agrupamento, são atualizados de acordo com a seguinte regra de alocação:

$$\begin{aligned} P_k = \big\{ & \boldsymbol{x}_i \in X : ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2 \leq ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_h)||^2, \\ & \forall h \neq k, h = 1, \cdots, K\big\} \end{aligned} \tag{3.32}$$

O algoritmo a seguir resume o passo a passo iterativo para obter o agrupamento pelos métodos baseado em um *kernel* adaptativo de Mahalanobis propostos nessa seção.

Algoritmo 3: Método baseado no *kernel* de Mahalanobis sob enfoque da kernelização da métrica

---

(1) Inicialização

Fixe K (o número de grupos), $2 \leq K < n$; escolha aleatoriamente uma partição inicial $\boldsymbol{P}$ de Ω em K grupos $P_1, \ldots, P_K$ ou, alternativamente, escolha K observações distintas $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ pertencentes a Ω como os centróides iniciais e aloque cada observação i ao centróide mais próximo $\boldsymbol{y}_h$ ($h = \arg\min_{1 \leq k \leq K} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2$ para obter a partição inicial $\boldsymbol{P} = \{P_1, \ldots, P_K\}$

(2) Definição dos melhores centróides

Calcular os centróides dos grupos $\boldsymbol{y}_k$ ($k = 1, \ldots, K$) de acordo com a Eq. (3.14).

(3) Cálculo da matriz $\boldsymbol{M}$

Calcular a matriz $\boldsymbol{M}$ de acordo com as equações. (3.16), (3.17), (3.18) ou (3.19).

(4) Definição da melhor partição

$test \leftarrow 0$

para $i = 1$ até n faça

defina o grupo vencedor P_h tal que

$h = \arg\min_{1 \leq k \leq K} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{y}_k)||^2$

se $i \in P_k$ e $h \neq k$

$test \leftarrow 1$

$P_h \leftarrow P_h \cup \{i\}$

$P_k \leftarrow P_k \setminus \{i\}$

(5) Critério de parada

Se $test = 0$, então, PARE, caso contrário, volte ao passo (2).

CAPÍTULO 4

Avaliação Experimental

4.1 Introdução

Neste capítulo é apresentado o esquema de avaliação experimental dos métodos de agrupamentos propostos. Os métodos propostos serão comparados a métodos clássicos de agrupamento como o k-médias e sua versão baseada no *kernel* gaussiano. Foi realizada uma simulação de Monte Carlo considerando quatro configurações de conjunto de dados simulados e 5 conjuntos de dados reais foram considerados. Todos os experimentos realizados nesse trabalho foram feitos utilizando a linguagem de programação R (R Core Team, 2013).

4.1.1 Conjunto de dados simulados

Cada conjunto de dados sintéticos foi simulado considerando classes com tamanhos e formas diferentes. Os conjuntos de dados simulados têm 500 pontos cada, divididos em quatro classes com tamanhos (n = 200, 150, 50 e 100), respectivamente. Cada classe desses dados foi desenhada de acordo com uma distribuição normal bivariada com vetor

médio e matriz de covariância representada por

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ e } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{4.1}$$

Foram consideradas quatro configurações de conjuntos de dados simulados:

1. As matrizes de covariância de classe são diagonais e aproximadamente iguais;
2. As matrizes de covariância são diagonais e diferentes para cada grupo;
3. As matrizes de covariância de classe são completas e aproximadamente iguais;
4. As matrizes de covariância de classe são completas e diferentes para cada grupo.

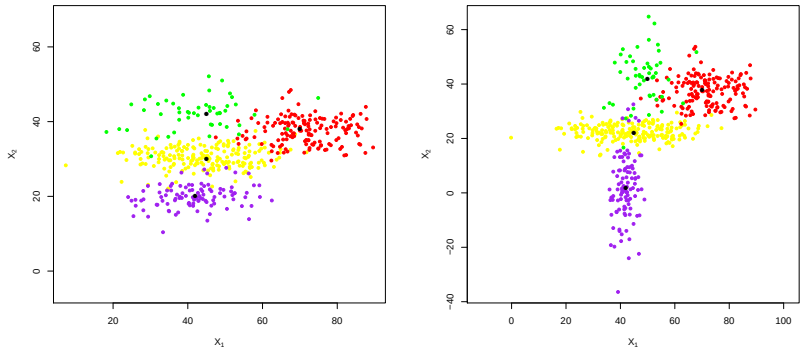
Os algoritmos foram aplicados aos conjuntos de dados sintéticos 1, 2, 3 e 4 para obtenção de 4 partições de grupo. As partições dos 4 grupos fornecidas pelos algoritmos de agrupamento foram comparadas com a partição a *priori* conhecida de 4 classes.

A Tabela 4.1 apresenta as quatro configurações dos dados simulados para cada classe e seus respectivos parâmetros, bem como os valores dos vetores de média e as matrizes de covariância (formadas pelas variâncias e as correlações de cada classe) para cada conjunto de dados simulados.

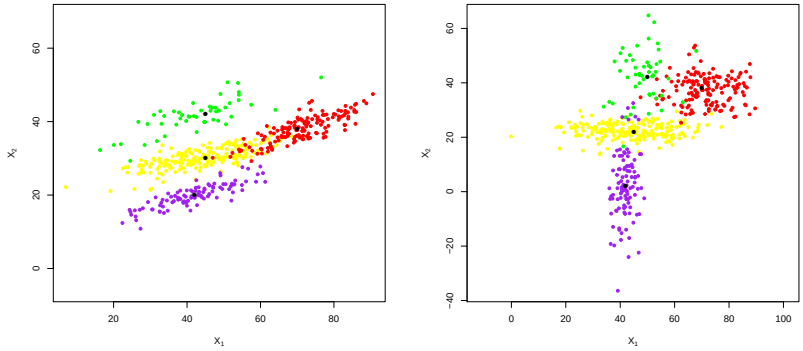
Tabela 4.1: Configurações dos cenários

Cenários	Classes	μ_1	μ_2	σ_1	σ_2	ρ
1	1	45	30	100	9	0.0
	2	70	38	81	16	0.0
	3	45	42	100	16	0.0
	4	42	20	81	9	0.0
2	1	45	22	144	9	0.0
	2	70	38	81	36	0.0
	3	50	42	36	81	0.0
	4	42	2	9	144	0.0
3	1	45	30	100	9	0.7
	2	70	38	81	16	0.8
	3	45	42	100	16	0.7
	4	42	20	81	9	0.8
4	1	45	22	144	9	0.7
	2	70	38	81	36	0.8
	3	50	42	36	81	0.7
	4	42	2	9	144	0.8

A Figura 4.1 mostra as configurações para os quatro conjunto de dados, em que consideramos: (a) Matrizes de covariância diagonais e globais (quase iguais), (b) Matrizes de covariância diagonais e locais (diferente para cada grupo), (c) Matrizes de covariância não diagonais (completas) e globais e (d) Matrizes de covariância não diagonais e locais.



(a) Cenário 1 (b) Cenário 2



(c) Cenário 3 (d) Cenário 4

Figura 4.1: Configuração dos cenários

4.1.2 Conjuntos de dados reais

Foram também realizados experimentos com cinco conjuntos de dados reais obtidos do Repositório de Aprendizagem de Máquina da Universidade da Califórnia, denominados aqui como *Wireless Indoor Localization, Breast Cancer, Banknote Authentication, Iris* e *Abalone*. A descrição de cada conjunto de dados é apresentada a seguir.

Conjunto de dados: Wireless Indoor Localization

O conjunto de dados *Wireless Indoor Localization* [1] consiste em uma coleta para realizar experiências sobre como os pontos fortes do sinal wifi podem ser usados para determinar um dos locais internos. Cada variável é a força do sinal wifi observada no smartphone. O banco de dados contém 2.000 observações e 7 variáveis.

Conjunto de dados: Breast Cancer

O conjunto de dados *Breast Cancer* [2] é um dos três domínios fornecidos pelo *Oncology Institute*. Esse conjunto de dados apresenta 286 observações e duas classes: eventos sem recorrência e eventos de recorrência. Do total de observações, 201 são da classe sem recorrência e 85 da classe com recorrência. Contém 9 variáveis: classe, idade, menopausa, tamanho do tumor, nós inv, caps do nó, deg-malig, mama e quadra da mama.

Conjunto de dados: Banknote Authentication

No conjunto de dados *Banknote Authentication* [3] os dados foram extraídos de imagens tiradas de espécimes genuínos e forjados do tipo notas de banco. Para a digitalização, foi usada uma câmera industrial normalmente usada para inspeção de impressão. As imagens finais têm 400x400 pixels. Devido à lente do objeto e à distância do objeto investigado, foram obtidas imagens em escala de cinza com uma resolução de cerca de 660 dpi. Uma ferramenta Wavelet Transform foi usada para extrair recursos dessas imagens. O banco de dados contém 1.372 observações e 5 variáveis: variação da imagem transformada Wavelet (contínua), distorção da imagem Wavelet Transformed (contínua), curtose da imagem transformada Wavelet (contínua), entropia da imagem (contínua) e classe (inteiro).

[1] https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization
[2] https://archive.ics.uci.edu/ml/datasets/breast+cancer
[3] https://archive.ics.uci.edu/ml/datasets/banknote+authentication

Conjunto de dados: Iris

o conjunto de dados *Iris* [4] contém 3 classes com 50 observações cada. Uma classe é linearmente separável das outras duas; estes não são linearmente separáveis um do outro. Cada classe se refere a um tipo de planta de íris: íris setosa, íris versicolor e íris virgínica. O banco apresenta ao todo 150 observações e 5 variáveis: comprimento da sépala, largura sépala, comprimento da pétala, largura da pétala e a variável referente as classes.

Conjunto de dados: Abalone

O conjunto de dados *Abalone* [5] foi construído para prever a idade do abalone a partir de medições físicas. A idade do abalone é determinada cortando a concha através do cone, manchando-a e contando o número de anéis através de um microscópio - uma tarefa chata e demorada. Outras medidas, mais fáceis de obter, são usadas para prever a idade, Informações adicionais, como padrões climáticos e localização, podem ser necessárias para resolver o problema. A partir dos dados originais, foram removidos exemplos com valores ausentes (a maioria com o valor previsto ausente) e os intervalos dos valores contínuos foram redimensionados para uso com uma RNA (dividindo por 200). Apresenta 4.177 observações e 8 variáveis: Sexo (nominal), Medição de comprimento (contínua), Diâmetro (contínuo), Altura (contínua), Peso total (contínuo), Peso com casca (contínuo), Peso das vísceras (contínuo), Peso da casca (contínuo) e Anéis (número inteiro).

4.2 Índices de avaliação

Para avaliar o desempenho dos diferentes algoritmos de agrupamento estudados e compará-los, foi adotado que a partição a *priori* é conhecida e duas medidas de avaliação externa foram consideradas: o Índice Corrigido de Rand (CR) (HUBERT; ARABIE, 1985) e a Taxa Total de Erro de Classificação (OERC) (BREIMAN et al., 1984).

Seja $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_i, \ldots, \mathcal{P}_c\}$ a partição a *priori* de $\Omega = \{1, \ldots, n\}$ em c classes e seja $P = \{P_1, \ldots, P_k, \ldots, P_K\}$ uma partição rígida de $\Omega = \{1, \ldots, n\}$ em K grupos fornecidos por um algoritmo de agrupamento. As quantidades n_{ik}, $i = 1, \ldots, c$, $k = 1, \ldots, K$, representam o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k e podem ser representadas na forma da Tabela 4.2, denominada matriz de confusão.

[4] https://archive.ics.uci.edu/ml/datasets/Iris
[5] https://archive.ics.uci.edu/ml/datasets/Abalone

Tabela 4.2: Matriz de confusão.

	Grupos					
Classes	P_1	$\cdots$	P_k	$\cdots$	P_K	$\sum$
$\mathcal{P}_1$	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}	$n_{1\bullet} = \sum_{k=1}^{K} n_{1k}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\mathcal{P}_i$	n_{i1}	$\cdots$	n_{ik}	$\cdots$	n_{iK}	$n_{i\bullet} = \sum_{k=1}^{K} n_{ik}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\mathcal{P}_c$	n_{c1}	$\cdots$	n_{ck}	$\cdots$	n_{cK}	$n_{c\bullet} = \sum_{k=1}^{K} n_{ck}$
$\sum$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$\cdots$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$\cdots$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$n = \sum_{i=1}^{c} \sum_{k=1}^{K} n_{ik}$

O Índice Corrigido de Rand (CR) é obtido como

$$CR = \frac{\sum_{i=1}^{c} \sum_{k=1}^{K} \binom{n_{ik}}{2} - \binom{n}{2}^{-1} \sum_{i=1}^{c} \binom{n_{i\bullet}}{2} \sum_{k=1}^{K} \binom{n_{\bullet k}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{c} \binom{n_{i\bullet}}{2} + \sum_{k=1}^{K} \binom{n_{\bullet k}}{2}\right] - \binom{n}{2}^{-1} \sum_{i=1}^{c} \binom{n_{i\bullet}}{2} \sum_{k=1}^{K} \binom{n_{\bullet k}}{2}}, \tag{4.2}$$

onde $\binom{n}{2} = \frac{n(n-1)}{2}$, n_{ik} representa o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k, $n_{i\bullet}$ representa o número de observações na classe $\mathcal{P}_i$, $n_{\bullet k}$ representa o número de observações no grupo P_k, e n é o número total de observações no conjunto de dados.

O CR avalia o grau de concordância (similaridade) entre uma partição *a priori* e uma partição fornecida por um método de agrupamento. Além disso, o CR não é sensível ao número de classes nas partições ou à distribuição das observações nos grupos. Finalmente, o CR assume valores no intervalo $[-1, 1]$, no qual o valor 1 indica concordância perfeita entre as partições, enquanto que valores próximos de zero ou negativos correspondem a concordância entre partições encontrada ao acaso (MILLIGAN; COOPER, 1986).

Em problemas de classificação, cada grupo P_k é associado a uma classe *a priori* $\mathcal{P}_i$ e esta associação deve ser interpretada como se a verdadeira classe *a priori* fosse $\mathcal{P}_i$. Dessa forma, para uma observação pertencente a um dado grupo P_k a decisão está correta se a classe *a priori* dessa observação é $\mathcal{P}_i$. Para obter uma taxa de erro de classificação mínima, precisamos encontrar uma regra de decisão que minimize a probabilidade de erro.

Seja $\ell(\mathcal{P}_i, P_k)$ a probabilidade *a posteriori* de que uma observação pertença à classe $\mathcal{P}_i$ quando associado ao grupo P_k. Seja $\ell(P_k)$ a probabilidade de que a observação pertença ao grupo P_k. A função ℓ é conhecida como função de verossimilhança.

A estimativa da máxima probabilidade *a posteriori* é a moda da probabilidade *a posteriori* $\ell(\mathcal{P}_i, P_k)$ e o índice da classe *a priori* associada a esta moda é dada por

$$MAP(P_k) = \arg\max_{1 \leq i \leq c} \ell(\mathcal{P}_i, P_k).$$

A regra de decisão de Bayes que minimiza a probabilidade média de erro é selecionar a classe *a priori* que maximiza a probabilidade *a posteriori*. A taxa de erro de alocação do grupo P_k é igual a $1 - \ell(\mathcal{P}_{MAP(P_k)}/P_k)$ e a taxa total de erro de classificação (OERC) é igual a

$$OERC = \sum_{k=1}^{K} \ell(P_k)(1 - \ell(\mathcal{P}_{MAP(P_k)}/P_k)).$$

Para uma amostra,

$$\ell(\mathcal{P}_{MAP(P_k)}/P_k) = \max_{1 \leq i \leq c} n_{ik}/n_{\bullet k}.$$

A taxa total de erro de classificação (OERC) foi concebida de modo a medir a habilidade de um algoritmo de agrupamento encontrar as classes *a priori* presentes em um conjunto de dados e é calculada da forma:

$$OERC = \sum_{k=1}^{K} \frac{n_{\bullet k}}{n} \left(1 - \max_{1 \leq i \leq c} n_{ik}/n_{\bullet k}\right) = 1 - \frac{\sum_{k=1}^{K} \max_{1 \leq i \leq c} n_{ik}}{n}. \tag{4.3}$$

O índice OERC assume valores no intervalo $[0, 1]$, no qual valores próximos de zero indicam maior habilidade de um algoritmo na detecção de classes *a priori*.

4.3 Resultados

Os algoritmos de agrupamentos discutidos nesse trabalho foram aplicados aos conjuntos de dados simulados e reais descritos anteriormente. Em cada aplicação foram calculados os índices CR e OERC e através destes ocorreu a comparação entre os métodos.

4.3.1 Avaliação com dados simulados

Para cada conjunto de dados sintéticos, os índices CR e OERC foram calculados no esquema de uma simulação de Monte Carlo com 100 repetições. Cada replicação fornece uma matriz de confusão, em que os grupos fornecidos pelo algoritmo de agrupamento são colocados em linhas e as classes a *priori* são colocadas em colunas.

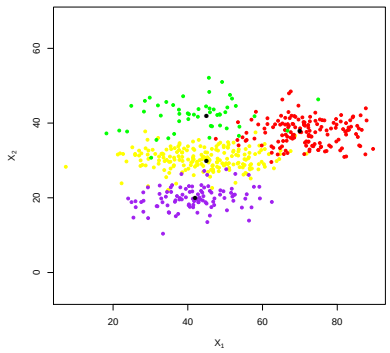
Em cada replicação, os algoritmos de agrupamento foram executados 100 vezes até a convergência para um valor estacionário do critério de adequação, depois disso, o melhor resultado para cada algoritmo de agrupamento foi selecionado de acordo com o critério de adequação. A média e o desvio padrão destas medidas foram computados. O termo $2\sigma^2$ nos *kernels* gaussiano e adaptáveis de Mahalanobis foi estimado como a média dos quantis 0,1 e 0,9 de $||\boldsymbol{x}_i - \boldsymbol{x}_k||^2, i \neq k$ segundo Caputo et al. (2001). Para todos os conjuntos de dados simulados, os algoritmos que obtiveram os melhores desempenhos são os que consideram o *kernel* de Mahalanobis adaptativo.

A tabela 4.3 apresenta o desempenho dos algoritmos KMeans, KKMeans, AMKK-MeansGD, AMKKMeansGF, AMKKMeansLD, AMKKMeansLF em valores de média e desvio padrão para o conjunto de dados sintéticos 1, de acordo com os índices CR e OERC. Nota-se que o algoritmo AMKKMeansGD apresenta melhor desempenho, pois tem o Índice Corrigido de Rand (0.717) e o menor erro de taxa total de classificação (0.129), portanto os dados são melhores explicados quando as matrizes de covariância são aproximadamente iguais e diagonais.

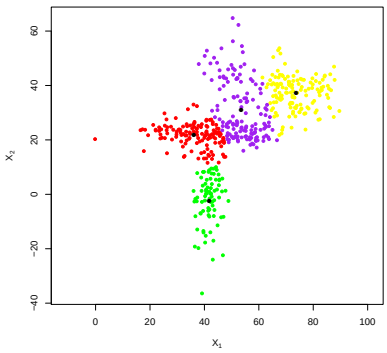
Tabela 4.3: Desempenho dos algoritmos no conjunto de dados sintéticos 1: média e desvio-padrão(entre parênteses) dos índices CR e OERC.

	CR	OERC
KMeans	0.341 (0.053)	0.388 (0.045)
KKMeans	0.346 (0.058)	0.385 (0.047)
AMKKMeansGD	**0.717** (**0.050**)	**0.129** (**0.035**)
AMKKMeansGF	0.652 (0.111)	0.174 (0.074)
AMKKMeansLD	0.692 (0.091)	0.143 (0.067)
AMKKMeansLF	0.638 (0.096)	0.183 (0.071)

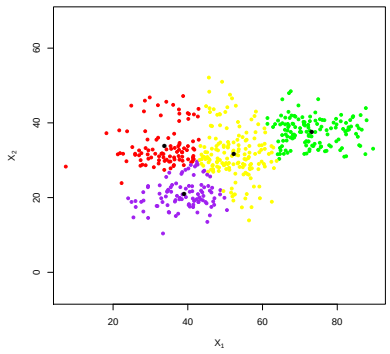
As Figuras 4.2(b), 4.2(c), 4.2(d), 4.2(e), 4.2(f) e 4.2(g) mostram os resultados de agrupamento fornecidos, respectivamente, pelos algoritmos KMeans, KKMeans, AMKK-MeansGD, AMKKMeansGF, AMKKMeansLD e AMKKMeansLF no conjunto de dados sintéticos 1 (Fig. 4.2(a)). Como pode ser visto na Figura 4.2(d), o algoritmo AMKKMe-ansGD fornece a melhor categorização do conjunto de dados sintéticos 1, com as matrizes de covariância de classe diagonais e única para todos os grupos.



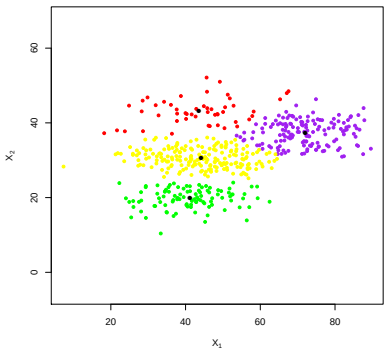
(a) Configuração 1



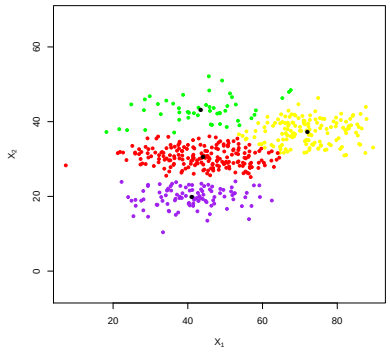
(b) KMeans



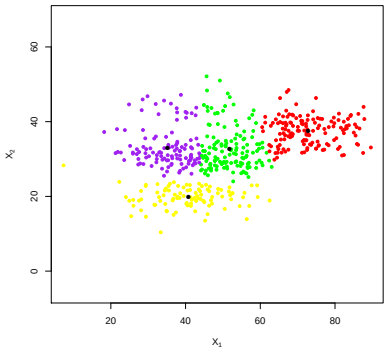
(c) KKMeans



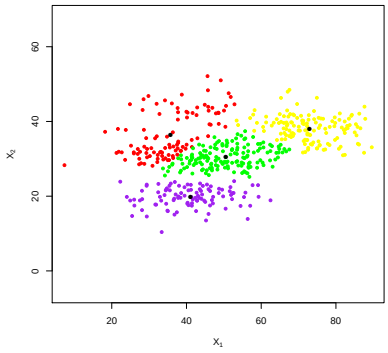
(d) AMKKMeansGD



(e) AMKKMeansGF



(f) AMKKMeansLD



(g) AMKKMeansLF

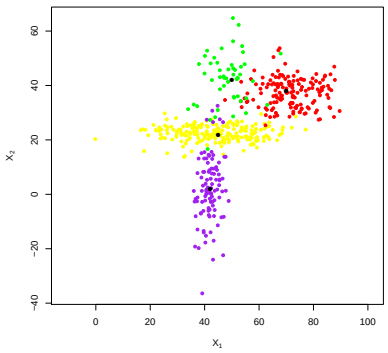
Figura 4.2: Dados simulados 1.

A tabela 4.4 apresenta o desempenho dos algoritmos de agrupamento para o conjunto de dados sintéticos 2. É possível notar que os algoritmos AMKKMeansLD e AMKK-MeansLF apresentam melhores resultados do que os algoritmos baseados em distâncias adaptativas globais (AMKKMeansGD e AMKKMeansGF) e os métodos convencionais (KMeans e KKMeans). Dentre os algoritmos com melhor desempenho o AMKKMeansLD fornece melhor explicação para o conjunto de dados 2 com o maior Índice Corrigido de Rand (0.770) e menor Erro de Taxa Total de Classificação (0.113).

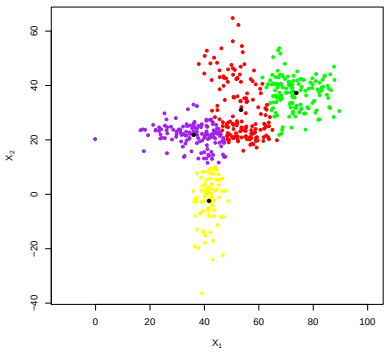
Tabela 4.4: Desempenho dos algoritmos no conjunto de dados sintéticos 2: média e desvio padrão (entre parênteses) dos índices CR e OERC.

	CR	OERC
KMeans	0.561 (0.077)	0.217 (0.059)
KKMeans	0.556 (0.079)	0.227 (0.063)
AMKKMeansGD	0.580 (0.098)	0.213 (0.066)
AMKKMeansGF	0.533 (0.078)	0.265 (0.064)
AMKKMeansLD	**0.770** (**0.039**)	**0.113** (**0.029**)
AMKKMeansLF	**0.745** (**0.050**)	**0.136** (**0.045**)

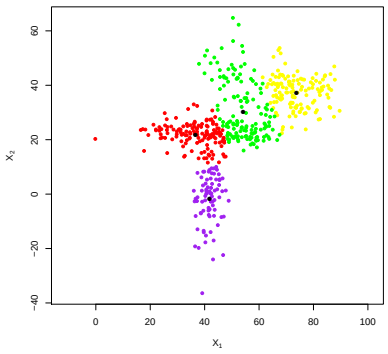
As Figuras 4.3(b), 4.3(c), 4.3(d), 4.3(e), 4.3(f) e 4.3(g) mostram os resultados de agrupamento fornecidos, respectivamente, pelos algoritmos KMeans, KKMeans, AMKK-MeansGD, AMKKMeansGF, AMKKMeansLD e AMKKMeansLF no conjunto de dados sintéticos 2 (Fig. 4.3(a)). Verifica-se na Figura 4.3(f) que o algoritmo AMKKMeansLD fornece uma melhor categorização dos dados, ou seja, matrizes de covariância de classe são diagonais e diferente para cada grupo.



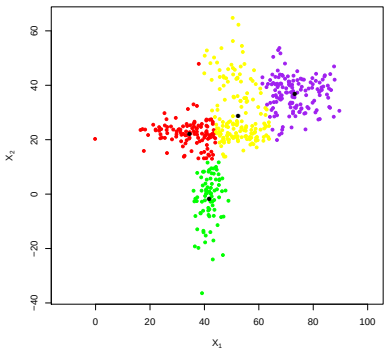
(a) Configuração 2



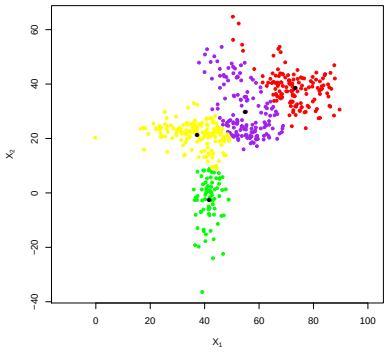
(b) KMeans



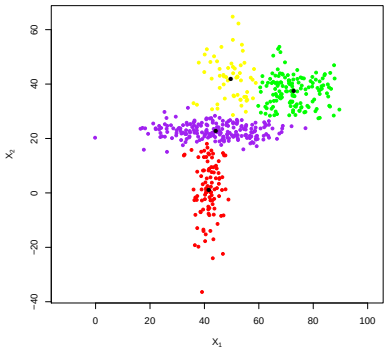
(c) KKMeans



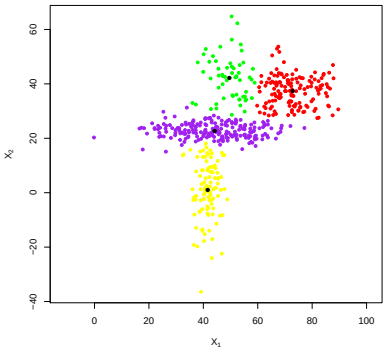
(d) AMKKMeansGD



(e) AMKKMeansGF



(f) AMKKMeansLD



(g) AMKKMeansLF

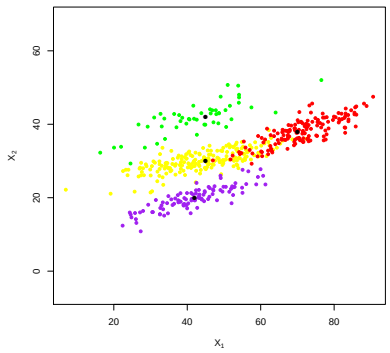
Figura 4.3: Dados simulados 2.

O desempenho dos algoritmos de agrupamento para o conjunto de dados sintéticos 3 é apresentado na Tabela 4.5. Verificam-se dois algoritmos com desempenhos semelhantes e bem superior aos demais algoritmos, são eles: AMKKMeansGF e AMKKMeansLF. Comparando ambos, o algoritmo AMKKMeansLF, no qual as matrizes de covariância de classe não são diagonais e diferente para cada grupo, apresenta o melhor desempenho, com maior Índice Corrigido de Rand (0.767) e menor Taxa Total de Erro de Classificação (0.080).

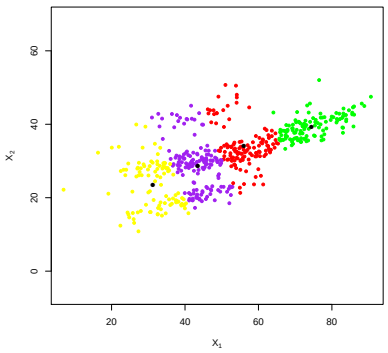
Tabela 4.5: Desempenho dos algoritmos no conjunto de dados sintéticos 3: média e desvio padrão (entre parênteses) dos índices CR e OERC.

	CR	OERC
KMeans	0.255 (0.029)	0.439 (0.031)
KKMeans	0.255 (0.031)	0.440 (0.033)
AMKKMeansGD	0.544 (0.110)	0.233 (0.097)
AMKKMeansGF	**0.760** (**0.037**)	**0.083** (**0.015**)
AMKKMeansLD	0.487 (0.153)	0.275 (0.130)
AMKKMeansLF	**0.767** (**0.032**)	**0.080** (**0.012**)

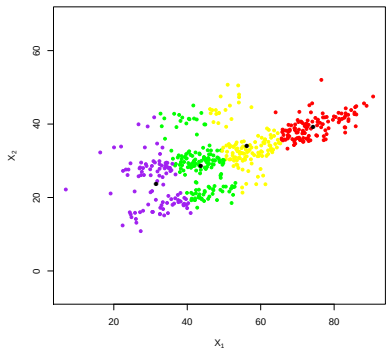
As Figuras 4.4(b), 4.4(c), 4.4(d), 4.4(e), 4.4(f) e 4.4(g) mostram os resultados dos agrupamentos fornecidos, respectivamente, pelos algoritmos KMeans, KKMeans, AMKKMeansGD, AMKKMeansGF, AMKKMeansLD e AMKKMeansLF no conjunto de dados sintéticos 3 (Fig. 4.4(a)). Como pode ser visto na Figura 4.4(e), o algoritmo AMKKMeansLF fornece uma melhor categorização dos dados. É possível verificar que os grupos estão dispostos com uma leve inclinação, o que indica direção, dado que existe uma relação de covariância entre as variáveis.



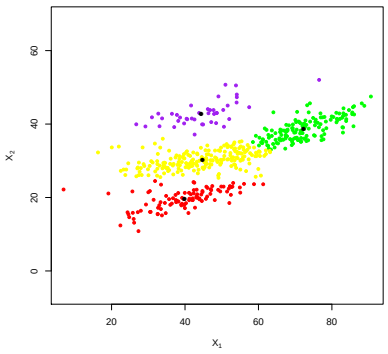
(a) Configuração 3



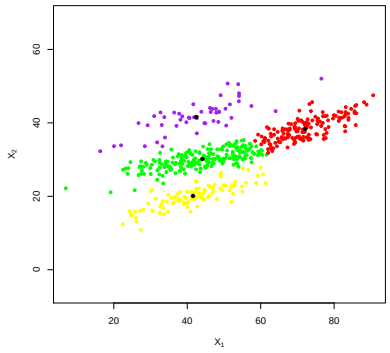
(b) KMeans



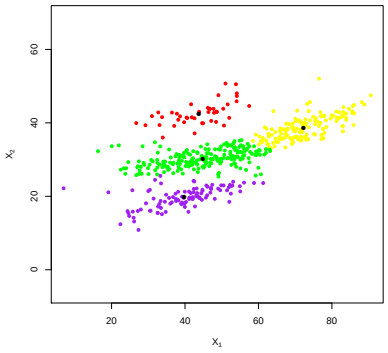
(c) KKMeans



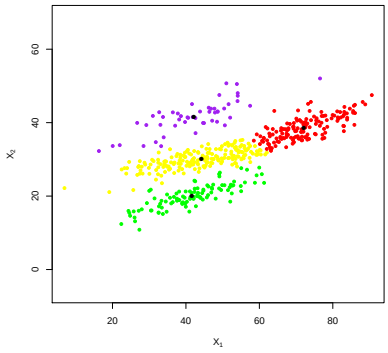
(d) AMKKMeansGD



(e) AMKKMeansGF



(f) AMKKMeansLD



(g) AMKKMeansLF

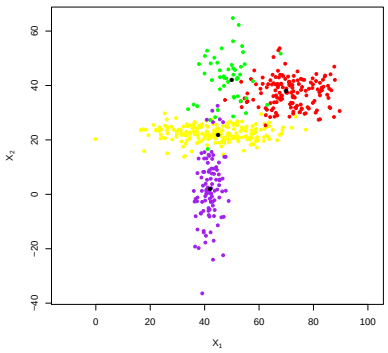
Figura 4.4: Dados simulados 3.

A tabela 4.6 apresenta o desempenho dos algoritmos KMeans, KKMeans, AMKK-MeansGD, AMKKMeansGF, AMKKMeansLD, bem como o desempenho do algoritmo AMKKMeansLF para o conjunto de dados sintéticos 4. É possível notar que o algoritmo AMKKMeansLF, matrizes de covariância diferentes para cada grupo e não diagonais, explica melhor o conjunto de dados, pois apresenta o maior Índice Corrigido de Rand (0.848) e menor Erro de Taxa Total de Classificação (0.057).

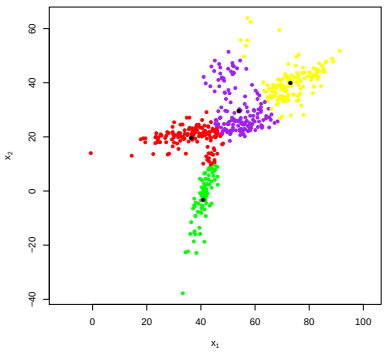
Tabela 4.6: Desempenho dos algoritmos no conjunto de dados sintéticos 4: média e desvio padrão (entre parênteses) dos índices CR e OERC.

	CR	OERC
KMeans	0.413 (0.026)	0.323 (0.023)
KKMeans	0.420 (0.026)	0.320 (0.023)
AMKKMeansGD	0.422 (0.027)	0.325 (0.024)
AMKKMeansGF	0.408 (0.071)	0.308 (0.056)
AMKKMeansLD	0.650 (0.108)	0.178 (0.073)
AMKKMeansLF	**0.848** (**0.025**)	**0.057** (**0.010**)

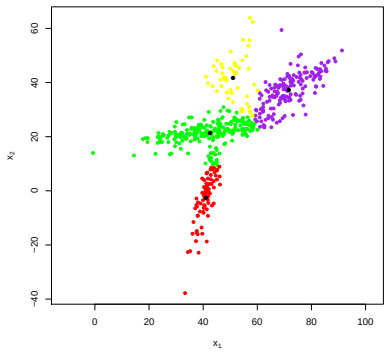
As Figuras 4.5(b), 4.5(c), 4.5(d), 4.5(e), 4.5(f) e 4.5(g) mostram os resultados de agrupamento fornecidos, respectivamente, pelos algoritmos KMeans, KKMeans, AMKK-MeansGD, AMKKMeansGF, AMKKMeansLD e AMKKMeansLF no conjunto de dados sintéticos 4 (Fig. 4.5(a)). Uma melhor caegorização dos dados pode ser visto na Figura 4.5(g) pelo algoritmo AMKKMeansLF que fornece matrizes de covariância de classe não diagonais e diferente para cada grupo. Assim como nas Figs. 4.4(b), 4.4(c), 4.4(d), 4.4(e), 4.4(f) e 4.4(g), os gupos estão levemente inclinados, característica da relação de covariância entre as variáveis.



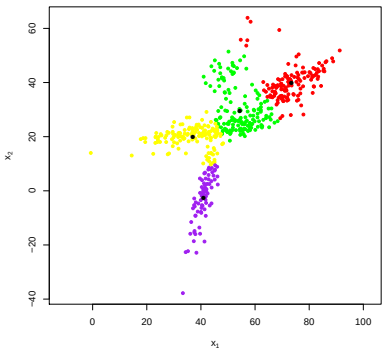
(a) Configuração 4



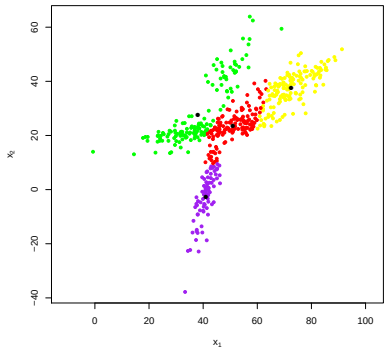
(b) KMeans



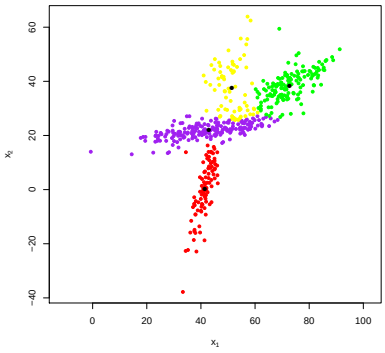
(c) KKMeans



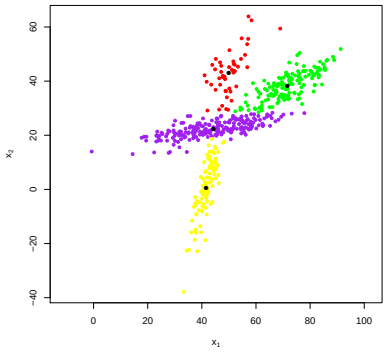
(d) AMKKMeansGD



(e) AMKKMeansGF



(f) AMKKMeansLD



(g) AMKKMeansLF

Figura 4.5: Dados simulados 4.

4.3.2 Avaliação com dados reais

Os algoritmos KMeans, KKMeans, AMKKMeansGD, AMKKMeansGF, AMKKMeansLD e AMKKMeansLF foram aplicados aos conjuntos de dados reais citados acima. Em cada caso os algoritmos de agrupamento foram executados 100 vezes (até a convergência para um valor estacionário de critério de adequação entre grupos e centróides) e selecionou-se o melhor resultado de acordo com os critérios de adequação.

Os critérios de adequação utilizados foram o Índice Corrigido de Rand (CR) e a Taxa Total de Erro de Classificação (OERC). As tabelas 4.7, 4.8, 4.9 e 4.10 apresenta o desempenho dos algoritmos para cada conjunto de dados reais. Em todos os conjunto considerados, os algoritmos propostos apresentam melhor desempenho.

A tabela 4.7 apresenta o desempenho dos algoritmos KMeans, KKMeans, AMKKMeansGD, AMKKMeansGF, AMKKMeansLD e AMKKMeansLF para o conjunto de dados Wirelles Indoor Localization. De acordo com os índices CR e OERC nota-se que o algoritmo que considera matriz de covariância completa e única para todos os grupos (AMKKMeansGF) apresentou o melhor desempenho.

Tabela 4.7: Desempenho dos algoritmos no conjunto de dados Wireless Indoor Localization

	CR	OERC
KMeans	0.8868	0.0455
KKMeans	0.9086	0.0360
AMKKMeans-GD	0.8947	0.0415
AMKKMeans-GF	**0.9250**	**0.0290**
AMKKMeans-LD	0.9223	0.0305
AMKKMeans-LF	0.8644	0.0565

A tabela 4.8 apresenta o desempenho dos algoritmos propostos e covencionais para o banco de dados Breast Cancer. Nesse conjunto o algoritmo com melhor desempenho é o que considera matriz de covariância diagonal e única para todos os grupos (AMKKMeansGD).

Tabela 4.8: Desempenho dos algoritmos no conjunto de dados Breast Cancer

	CR	OERC
KMeans	0.1732	0.5472
KKMeans	0.1864	0.5660
AMKKMeans-GD	**0.3413**	**0.3774**
AMKKMeans-GF	0.2488	0.4906
AMKKMeans-LD	0.2956	0.4434
AMKKMeans-LF	0.2743	0.4623

A tabela 4.9 apresenta o desempenho dos algoritmos propostos e covencionais para o banco de dados Banknote Authentication. O algoritmo com matrizes de covariância completa e única para todos os grupos (AMKKMeansGF) foi o melhor para representar os dados. Seu desempenho foi extremamente superior a todos os demais algoritmos.

Tabela 4.9: Desempenho dos algoritmos no conjunto de dados Banknote Authentication

	CR	OERC
KMeans	0.0485	0.3878
KKMeans	0.0307	0.4111
AMKKMeans-GD	0.0193	0.4293
AMKKMeans-GF	**0.8703**	**0.0335**
AMKKMeans-LD	0.0209	0.4264
AMKKMeans-LF	0.0046	0.4446

A tabela 4.10 apresenta o desempenho dos algoritmos propostos e covencionais para o banco de dados Iris. Nesse conjunto de dados, os métodos de agrupamento *hard* com matrizes de covariância completa (AMKKMeansGF e AMKKMeansLF) foram os melhores para separação das classes. Os dois algoritmos apresentam os mesmos valores de Índice Corrigido de Rand e Taxa Total de Erro de Classificação.

Tabela 4.10: Desempenho dos algoritmos no conjunto de dados Iris

	CR	OERC
KMeans	0.7163	0.1133
KKMeans	0.7302	0.1067
AMKKMeans-GD	0.8857	0.0400
AMKKMeans-GF	**0.9410**	**0.0200**
AMKKMeans-LD	0.8857	0.0400
AMKKMeans-LF	**0.9410**	**0.0200**

A tabela 4.11 apresenta o desempenho dos algoritmos propostos e covencionais para o banco de dados Abalone. Os algoritmos de agrupamento AMKKMeansLD e AMKKMeansLF (que consideram matrizes diferentes para cada grupo), foram os que se adequaram bem ao conjunto de dados. Pode-se notar que apresentam o mesmo valor do Índice Corrigido de Rand e uma diferença mínima na Taxa Total de Erro de Classificação. Contudo, temos que o algoritmo AMKKMeansLF (matrizes de covariância completa e diferente) apresenta desempenho superior aos demais.

Tabela 4.11: Desempenho dos algoritmos no conjunto de dados Abalone

	CR	OERC
KMeans	0.1331	0.4850
KKMeans	0.1294	0.4850
AMKKMeans-GD	0.1323	0.4953
AMKKMeans-GF	0.0929	0.5107
AMKKMeans-LD	**0.1425**	**0.4891**
AMKKMeans-LF	**0.1425**	**0.4750**

Neste capítulo foi avaliado o desempenho dos métodos de agrupamento *hard* baseados no *kernel* de Mahalanobis introduzido no Capítulo 3 através de experimentos com dados simulados e dados reais. Os resultados obtidos com a simulação de Monte Carlo evidenciou a superioridade dos algoritmos aqui propostos. Os resultados com dados reais também indicaram superioridade dos algoritmos propostos, o que mostra a utilidade desses algoritmos em diversas situações práticas da realidade.

CAPÍTULO 5

Conclusões

Nesse trabalho foram apresentados algoritmos de agrupamento *hard* baseados no *kernel* de Mahalanobis com distâncias quadráticas adaptativas considerando a abordagem da kernelização da métrica. A abordagem proposta foi superior em todas as análises em relação aos métodos convencionais, pois tem a vantagem da matriz simétrica positiva definida $\boldsymbol{M}$ em que as covariâncias entre os elementos são levadas em conta. Essas matrizes foram construídas com base em distâncias quadráticas adaptativas, que mudam a cada iteração dos algoritmos e tanto podem ser a mesma para todos os grupos (distâncias adaptativas globais), quanto serem diferentes de um grupo para outro (distâncias adaptativas completas). O uso de funções *kernel* se justifica pelo fato de que em situações de disposição não linear dos grupos, essas funções auxiliam a identifica-los e o uso da distância de Mahalanobis se justifica pelo fato de que leva em consideração a covariância entre as variáveis e é invariante em medida de escala, abrangendo uma classe muito maior de problemas.

Para cada algoritmo proposto, foram obtidas as expressões para o cálculo dos centróides dos grupos, expressões para as matrizes de covariância para cada método por meio da extração de sua derivada da função objetivo correspondente para cada matriz de covariância. A utilidade dos métodos propostos foi demonstrada através de experimentos númericos com conjuntos de dados simulados por meio de simulaçõs de Monte Carlo e conjunto de dados reais.

Os conjuntos foram avaliados considerando os índices CR e OERC. Com relação aos conjuntos de dados simulados os métodos de agrupamento *hard* baseados em distâncias quadráticas adaptativas globais e locais com matrizes de covariância diagonais/e ou completas apresentaram desempenho superior aos métodos convencionais nos quatro cenários simulados.

Destacamos assim os algoritmos que apresentaram bons desempenhos nas simulações para cada cenário. Para o primeiro conjunto de dados simulados o algoritmo AMKKMeansGD (matrizes de covariância de classe são diagonais e aproximadamente iguais) foi o mais adequado. No segundo conjunto de dados, o algoritmo AMKKMeansLD (matrizes de covariância são diagonais e diferentes para cada grupo) apresentou o melhor desempenho. Para o terceiro e quarto conjunto de dados simulados o algoritmo AMKKMeansLF (matrizes de covariância completas e diferente para cada grupo) foi o mais adequado em ambos os cenários.

Com relação aos conjuntos de dados reais, mais uma vez, os algoritmos baseados no *kernel* de Mahalanobis apresentaram desempenho superior em relação aos métodos convencionais. De maneira geral, as melhores classificações foram feitas pelos algoritmos de matrizes não diagonais (AMKKMeansGF e AMKKMeansLF). Para os dados Wireless Indoor Localization e Banknote Authentication, o algoritmo que considera matrizes de covariância completa e única para todos os grupos (AMKKMeansGF) foi o que teve melhor desempenho. No conjunto Breast Cancer o melhor algoritmo foi com matrizes de covariância diagonais e diferentes para cada grupo (AMKKMeansLD). Para o conjunto de dados Abalone o algoritmo com melhor desempenho foi com matrizes de covariância completa e diferente para cada grupo (AMKKMeansLF). E, no caso do conjunto de dados Iris, os melhores algoritmos foram AMKKMeansGF e AMKKMeansLF que apresentaram o mesmo desempenho.

Referências Bibliográficas

BREIMAN, L. et al. Classification and regression trees chapman & hall. *New York*, 1984.

CAMPS-VALLS, G. et al. Hyperspectral image classification with mahalanobis relevance vector machines. In: IEEE. *2007 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.], 2007. p. 3802–3805.

CAPUTO, B. et al. Appearance-based object recognition using svms: which kernel should i use? 2001.

COSTA, J. A. F. Segmentação do som por métodos de agrupamentos hierárquicos com conectividade restrita. In: *VII Congresso Brasileiro de Redes Neurais*. [S.l.: s.n.], 2005.

DONI, M. V. Análise de cluster: métodos hierárquicos e de particionamento. *São Paulo: Universidade Presbiteriana Mackenzie, São Paulo*, 2004.

FERREIRA, M. R.; CARVALHO, F. d. A. de. A kernel k-means clustering algorithm based on an adaptive mahalanobis kernel. In: IEEE. *2014 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2014. p. 1885–1892.

FERREIRA, M. R. P. Agrupamento baseado em kernel com ponderação automática das variáveis via distâncias adaptativas. Universidade Federal de Pernambuco, 2013.

FILIPPONE, M. et al. A survey of kernel and spectral methods for clustering. *Pattern recognition*, Elsevier, v. 41, n. 1, p. 176–190, 2008.

GIROLAMI, M. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, IEEE, v. 13, n. 3, p. 780–784, 2002.

HÖPPNER, F. et al. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. [S.l.]: John Wiley & Sons, 1999.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985.

MERCER, J. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, The Royal Society London, v. 209, n. 441-458, p. 415–446, 1909.

MILLIGAN, G. W.; COOPER, M. C. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate behavioral research*, Taylor & Francis, v. 21, n. 4, p. 441–458, 1986.

MULLER, K.-R. et al. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, IEEE, v. 12, n. 2, p. 181–201, 2001.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>.

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, MIT Press, v. 10, n. 5, p. 1299–1319, 1998.

WANG, D.; YEUNG, D. S.; TSANG, E. C. Weighted mahalanobis distance kernels for support vector machines. *IEEE Transactions on Neural Networks*, IEEE, v. 18, n. 5, p. 1453–1462, 2007.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, Springer, v. 2, n. 2, p. 165–193, 2015.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, Ieee, v. 16, n. 3, p. 645–678, 2005.

ZAKI, M. J.; JR, W. M.; MEIRA, W. *Data mining and analysis: fundamental concepts and algorithms*. [S.l.]: Cambridge University Press, 2014.