Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Estatística

Classificação otimizada baseada em U-estatísticas

Mayara Belló Soares

Porto Alegre, Abril de 2023.

CIP - Catalogação na Publicação

Soares, Mayara Belló
Classificação otimizada baseada em U-estatísticas / Mayara Belló Soares. -- 2023.
52 f.
Orientadora: Gabriela Bettella Cybis.

Coorientadora: Marcio Valk.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística, Programa de Pós-Graduação em Estatística, Porto Alegre, BR-RS, 2023.

1. Alta dimensão e baixo tamanho de amostra. 2. U-estatística. 3. classificador. 4. teste de agrupamento. 5. teste de classificação. I. Cybis, Gabriela Bettella, orient. II. Valk, Marcio, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Dissertação submetida por Mayara Belló Soares como requisito parcial para a obtenção do título de Mestre em Estatística pelo Programa de Pós-Graduação em Estatística da Universidade Federal do Rio Grande do Sul.

Orientador(a):

Prof. Dra. Gabriela Bettella Cybis

Co-orientador(a):

Prof. Dr. Marcio Valk

Comissão Examinadora:

Prof. Dr. Guilherme Pumi (PPGEst - UFRGS)

Prof. Dr. Danilo Marcondes Filho (PPGEst - UFRGS)

Prof. Dr. Anderson Luiz Ara Souza (PPGMNE - UFPR)

Data de Apresentação: 19 de Abril de 2023

"You cannot hope to build a better world without improving the individuals. To that end each of us must work for his own improvement, and at the same time share a general responsibility for all humanity, our particular duty being to aid those to whom we think we can be most useful."

(Marie Curie)

Agradecimentos

Agradeço primeiramente aos meus pais Luiz Felipe Soares e Arivâne Soares por me incentivarem a ir em busca dos meus sonhos, e também à minha irmã Alessandra Soares, por estar, junto com meus pais, sempre presentes na minha vida. Ao meu amigo e colega Leonardo Pinheiro, por todo o suporte ao longo desses anos. Agradeço à professora Dra. Gabriela Cybis, por aceitar a orientação e também por toda sua dedicação a este trabalho. Ao professor Dr. Márcio Valk pela coorientação. Agradeço aos membros da banca, prof. Dr. Guilherme Pumi, prof. Dr. Danilo Marcondes e prof. Dr. Anderson Ara pela disponibilidade. Agradeço aos professores do programa de Pós-Graduação em Estatística e também à UFRGS pela educação pública e de qualidade que representa.

Dedico este trabalho à memória de minha avó Iria Rosa Belló.

Resumo

A modelagem de dados visando agrupamento e classificação em ambientes de alta dimensão e baixo tamanho de amostra (HDLSS - *High-dimension low-sample size data*) é um desafio em diferentes áreas do conhecimento. Uma alternativa é a utilização de métodos não paramétricos, por permitir uma abordagem de inferência dependendo de poucos pressupostos sobre os dados. Em particular, uma série de métodos de inferência para problemas de agrupamento e classificação baseados em U-estatísticas, implementados no pacote uclust do *software* R, tem gerado resultados promissores no contexto HDLSS. Buscando tornar essa abordagem melhor adaptada a diferentes estruturas de dados, o foco desta dissertação é propor um método otimizado de classificação dentro desse contexto. A classificação é realizada em duas etapas: primeiro encontramos a distância ponderada que maximiza a separação entre dois grupos de referência, medida pela estatística B_n; e em seguida utilizamos essa distância para classificar novas observações, através de um enfoque comparativo. Estudos de Monte Carlo, no contexto de HDLSS, mostram que o método otimizado apresenta melhores taxas de classificações corretas quando a diferença entre grupos está concentrada em uma fração das componentes do vetor de dados. O uso desta distância otimizada também serve como base para a proposta de um novo teste U otimizado, que verifica se dois grupos de observações são de fato distintos, e também de um novo teste de hipóteses para a classificação. Estudos de simulação mostram que nos cenários simulados, onde o classificador está bem adaptado, ambos os testes apresentam mais poder que os métodos originais. É apresenta uma aplicação dos métodos a um conjunto de dados HDLSS de pacientes de linfoma, no qual o classificador otimizado apresenta resultados favoráveis.

Palavras-chave: Alta dimensão e baixo tamanho de amostra; U-estatística; classificador; teste de agrupamento; teste de classificação.

Abstract

Data modeling for clustering and classification in high dimension and low sample size (HDLSS) environments is a challenge in different areas of knowledge. An alternative is the use of non-parametric methods, because they allow for an inferencial approach depending on a few assumptions about the data. In particular, a series of inferencial methods for clustering and classification problems based on U-statistics, implemented in the Uclust R-package, has generated promising results in the HDLSS context. With the objective of making this approach better adapted to different data structures, this work proposes an optimized classification method within this context. The classification is carried out in two stages: first, we find the weighted distance that maximizes the separation between two reference groups, measured by the B_n statistic; and then, we use this distance to classify new observations, through a comparative approach. Monte Carlo studies, in the HDLSS context, show that the optimized method presents better rates of correct classifications when the difference between groups is concentrated in a few components of the data vector. This optimized distance also serves as the basis for the proposal of a new optimized U test, which verifies whether two groups of observations are in fact distinct, and also for a new hypothesis test for classification. Simulation studies show that in scenarios where the classifier is well adapted, both tests are more powerfull than the original methods. An application of these methods in a HDLSS dataset is presented.

Keywords: High dimension low sample size; U-statistics; classifier; clustering test; classification test.

Índice

Capítulo 1

Introdução

Nesta dissertação apresentamos uma proposta de métodos otimizados de agrupamento e classificação baseados em U-estatísticas. O trabalho se insere no contexto dos métodos de inferência para agrupamento com base em U-estatísticas do pacote $uclust$ do software R (Cybis et al., 2018; Valk and Cybis, 2021). Esta é uma abordagem não paramétrica, com poucos pressupostos sobre a distribuição dos dados, que pode ser aplicada a uma ampla gama de problemas e adaptada a diferentes tipos de dados. Adicionalmente, a abordagem apresenta bons resultados em um ambiente de alta dimensão e baixo tamanho amostral (High Dimension Low Sample Size - HDLSS), onde os métodos tradicionais de classificação são inadequados e os procedimentos multivariados (assim como os procedimentos de mineração de dados) encontram dificuldades devido à alta dimensionalidade (Kalina, 2014).

Para tornar os métodos ainda mais bem adaptados a diferentes estruturas de dados, sugerimos o uso de uma distância ponderada otimizada, que maximiza a separação entre dois grupos de referência e é utilizada para classificar um novo indivíduo em um deles. Além disso, propomos adaptações de testes para separação de dois grupos e de classificação, considerando o uso desta distância otimizada.

Revisão bibliográfica

A modelagem de dados visando agrupamento e classificação em ambientes de alta dimensão e baixo tamanho de amostra vem se tornando um desafio em diferentes áreas do conhecimento. Na genética, por exemplo, a tecnologia de microarrays tornou possível avaliar simultaneamente milhares de genes de uma amostra biológica. Cybis et al. (2018) discutem o aumento do uso de dados com alta complexidade no campo da genética e por consequência o aumento no desenvolvimento de métodos estatísticos adaptados para responder a esse tipo de questões envolvendo técnicas de agrupamento e classificação. Izbicki and dos Santos (2022) aponta que os métodos tradicionais já não são capazes de lidar de forma satisfatória nesse contexto. Além disso afirma que os avanços computacionais possibilitam uma maior capacidade de armazenamento de dados, e também permitem que novas metodologias sejam exploradas.

Um exemplo interessante nesse contexto é o trabalho proposto por Wu et al. (2023), um modelo para a predição de câncer baseado em aprendizado profundo. Nesse cenário é fundamental obter modelos mais precisos e utilizar dados baseados em microarrays pode facilitar a exploração molecular. Contudo uma das dificuldades encontradas é a complexidade dos padrões biológicos, relacionados ao

baixo tamanho amostral e alta dimensão dos dados de microarray, prejudicando assim o ajuste do modelo. Wu et al. (2023) comenta que a alta dimensionalidade exige que muitos parâmetros sejam estimados, mas o número pequeno de amostras é insuficiente para o processo de treinamento do modelo, produzindo um sobreajuste. No problema particular, os autores precisaram empregar uma estratégia complexa, composta de várias etapas para mitigar essa questão.

Frequentemente, um dos primeiros passos de análises exploratórias é o agrupamento dos dados, utilizado para encontrar uma estrutura de dados semelhantes. Segundo Clémençon (2014) a análise de agrupamento tem como objetivo associar os pontos mais similares uns dos outros em um mesmo grupo, então esse grupo será mais similar entre esses elementos do que quando comparado com os demais grupos, segmentando a base de dados em subgrupos. Estes são métodos de aprendizado não supervisionado e que se mostram uma ferramenta estatística muito útil na identificação de padrões sistemáticos. Na maioria das vezes pode ser aplicado em qualquer tipo de dados onde há evidências de similaridades ou dissimilaridades.

É importante notar que na abordagem por inferência estatística clássica para o agrupamento, supõe-se que a amostra representa a população e os resultados assintóticos baseiam-se no crescimento da amostra. Tais resultados não podem ser aplicados no cenário de HDLSS pela característica de pequeno tamanho amostral (Lacerda, 2022). Segundo Sen (2006), o fato de muitas das ferramentas da estatística inferencial serem limitadas nesse contexto, resulta em desafios para a validação estatística em estudos genômicos. Assim torna-se problemático o uso de modelos multivariados, onde o número de parâmetros pode superar o tamanho de amostra. Nesse contexto, Sen (2006) e Cybis et al. (2018) utilizaram as U-estatísticas como alternativa, pois tal abordagem é bem adaptada ao contexto HDLSS e requer poucos pressupostos específicos para esta modelagem, considerando a complexidade presente.

Com o objetivo de avaliar os agrupamentos para dados de microarrays McShane et al. (2002) apresentaram um método para avaliar a existência de agrupamentos significativos e também medidas úteis para entender a estrutura do agrupamento. Tal resultado é motivado pelo fato de que para esse tipo de dado, as técnicas de agrupamento frequentemente detectam agrupamentos mesmo em dados aleatórios, podendo provocar uma má interpretação dos resultados. Já Shimodaira (2004) propôs uma abordagem para avaliar a significância para um agrupamento hierárquico calculando o p-valor através de um método múltiplo de reamostragem por bootstrap, que mais tarde foi implementado no pacote pvclust do R (Suzuki and Shimodaira, 2006).

Liu et al. (2008) propõem um método para avaliar a significância do agrupamento de dados com alta dimensão e pequeno tamanho de amostra, conhecido como significância estatística de agrupamento (SigClust). Essa abordagem considera que se um grupo, resultante do agrupamento realizado, advém de uma única distribuição normal multivariada, então qualquer subdivisão desse grupo é não significativa. Considerando sob a hipótese nula a distribuição normal dos dados, é possível quantificar a significância do agrupamento através do p-valor. A simulação por Monte Carlo é usada para estimar p-valores baseados em percentil. Huang et al. (2015) aprimoram o método SigClust, propondo uma melhoria na estimação dos autovalores, levando a um melhor controle do erro do tipo I. Além disso, Kimes et al. (2017) propôs uma extensão do método SigClust para testar a significância estatística no agrupamento hierárquico.

Maitra et al. (2012) também apresentaram uma abordagem via reamostragem por bootstrap para avaliar a significância do agrupamento realizado, em grupos compactos, no ambiente multidimensional.

Desenvolveram uma metodologia para grupos não homogêneos e elipsoidal, além de um mapa de quantificação, baseado em p-valores e q-valores. Com os grupos definidos, o próximo passo é classificar um novo elemento em um dos grupos.

Conforme Zhang et al. (2023), o problema de classificação, no ambiente de HDLSS, é muito difícil separar de forma efetiva os diferentes grupos, pois há dificuldades em lidar com a vasta quantidade de dados e problemas de estimação de distâncias apropriadas. Foram propostos diversos métodos baseados em otimização para resolver os problemas de baixa precisão e eficiência na predição, sendo possível categorizar esses métodos em dois grupos, métodos pré-processados e métodos integrados. Os métodos pré-processados utilizam uma etapa de pré-processamento, realizando a seleção de variáveis ou redução de dimensão para então utilizar os modelos de classificação, como por exemplo o método de análise de componentes principais. Já os métodos integrados utilizam uma função de regularização ou decomposição em valores singulares (Singular Value Decomposition - SVD) dentro dos modelos de classificação para obter pesos de atributos e então realizar a classificação (Zhang et al., 2023). Hallajian et al. (2022) comentam que muitas variáveis afetam negativamente o processo de classificação por conter informações redundantes ou irrelevantes, competindo à seleção de variáveis encontrar o subconjunto ótimo, além de transformar os modelos de aprendizagem em modelos mais compreensíveis e economizar esforço computacional.

Liao and Akritas (2007) propõem uma abordagem diferente para classificação, baseada em teste. Eles trabalham com classificação binária univariada ou multivariada, além de apresentar uma extensão para a classificação multiclasse. O método considera que existem duas populações ou classes, com médias distintas, e então dois testes são realizados. O primeiro alocando a nova observação na população 1, sob hipótese nula de igualdade de médias, e o segundo teste alocando a nova observação na população 2, sob a hipótese nula de igualdade de médias. Então, os p-valores dos testes são comparados e assim a classificação é realizada naquele que apresentar menor p-valor.

U-estatísticas para classificação e agrupamento

Muitas técnicas de agrupamento utilizam a otimização de critérios empíricos que podem ser expressos como U-estatísticas, para minimizar a dispersão dentro do conjunto. Clémençon (2014) definiu uma abordagem estatística para estudar as propriedades teóricas desses métodos para realizar a separação em grupos. Em muitas dessas técnicas o cálculo da matriz de dissimilaridades nxn, contendo todas as medidas de distância pareadas entre os n pontos da amostra é necessário, podendo ser o ponto computacionalmente limitante (Modarres, 2022).

De acordo com Halmos (1946), as U-estatísticas são estimadores não viesados e simétricos que foram apresentados por Wassily Hoeffdings, sendo essa uma contribuição muito importante para a teoria de estimação não paramétrica. Embora outras abordagens já haviam sido propostas, como por exemplo, a de Halmos que justificou o uso de estimadores simétricos e não viesados, a generalização deste assunto só foi apresentada por Hoeffding. Ele formalizou a formulação básica dos parâmetros estatísticos e a construção adequada desses estimadores, apresentando suas propriedades de distribuição, com fácil acesso à teoria assintótica relacionada (Halmos, 1946).

Motivados por um cenário de análises de diversidade complexa, especialmente aquela que surge

em genética, genômica, ecologia, em que um grande desafio é dado pela alta dimensionalidade e por vezes baixo tamanho amostral, Pinheiro et al. (2009) estudaram uma classe de quasi U-estatísticas, denominada B_n. Desenvolveram a teoria assintótica dessas estatísticas sob a hipótese de homogeneidade com o uso de uma propriedade de Martingale. Nesse contexto, Valk and Pinheiro (2012) utilizam a estrutura da U-estatística generalizada para propor um método não paramétrico de inferência para agrupamento dentro do contexto de séries temporais. A teoria assintótica foi baseado na decomposição das medidas de dissimilaridade, mas a variância é estimada por um procedimento de reamostragem.

Dentro desse arcabouço teórico da estatística B_n, Cybis et al. (2018) propuseram dois testes de hipóteses para dados HDLSS. Um dos testes avalia a homogeneidade dos grupos e o outro teste verifica a significância estatística da classificação realizada, utilizando para isto a versatilidade da abordagem por U-estatísticas. O artigo está dentro do contexto da análise de dados genéticos, onde por apresentar dados frequentemente categóricos e estruturas complexas é um contexto desafiador para as técnicas de agrupamento e classificação.

A seguir vieram trabalhos como o de Valk and Cybis (2021), que utilizaram a inferência por essas U-estatísticas para realizar agrupamento hierárquico em um cenário de HDLSS. O principal resultado foi a obtenção de significância estatística para esse tipo de agrupamento, com maior poder estatístico do que os poucos métodos alternativos existentes para esse tipo de dado. Já Bello (2021), traz uma proposta de extensão dessas metodologias para realizar agrupamento em três grupos, além de verificar se a separação é estatisticamente significativa. Também recomenda sua utilização em casos de alta dimensão e pequeno tamanho de amostra. Os códigos do para esses métodos de agrupamento e classificação baseados em U-estatísticas estão disponíveis no pacote $uclust$ do software R. E ainda considerando este referencial teórico no contexto de HDLSS, Lacerda (2022) estudou o problema de classificação e propôs um novo método de classificação com inferência para dois ou mais grupos.

Seguindo uma abordagem alternativa, Rauf Ahmad and Pavlenko (2018) propuseram um U-classificador, com ajuste de viés para dados de alta dimensão construído a partir de uma combinação linear dois componentes relacionados a U-estatísticas. O método foi proposto para dois ou mais grupos mesmo que suas distribuições sejam não-normais.

Motivação e Objetivo

É muito comum em contextos de alta dimensionalidade que nem todas as componentes do vetor de dados contribuam igualmente para a diferenciação dos grupos. De fato, frequentemente, apenas uma pequena fração das componentes são determinantes para esta distinção. Embora os métodos baseados em U-estatística sejam apropriados para diferentes estruturas de dados e possam utilizar uma ampla gama de medidas de dissimilaridade, na prática geralmente se utiliza distâncias que tratam todas as componentes de forma homogênea. O objetivo desse trabalho é propor uma nova abordagem, encontrando a distância ótima, que melhor separa dois grupos, no contexto dos métodos baseados na estatística B_n. Essa abordagem busca se adaptar melhor a heterogeneidade entre as componentes do vetor de dados, resultando em classificadores mais eficientes e testes mais poderosos.

Novidades do trabalho

O trabalho propõe um método de classificação de um novo elemento em um de dois grupos, por meio da utilização de uma distância otimizada. O procedimento de classificação é realizado em duas etapas: 1 - encontrar a distância otimizada que maximiza a separação entre dois grupos de referência, medido pela estatística B_n; 2 - a partir dos pesos encontrados utilizar o classificador DB_n para classificar novas observações. Além disso, são propostos os testes U e o teste de classificação otimizados, ambos adaptados considerando a distância otimizada.

Suporte computacional

O *software* R versão 4.1.3, por meio da interface do RStudio versão 2022.02.0, foi utilizado para as simulações e aplicação dos métodos abordados. O pacote $uclust$ também foi fundamental para os estudos realizados neste trabalho e está disponível em: https://CRAN.R-project.org/package=uclust.

Organização do trabalho

O restante do trabalho está organizado da seguinte forma: O Capítulo 2 apresenta a revisão da metodologia, abordando os principais conceitos dos métodos que serão utilizados ao longo do trabalho; O Capítulo 3 traz o desenvolvimento do cálculo da distância otimizada e as adaptações nos testes propostos. No Capítulo 4 é detalhado o processo das simulações de Monte Carlo, apresentando os resultados da performance do método proposto comparada ao método utilizando a distância euclidiana. Também é comparada a performance dos testes otimizados com os testes na forma padrão. O Capítulo 5 descreve um problema real, um conjunto de dados de pacientes com linfoma, em que o método proposto será aplicado e comparado a outros métodos. Por fim, o Capítulo 6 faz um retoma a proposta do trabalho, discutindo os resultados e dificuldades encontrados.

Capítulo 2

Revisão de metodologia

Neste Capítulo, apresentamos os principais conceitos dos métodos utilizados para sustentar este trabalho.

2.1 U-estatística e definição da estatística B_n

Nesta seção será apresentada uma breve contextualização da teoria de U-estatísticas. Conforme mencionado anteriormente, elas desempenham um papel importante na teoria da estimação não-paramétrica, em particular, para a construção de Estimadores Não Viesados de Variância Uniformemente Mínima (ENVVUM).

Para a definição formal da U-estatística, considere $\mathbf{X} = (\mathbf{X}_{11}, ..., \mathbf{X}_{1n_1}, \mathbf{X}_{2(n_1+1)}, ..., \mathbf{X}_{2n})$, uma amostra aleatória de n vetores L-dimensionais, dividida em dois grupos, G_1 e G_2, de tamanhos n_1 e n_2 respectivamente, sendo $n = n_1 + n_2$. Assim, $\mathbf{X}_{gi}$ representa a i-ésima observação, que pertence ao grupo g. Assume-se que os elementos de cada grupo constituem amostras aleatórias e identicamente distribuídas (iid), obtidas a partir das distribuições L-dimensionais F_1 e F_2. Seja $\phi(\cdot, \cdot)$ um kernel não negativo, frequentemente representado por uma medida de distância ou dissimilaridade e defina o parâmetro funcional $\theta(F_1, F_2)$ como a esperança (Pinheiro et al., 2009), assim

$$\theta(F_1, F_2) = \int \int \phi(\mathbf{X}_{1i}, \mathbf{X}_{2j}) dF_1(\mathbf{X}_{1i}) dF_2(\mathbf{X}_{2j}), \text{ para } \mathbf{X}_{1i}, \mathbf{X}_{2j} \in \mathbb{R}^L. \tag{2.1}$$

Assumindo $\phi(\cdot, \cdot)$ uma função convexa linear das componentes marginais, temos que

$$\theta(F_1, F_2) \geq \frac{1}{2}(\theta(F_1, F_1) + \theta(F_2, F_2)), \tag{2.2}$$

para as distribuições F_1 e F_2. Se tomamos $\mathbb{E}_{F_g}(\mathbf{X}_{gi}) = \boldsymbol{\mu}_g$, para os grupos $g \in \{1, 2\}$, então a desigualdade vale quando $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

Note que o parâmetro $\theta(F_1, F_2)$ pode ser interpretado como a distância esperada entre observações de F_1 e F_2. Nesse contexto a U-estatística generalizada correspondente a $\theta(F_g, F_g)$ é dada por

$$U_{n_g}^{(g)} = \begin{pmatrix} n_g \\ 2 \end{pmatrix}^{-1} \cdot \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}). \tag{2.3}$$

Aqui $U_{n_g}^{(g)}$ é o estimador não viesado para o parâmetro funcional dentro do grupo, $\theta(F_g, F_g)$, que pode ser interpretado como a distância esperada dentro do grupo. Já a U-estatística para o parâmetro $\theta(F_1, F_2)$, representando o estimador não viesado para a distância esperada entre grupos pode ser escrita como

$$U_{n_1,n_2}^{(1,2)} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(\mathbf{X}_{1i}, \mathbf{X}_{2j}). \tag{2.4}$$

Assim, a U-estatística para a amostra combinada que corresponde à Expressão (2.3) quando todas as observações são tratadas como pertencentes ao mesmo grupo, pode ser decomposta como

$$U_n = \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)} + \frac{n_1 n_2}{n(n-1)} (2U_{n_1,n_2}^{(1,2)} - U_{n_1}^{(1)} - U_{n_2}^{(2)}) = W_n + B_n, \tag{2.5}$$

em que a estatística B_n é escrita como

$$B_n = \frac{n_1 n_2}{n(n-1)} (2U_{n_1,n_2}^{(1,2)} - U_{n_1}^{(1)} - U_{n_2}^{(2)}). \tag{2.6}$$

Deste modo, a estatística B_n pode ser vista como a comparação entre a distância média entre grupos e as distâncias médias dentro de cada grupo. A estatística B_n é uma quasi U-estatística, pertencendo à classe das U-estatísticas degeneradas (Pinheiro et al., 2009). Além disso, temos que B_n é assintoticamente normal, tanto quando $n \to \infty$ quanto quando $L \to \infty$ (Pinheiro et al., 2009).

Este trabalho se utiliza da estatística B_n como função objetiva para encontrar uma distância otimizada que melhor separa dois grupos, e em seguida utiliza-se dessa distância para classificar novas observações, através de uma abordagem comparativa, o classificador DB_n.

2.2 *Teste U para separação de dois grupos*

O teste U para separação de grupos, baseado na U-estatística B_n, é empregado para verificar significância estatística da separação de dois grupos de observações previamente definidos G_1 e G_2. O principal pressuposto do teste é a homogeneidade dentro de cada grupo. A hipótese nula afirma que os dois grupos advém da mesma distribuição, ou seja $F_1 = F_2$. Nesse caso temos que $\mathbb{E}(B_n) = 0$. Já a hipótese alternativa afirma que os grupos são de fato separados com $F_1 \neq F_2$ (e $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$), e assim, a desigualdade (2.2) garante que $\mathbb{E}(B_n) > 0$ (Sen, 2006).

A normalidade assintótica de B_n é utilizada para construir a distribuição da estatística do teste sob H_0, na qual a variância de B_n é estimada por um procedimento de reamostragem sob H_0. Nesse procedimento as observações dos diferentes grupos são reunidas e posteriormente rearranjadas em dois grupos de tamanhos n_1 e n_2 para o cálculo da estatística B_n^r. O processo é repetido R vezes, e os valores de B_n^r, para $r : \{1, \cdots, R\}$, são utilizados para estimação da variância Pinheiro et al. (2009).

2.3 Teste de homogeneidade

Cybis et al. (2018) consideram o problema de identificar se uma amostra é homogênea. Nesse contexto de U-estatística Valk and Pinheiro (2012) definem uma amostra como homogênea se ela não pode ser separada em dois subgrupos significativamente distintos de acordo com o teste U. A principal diferença do teste anterior, é que aqui não há uma designação pré-definida dos grupos de modo que todas as observações estão reunidas. Para verificar se a amostra é homogênea, deve se considerar todas as possíveis combinações dos dados em dois subgrupos e realizar o teste U para ver se alguma partição dos dados é significativa.

Em seu trabalho, Cybis et al. (2018) propôs avaliar primeiramente a configuração que melhor separa os dois grupos e somente depois aplicar o teste U apenas na configuração de melhor separação, reduzindo assim drasticamente o grande custo computacional. Caso a hipótese nula não seja rejeitada para um subgrupo, todos os demais arranjos seriam considerados necessariamente homogêneos, já com a rejeição da hipótese nula seria possível afirmar que os dois grupos são de fato dissimilares.

A configuração que melhor separa os dois grupos é definida como aquela que tem maior valor de B_n padronizado. Então, sob H_0 vamos considerar uma aproximação para a distribuição do máximo de B_n padronizado que assume que as estatísticas B_n são independentes para os diferentes arranjos de grupo. Temos que

$$F_{\max}(x) = \mathbb{P}\left(\max\left(\frac{B_n}{\sqrt{Var(B_n)}}\right) < x\right) = \Phi(x)^{\gamma}, \tag{2.7}$$

para $\Phi(\cdot)^{\gamma}$ função da distribuição acumulada da normal padrão elevado na potência γ, onde $\gamma = 2^{n-1} - n - 1$ é o número possíveis configurações separando os dados em dois grupos. Rejeita-se a hipótese nula de homogeneidade se $F_{\max}(x) > (1 - \alpha)$.

2.4 Classificador DB_n e teste de hipóteses para classificação

Nesta seção será apresentado o método de classificação que será foco desse trabalho. Além disso, apresentaremos o teste de hipóteses para verificar a significância da classificação proposto por Lacerda (2022), uma reformulação do método publicado por Cybis et al. (2018). Tanto o classificador DB_n quanto o teste de hipóteses para a classificação estão definidos no contexto de alta dimensionalidade e baixo tamanho amostral (HDLSS) e estão baseados na estatística B_n, utilizando da teoria das U-estatísticas para apresentar suas propriedades.

Levando em consideração a rejeição da hipótese nula do teste U de separação, ou seja, G_1 e G_2 são grupos dissimilares, o próximo passo é a classificação de uma nova observação. Valk and Pinheiro (2012) já haviam proposto um método empírico para a classificação baseado na comparação de duas estatísticas B_n. Suponha que $\mathbf{X}$ seja composta de n_1 observações sabidamente pertencentes ao grupo 1 e n_2 observações pertencentes ao grupo 2, e seja $\mathbf{X}^*$ uma nova observação que gostaríamos de classificar em um dos grupos. Nesse contexto, defina B_{n_1} como a estatística B_n calculada quando a nova amostra $\mathbf{X}^*$ é colocada no grupo 1 e B_{n_2} como a estatística calculada quando $\mathbf{X}^*$ é colocada no grupo 2. Note que quanto maiores as distâncias entre grupos em relação às dentro de grupos, ou

seja, mais compactos e separados os grupos, maior será o valor de B_n. Assim, se $\mathbf{X}^*$ não estiver bem classificada no grupo 2 é esperado que a estatística B_{n_2} seja menor que a estatística B_n calculada sem $\mathbf{X}^*$. Portanto, caso, B_{n_1} for maior que B_{n_2}, então $\mathbf{X}^*$ produz um melhor agrupamento quando está classificada no grupo 1. Desse modo, considere a estatística

$$DB_n = B_{n_1} - B_{n_2}, \tag{2.8}$$

que será a base do classificador. Se o valor da estatística DB_n for maior que zero, $\mathbf{X}^*$ é classificado como pertencente ao grupo 1, caso contrário $\mathbf{X}^*$ deve ser classificado como grupo 2.

Levando em conta essa abordagem, foi proposto por Cybis et al. (2018) um teste para avaliar a significância estatística da classificação, baseado na estatística DB_n. Considerando $\mu_{B_{n_1}}$ o valor esperado para a estatística B_{n_1} e $\mu_{B_{n_2}}$ para a estatística B_{n_2}, podemos dizer que $E(DB_n) = \mu_{B_{n_1}} - \mu_{B_{n_2}} \equiv \mu_{DB_n}$, com isso as hipóteses do teste são definidas como

$$H_0 : \mu_{DB_n} \leq 0 \text{ versus } H_1 : \mu_{DB_n} > 0. \tag{2.9}$$

A hipótese nula afirma que B_{n_2}, em que $\mathbf{X}^* \in G_2$, produz um melhor agrupamento que B_{n_1}, já a hipótese alternativa é que $\mathbf{X}^*$ está corretamente classificado em G_1. Para a estimação do p-valor associado ao teste é utilizado um método de reamostragem, em que n_1 elementos do grupo G_1 e n_2 do grupo $G_2 \cup \mathbf{X}^*$ são amostrados repetidas vezes, em um processo de amostragem com reposição, com $\mathbf{X}^*$ sempre sorteado a partir de $G_2 \cup \mathbf{X}^*$ (hipótese nula). Para cada replicação, o valor de DB_n^r é calculado. A rejeição da hipótese nula se dá quando a estatística DB_n é maior que o percentil $(1-\alpha)$ da distribuição reamostrada. Note que a distribuição empírica é utilizada para verificar significância considerando que a distribuição da estatística DB_n é desconhecida.

Em seu trabalho Lacerda (2022) retomou o teste de classificação aprofundando o conhecimento sobre a estatística DB_n, demonstrando a teoria assintótica. As estatísticas B_{n_1} e B_{n_2} são assintoticamente normais em n e/ou L, resultado que pode ser verificado em (Pinheiro et al., 2009), Lacerda (2022) mostra que dessa forma a estatística DB_n por ser uma diferença entre normais será assintoticamente normal.

Para a estimação da média e variância da estatística DB_n sob H_0, Lacerda (2022) propôs um método utilizando um processo de reamostragem, remodelado para evitar problemas associados a distâncias nulas nas replicações. Para aplicá-lo precisamos que os grupos tenham tamanhos $n_1 - 1$ e $n_2 - 1$, de modo que uma observação de cada grupo deve ser desconsiderada no cálculo da DB_n. Nele $n_1 - 1$ elementos do grupo G_1 são amostrados sem reposição, já os elementos do grupo G_2 são rearranjados de modo que $n_2 - 1$ elementos são designados ao grupo G_2 da replicação e o elemento restante é designado como $\mathbf{X}^*$ na replicação (hipótese nula). Para cada replicação, o valor de DB_n^r é calculado. O conjunto dos R valores de DB_n^r é utilizado para estimar média e variância de DB_n.

Em Lacerda (2022), o teste de hipóteses para a classificação é realizado em duas etapas, em que no primeiro momento identifica-se qual o grupo mais verossímil para o elemento $\mathbf{X}^*$, utilizando para isso a estatística DB_n. E posteriormente, define-se como G_1 o grupo no qual $\mathbf{X}^*$ foi classificado.

Então, o teste de hipóteses para a classificação é enunciado da seguinte forma

$$\begin{cases} H_0 : \mathbf{X}^* \in G_2; \\ H_1 : \mathbf{X}^* \in G_1. \end{cases}$$

Sob H_0, temos que $E(DB_n) \leq 0$, deste modo ao rejeitar H_0 temos evidências suficientes para garantir que o novo elemento está corretamente classificado em G_1 a um nível de significância α.

Para elucidar melhor o contexto do teste de classificação em relação as hipóteses, ainda que represente uma simplificação para o contexto unidimensional, abaixo apresentamos o comportamento das hipóteses em um gráfico apresentado na figura 2.1. Observe os grupos G_1 e G_2, e considere H_0: $\mathbf{X}^* \in G_2$ e H_1: $\mathbf{X}^* \in G_1$. Logo, espera-se que se $\mathbf{X}^*$ está bem classificado no grupo G_2, ele fique sob a região de H_0 representada pela curva em azul, já se $\mathbf{X}^*$ estiver bem classificado no grupo G_1, ele fique sob a região de H_1 representada pela curva em preto.

Pontos na região de sobreposição das curvas podem ser oriundos de qualquer uma das distribuições, mas o classificador DB_n designará $\mathbf{X}^*$ para um dos dois grupos: grupo G_1 se $\mathbf{X}^* > 5.25$ e grupo G_2 se $\mathbf{X}^* < 5.25$. Para pontos próximos dessa linha, a confiança na classificação pode ser baixa. Só afirmamos que uma classificação é significativa no grupo G_1 (rejeitando H_0) se $\mathbf{X}^*$ está além de determinado quantil da distribuição de G_2 (região B). Assim, consideramos quatro possibilidades:

- $\mathbf{X}^* \in$ Região A - Classificado significativamente no grupo G_2;
- $\mathbf{X}^* \in$ Região A* - Classificado no grupo G_2, mas não de modo significativo;
- $\mathbf{X}^* \in$ Região B* - Classificado no grupo G_1, mas não de modo significativo;
- $\mathbf{X}^* \in$ Região B - Classificado significativamente no grupo G_1.

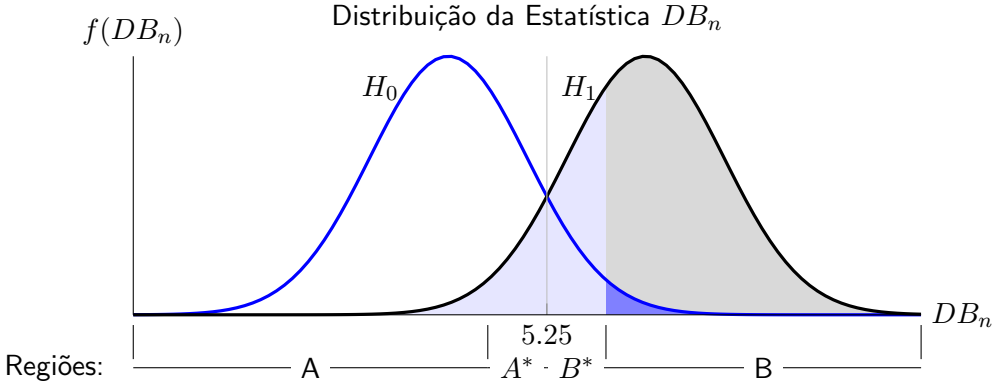

Figura 2.1: Representa a distribuição da estatística DB_n sob as hipóteses

Ainda é possível mensurar a probabilidade de cometer um erro na tomada de decisão, esse erro pode ser compreendido de duas formas: erro do tipo I e erro do tipo II. O erro do tipo I é quando rejeitamos a hipótese nula H_0 e ela é verdadeira, essa probabilidade também é chamada de nível de significância e está representada sob a curva em azul, destacado em azul escuro no gráfico,

$$\alpha = \mathbb{P}(\text{Erro do tipo I}) = \mathbb{P}(\text{rejeitar } H_0 | H_0 \text{ é verdadeira}). \tag{2.10}$$

Já o erro do tipo II, é cometido quando não rejeitamos a hipótese nula H_0 e ela é falsa, representado no gráfico pela área sob H_1, representada pela área sob a curva preta, destacado em azul claro no gráfico,

$$\beta = \mathbb{P}(\text{Erro do tipo II}) = \mathbb{P}(\text{não rejeitar } H_0 | H_0 \text{ é falsa}). \tag{2.11}$$

E, além disso, é possível avaliar o poder do teste, que representa a probabilidade de rejeitar a hipótese nula H_0, quando a hipótese alternativa H_1 é verdadeira, sendo a área sob H_1, curva preta, destacado em cinza,

$$1 - \beta = \mathbb{P}(\text{rejeitar } H_0 | H_1 \text{ é verdadeira}). \tag{2.12}$$

Note que, apesar da estratégia de definição de H_0 de Lacerda (2022), assim como definimos um teste com H_0: $\mathbf{X}^* \in G_2$, teoricamente poderíamos ter definido o teste simétrico que designaria classificações significativas no grupo G_1, em que H_0: $\mathbf{X}^* \in G_1$. Inclusive, o procedimento de classificação com inferência possui uma definição alternativa com base em p-valores, nos moldes do apresentado em Liao and Akritas (2007). Nele, dois testes de hipóteses são realizados, um com cada grupo representando a hipótese alternativa. A classificação ocorre naquele grupo cujo teste apresentou menor p-valor, se esse for menor que o nível de significância. Note entretanto que assim estamos trabalhando com dois testes de hipótese simétricos, e portanto podemos corrigir o nível de significância para múltiplos testes por Bonferroni, utilizando $\alpha/2$.

A partir da revisão de literatura exposta aqui, no próximo capítulo será apresentado o desenvolvimento da proposta desta dissertação.

Capítulo 3

Metodologia

Neste Capítulo, começaremos a explorar a contribuição deste trabalho. A motivação é a proposta de um classificador otimizado organizado em duas etapas: na primeira buscamos a distância euclidiana ponderada que melhor separa dois grupos de dados, em seguida classificamos novas observações em um dos dois grupos com base nessa distância ótima. Iniciamos o Capítulo apresentando o desenvolvimento do problema de otimização para maximizar a distância euclidiana ponderada entre dois grupos. Na sequência apresentaremos versões modificadas do teste U, do classificador e do teste de classificação, levando em consideração essa distância otimizada.

3.1 Distância otimizada

Nessa seção vamos apresentar o processo de otimização do método proposto que se utiliza da estatística B_n como função objetiva para maximizar a distância entre grupos, associando um peso a cada componente do vetor de dados.

Para o cálculo da estatística B_n e o emprego dos métodos descritos no Capítulo 2 é necessária a escolha de um kernel $\phi(\cdot,\cdot)$. Embora comumente a estatística B_n emprega a distância euclidiana, neste trabalho propomos a utilização de uma distância otimizada, baseada na distância euclidiana ponderada. O intuito é encontrar a configuração de pesos que maximiza a separação entre grupos, medida pela estatística B_n, em que as componentes do vetor de dados mais relevantes para a separação dos grupos recebam um maior peso. A partir dos pesos encontrados nessa otimização estamos interessados em realizar a classificação de uma nova observação em um de dois grupos, empregando o classificador DB_n, após isso, um teste hipóteses para a classificação, verificando se a classificação é estatisticamente significativo.

Para fins dessa sessão, sejam $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{n_1}$ observações L dimensionais pertencentes ao grupo G_1, e $\boldsymbol{X}_{n_1+1}, \cdots, \boldsymbol{X}_{n_1+n_2}$ observações L dimensionais pertencentes ao grupo G_2. Considere como kernel a distância euclidiana ponderada ao quadrado entre as observações $\boldsymbol{X}_i$ e $\boldsymbol{X}_j$, para $i, j \in \{1, \cdots, n\}$, tal que

$$\phi(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{\ell=1}^{L} a_\ell (X_{i\ell} - X_{j\ell})^2 \text{ para } a_\ell > 0, \tag{3.1}$$

em que $X_{i\ell}$ é a ℓ-ésima componente do vetor $\boldsymbol{X}_i$ e a_ℓ representam os pesos associados à ℓ-ésima componente do vetor $\boldsymbol{X}_i$. Adicionalmente, para regularizar o problema de otimização, adicionaremos a restrição de que a norma euclidiana dos pesos seja igual a 1. A norma euclidiana do vetor de pesos $\boldsymbol{a} = (a_1, \cdots, a_L)$ é dada por

$$|\boldsymbol{a}| = \sqrt{\boldsymbol{a} \cdot \boldsymbol{a}} = \sqrt{a_\ell^2 + \cdots + a_L^2}, \tag{3.2}$$

em que $\cdot$ representa o produto interno. Note, entretanto, que $|\boldsymbol{a}| = 1 \Leftrightarrow |\boldsymbol{a}|^2 = 1$. Assim, adicionar uma restrição à norma euclideana ao quadrado nos leva a resultado equivalente, portanto assumimos

$$\sum_{\ell=1}^{L} a_\ell^2 = 1. \tag{3.3}$$

Queremos encontrar a configuração dos a_ℓ que maximiza a estatística B_n, definida em (2.6), sujeita à restrição (3.3). Portanto, o lagrangiano deste problema de otimização é dado por

$$\mathcal{L}(\boldsymbol{a}, \lambda) = B_n + \lambda \left(\sum_{\ell=1}^{L} a_\ell^2 - 1 \right), \tag{3.4}$$

assim, temos

$$\mathcal{L}(\boldsymbol{a}, \lambda) = \frac{n_1 n_2}{n(n-1)} \cdot (2U_{n_1,n_2}^{(1,2)} - U_{n_1}^{(1)} - U_{n_2}^{(2)}) + \lambda \left(\sum_{\ell=1}^{L} a_\ell^2 - 1 \right),$$

substituindo $U_{n_1,n_2}^{(1,2)}$, $U_{n_1}^{(1)}$, $U_{n_2}^{(2)}$ definidas em (2.4) e (2.3), temos

$$\begin{aligned} \mathcal{L}(\boldsymbol{a}, \lambda) \quad &= \frac{n_1 n_2}{n(n-1)} \cdot \left[2 \left(\frac{1}{n_1 n_2} \cdot \sum_{i \in G_1} \sum_{j \in G_2} \phi(\boldsymbol{X}_i, \boldsymbol{X}_j) \right) - \binom{n_1}{2}^{-1} \cdot \sum_{i \in G_1} \phi(\boldsymbol{X}_i, \boldsymbol{X}_j) \right. \\ &\quad \left. - \binom{n_2}{2}^{-1} \cdot \sum_{i \in G_2} \phi(\boldsymbol{X}_i, \boldsymbol{X}_j) \right] + \lambda \left(\sum_{\ell=1}^{L} a_\ell^2 - 1 \right). \end{aligned}$$

Notando que $\phi(\boldsymbol{X}_i, \boldsymbol{X}_j)$ é a distância ponderada (3.1), e reorganizando os termos na equação

$$\begin{aligned} \mathcal{L}(\boldsymbol{a}, \lambda) \quad &= \sum_{\ell=1}^{L} a_\ell \left[\frac{n_1 n_2}{n(n-1)} \cdot \left(\frac{2}{n_1 n_2} \cdot \sum_{i \in G_1} \sum_{j \in G_2} (X_{i\ell} - X_{j\ell})^2 \right. \right. \\ &\quad - \binom{n_1}{2}^{-1} \cdot \sum_{i \in G_1} (X_{i\ell} - X_{j\ell})^2 \\ &\quad \left. \left. - \binom{n_2}{2}^{-1} \cdot \sum_{i \in G_2} (X_{i\ell} - X_{j\ell})^2 \right) \right] + \lambda \left(\sum_{\ell=1}^{L} a_\ell^2 - 1 \right). \end{aligned} \tag{3.5}$$

De maneira a facilitar a visualização dos cálculos, podemos definir $U_{n\ell}$ considerando apenas os termos dentro do colchetes e que não dependem de a_ℓ, assim

$$\begin{aligned} U_{n\ell} = \frac{n_1 n_2}{n(n-1)} \cdot &\left(\frac{2}{n_1 n_2} \cdot \sum_{i \in G_1} \sum_{j \in G_2} (X_{i\ell} - X_{j\ell})^2 \right. \\ &- \binom{n_1}{2}^{-1} \cdot \sum_{i \in G_1} (X_{i\ell} - X_{j\ell})^2 \\ &\left. - \binom{n_2}{2}^{-1} \cdot \sum_{i \in G_2} (X_{i\ell} - X_{j\ell})^2 \right). \end{aligned} \tag{3.6}$$

Note que $U_{n\ell}$ corresponde à estatística B_n calculada apenas para a coordenada ℓ.

Então é possível reescrever o lagrangiano como

$$\mathcal{L}(\boldsymbol{a}, \lambda) = \sum_{\ell=1}^{L} a_\ell \cdot U_{n\ell} + \lambda \left(\sum_{\ell=1}^{L} a_\ell^2 - 1 \right). \tag{3.7}$$

Para resolver o problema de otimização, igualamos o gradiente do lagrangiano $\nabla\mathcal{L}$ ao vetor nulo

$$\nabla\mathcal{L}(\boldsymbol{a}, \lambda) = \mathbf{0}, \tag{3.8}$$

de modo que para cada componente do vetor temos

$$\frac{\partial\mathcal{L}}{\partial a_\ell} = U_{n\ell} + 2a_\ell\lambda = 0. \tag{3.9}$$

Isolando a_ℓ obtemos o ponto crítico deste problema,

$$a_\ell = -\frac{1}{2\lambda} \cdot U_{n\ell}, \tag{3.10}$$

para $\ell \in \{1, ..., L\}$. Implementando a restrição, $\sum_{\ell=1}^{L} a_\ell^2 = 1$, e substituindo a_ℓ, temos

$$\frac{1}{4\lambda^2} \sum_{\ell=1}^{L} U_{nl}^2 = 1, \tag{3.11}$$

então λ assume os valores de

$$\lambda = \pm\frac{1}{2}\sqrt{\sum_{\ell=1}^{L} U_{nl}^2}, \tag{3.12}$$

onde a função será maximizada para os valores negativos de λ.

Assim, os a_ℓ que maximizam a função objetivo são dados por

$$a_\ell = \frac{1}{\sqrt{\sum_{\ell=1}^{L} U_{n\ell}^2}} \cdot U_{n\ell}. \tag{3.13}$$

Note que esses pesos ótimos podem ser calculados analiticamente. Entretanto, cada um dos U_{nl} requer o cálculo de uma matriz de distâncias $n \times n$, e temos L pesos para calcular. Esse custo computacional de ordem n^2L se torna relevante nos métodos de inferência apresentados abaixo, quando os pesos devem ser recalculados em cada iteração de um procedimento de reamostragem.

Vale ressaltar que em um primeiro momento observamos a atribuição de pesos negativos à algumas coordenadas do vetor. Pesos negativos indicam componentes do vetor de dados que são mais semelhantes entre observações de grupos distintos do que dentro dos grupos. Como isso é problemático para classificação, e na prática os pesos negativos eram muito pequenos, decidimos zerar os pesos negativos e observar a performance do método. Comparando os resultados obtidos em um estudo piloto, observamos o percentual de classificações corretas utilizando os pesos calculados, e potencialmente menores que zero, versus o uso de apenas pesos positivos (zerando as demais componentes). Com base nessa análise, optamos por zerar os pesos negativos, utilizando apenas pesos maiores ou iguais a zero.

3.2 Teste U otimizado

O primeiro passo antes de fazer a classificação de uma nova observação é verificar se os dois grupos de referência são de fato distintos e para isso é adequado utilizar o teste U da seção 2.2. Dado que o teste U é fundamentado na estatística B_n e que estamos sugerindo que essa estatística seja baseada na distância otimizada, propomos uma variante do teste U, que utiliza a distância otimizada para avaliar a separação de dois grupos pré-definidos.

Retomando o conceito das hipóteses do teste U, temos que a hipótese nula afirma que os dois grupos advém da mesma distribuição, portanto $\mathbb{E}(B_n) = 0$, a hipótese alternativa afirma que os grupos são de fato separados com $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ e consequentemente $\mathbb{E}(B_n) > 0$. A distribuição da estatística do teste sob H_0 é construída levando em conta a normalidade assintótica de B_n.

Note que como os pesos a_ℓ são estatísticas calculadas com base na própria amostra, esse procedimento deve ser levado em consideração na hora de determinar a distribuição de B_n sob H_0, sob pena de inflar o erro do tipo I. Enquanto Pinheiro et al. (2009) e Valk and Pinheiro (2012) mostram a normalidade assintótica de B_n para uma ampla gama de distâncias $\phi(\cdot,\cdot)$ e distribuições de $\mathbf{X}$, no caso da distância otimizada da seção 3.1 não dispomos de tal resultado teórico. As simulações apresentadas na seção 4.3 dão indícios de que possivelmente a normalidade assintótica seja válida, mas não está claro sob quais condições. Assim, a significância do teste é medida com base em comparação com quantil empírico da distribuição obtida através do procedimento de reamostragem sob H_0. Como no procedimento original, para cada uma das R replicações, as observações dos diferentes grupos são reunidas e posteriormente rearranjadas em dois grupos de tamanhos n_1 e n_2. Na sequência, os pesos a_ℓ^r e a estatística B_n^r são calculados novamente para cada iteração. O p-valor empírico é estimado observando a proporção de vezes que $B_n > B_n^r$. Assim, rejeita-se H_0 se o p-valor empírico calculado for maior que α, desse modo podemos afirmar que, ao nível de significância α, há evidências de que os grupos são de fato distintos.

3.3 Teste de classificação otimizado

Esta seção traz o teste de classificação otimizado, que se baseia na estrutura do teste proposto por Lacerda (2022), referenciado na seção 2.4, mas ao invés de utilizar uma distância $\phi(\cdot,\cdot)$ qualquer, o cálculo das distâncias utiliza a distância otimizada, visto na seção 3.1.

O teste de hipóteses para a classificação é realizado em duas etapas, sendo a primeira o processo de identificação do grupo mais verossímil para $\mathbf{X}^*$, utilizando para isso o classificador otimizado DB_n. Em seguida, esse grupo será chamado de G_2 e o grupo complementar de G_1, assim temos que

$$\begin{cases} H_0 : \mathbf{X}^* \in G_2; \\ H_1 : \mathbf{X}^* \in G_1. \end{cases}$$

Caso haja a rejeição de H_0, temos evidências suficientes para garantir que o novo elemento está corretamente classificado em G_1 a um nível de significância α.

No teste original a distribuição da estatística do teste considera normalidade assintótica e média e variância da estatística DB_n sob H_0 é estimada por reamostragem. Mas, como visto na seção 3.2, para o teste otimizado não dispomos do resultado referente a normalidade assintótica. Portanto, a significância estatística do teste será baseada no quantil empírico da distribuição obtida por meio de reamostragem. O processo de reamostragem segue os passos descritos na seção 2.4 para o método de Lacerda (2022), entretanto é necessário que a cada repetição os pesos associados ao vetor de dados sejam recalculados antes de calcular DB_n.

Capítulo 4

Simulações

Este Capítulo do trabalho irá avaliar os resultados do método proposto por meio de simulações de Monte Carlo comparando-os com outros métodos. Para isso, foram avaliados diferentes cenários, que variam em tamanho de amostra n, grau de separação de grupos μ, além de variar o número de componentes do vetor de dados ℓ^*, onde há diferença média entre grupos.

4.1 Classificador DB_n otimizado

Nesta seção os resultados obtidos a partir do uso do classificador DB_n otimizado serão apresentados. O primeiro resultado a ser verificado é a performance do DB_n com a distância otimizada como classificador. Para isso, a performance do método otimizado foi compara ao método que emprega o DB_n original com a distância euclidiana. Com essa finalidade, conduziu-se o seguinte processo de simulação, onde foram geradas amostras aleatórias de tamanhos $n \in \{30, 50, 100\}$, com dimensão $L = 1.000$ com entradas do vetor independentes. Para cada cenário foram consideradas $R = 500$ replicações. O grupo G_1 foi composto por ℓ^* entradas com distribuição $N(\mu, 1)$, onde $\mu \in \{0.05, 0.10, 0.20, 0.40, 0.80, 1.00, 1.50, 2.00, 2.50\}$, e as demais ℓ^- entradas com distribuição $N(0, 1)$. Já no grupo G_2 todas entradas foram geradas a partir da distribuição $N(0, 1)$. Com o objetivo de verificar a classificação de $\mathbf{X}^*$, um novo elemento foi gerado a partir de G_1, com ℓ^* entradas de distribuição $N(\mu, 1)$, e sua correta classificação foi observada.

A partir do processo de simulação descrito acima, o primeiro passo para o cálculo da distância otimizada é a estimação dos pesos associados ao vetor de dados dos grupos pré-existentes, calculando os a_ℓ's que maximizam a separação dos grupos. Em seguida, o novo elemento $\mathbf{X}^*$, gerado a partir de G_1, foi integrado à amostra e nesse momento a distância otimizada foi calculada utilizando os a_ℓ's previamente definidos. Inicialmente foi considerando $\mathbf{X}^* \in G_1$ e a seguir $\mathbf{X}^* \in G_2$ no calculo da estatística B_n e após isso, empregou-se o classificador DB_n para indicar em qual dos grupos o $\mathbf{X}^*$ deveria ser classificado.

Ao final de cada simulação, foi calculado o percentual de classificações corretas, considerando a distância otimizada e também o método original com a distância euclidiana, obtendo assim a performance dos dois métodos. Os resultados detalhados podem ser encontrados na Tabela A.1 do anexo.

As Figuras 4.1, 4.2 e 4.3 apresentam, para diferentes valores de n, o comportamento do percentual de classificações corretas para os dois métodos, observados tanto quando variamos o número de coordenadas relevantes para separação de grupos ℓ^*, quanto à medida que a diferença média entre grupos cresce.

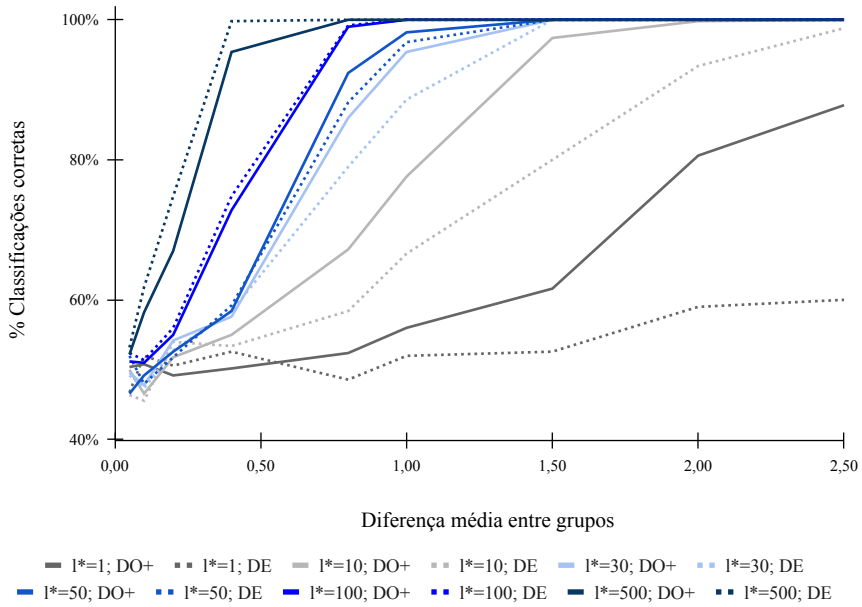

Figura 4.1: Comparativo dos percentuais de classificação correta para $n = 30$. As linhas contínuas representam a performance a partir do método otimizado (DO$^+$) e as linhas pontilhadas o método considerando a distância euclidiana (DE).

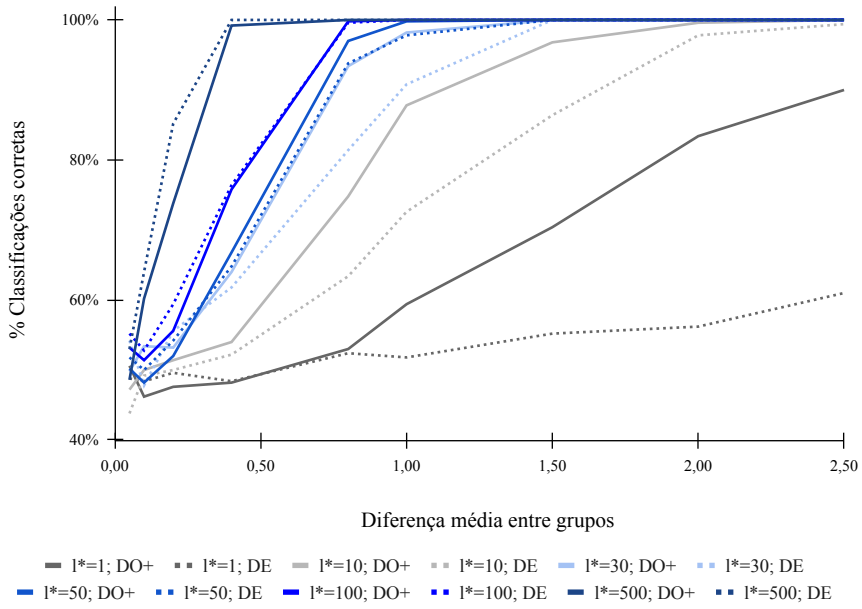

Figura 4.2: Comparativo dos percentuais de classificação correta para $n = 50$. As linhas contínuas representam a performance a partir do método otimizado (DO^+) e as linhas pontilhadas o método considerando a distância euclidiana (DE).

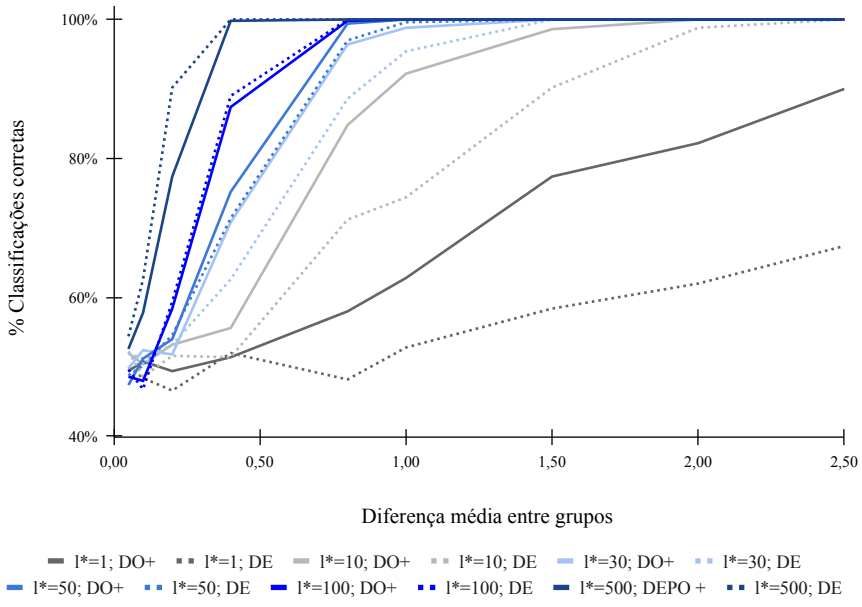

Figura 4.3: Comparativo dos percentuais de classificação correta para $n = 100$. As linhas contínuas representam a performance a partir do método otimizado (DO^+) e as linhas pontilhadas o método considerando a distância euclidiana (DE).

Pode-se observar que, para os três tamanhos de amostra, há um ganho no percentual de classificações corretas quando empregamos o classificador otimizado (DO^{+}), em relação ao método que emprega a distância euclidiana (DE). Essa diferença acontece principalmente nos cenários onde há menos componentes relevantes para separação ℓ^{*}, ou seja, onde a diferença média entre os grupos está concentrada. Já para os casos onde a diferença média entre grupos está mais distribuída entre os componentes de vetores, o método tradicional se mostra melhor em termos de performance.

Um resultado já esperado e que pode ser observado nas Figuras 4.4, 4.5 e 4.6 é que a média dos pesos associados aos componentes de vetores de dados ℓ^{*}, gerados a partir de $N(\mu, 1)$, apresentam maiores valores, enquanto para os demais componentes ℓ^{-}, que não contribuem para a separação dos grupos, esses valores giram em torno de zero. Os resultados detalhados podem ser encontrados na Tabela A.2 do anexo.

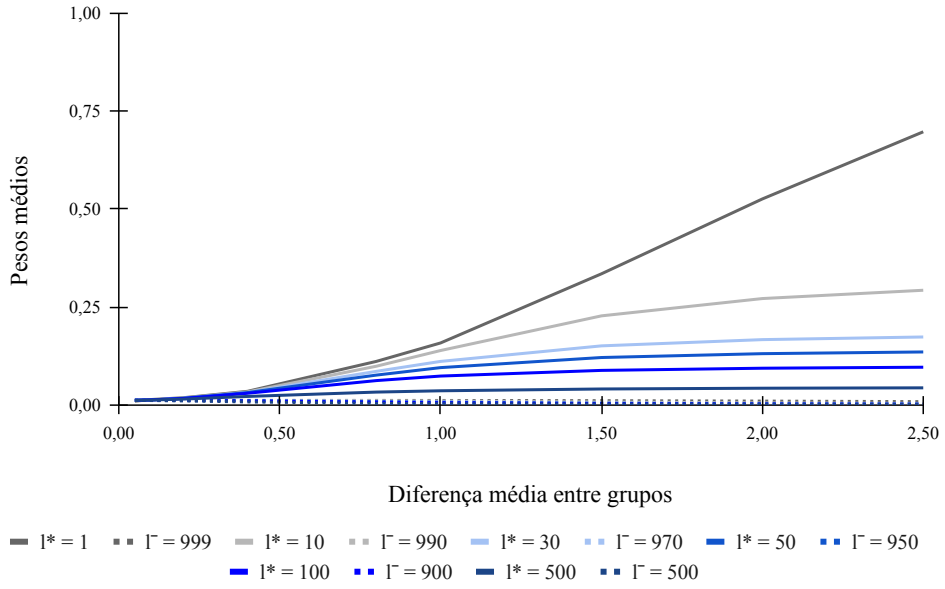

Figura 4.4: Pesos médios associados às ℓ^{*} componentes do vetor relevantes para separação dos grupos e às demais $\ell^{-} = 1.000 - \ell^{*}$ componentes, para $n = 30$

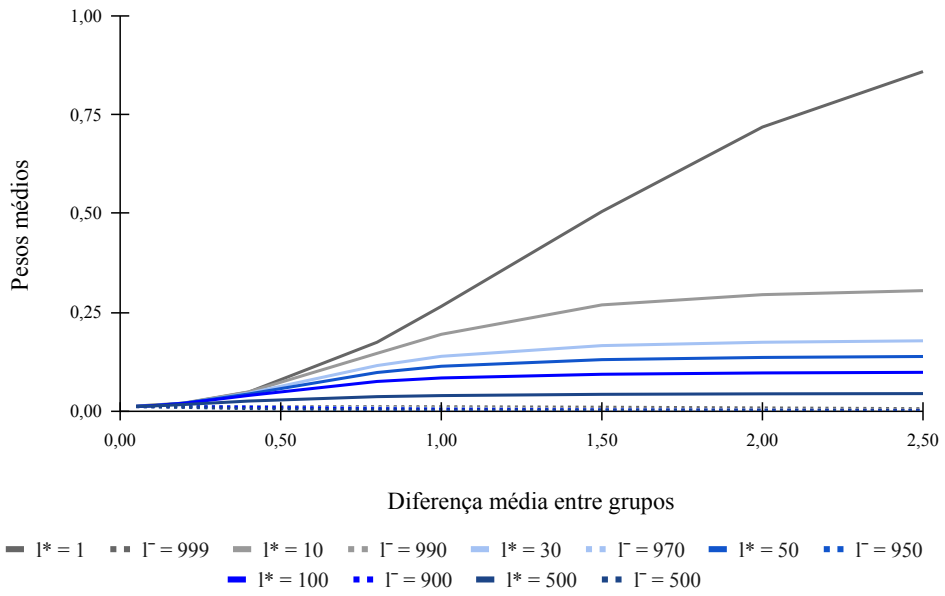

Figura 4.5: Pesos médios associados às ℓ^* componentes do vetor relevantes para separação dos grupos e às demais $\ell^- = 1.000 - \ell^*$ componentes, para $n = 50$

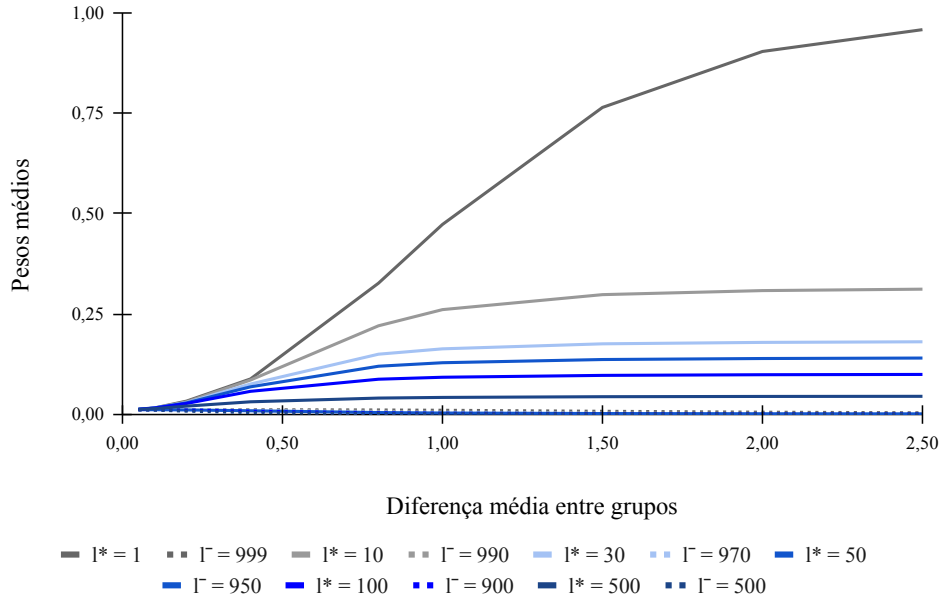

Figura 4.6: Pesos médios associados às ℓ^* componentes do vetor relevantes para separação dos grupos e às demais $\ell^- = 1.000 - \ell^*$ componentes, para $n = 100$

4.2 Teste U otimizado

Nesta seção iremos avaliar a performance do teste U otimizado, os cenários para a simulação de Monte Carlo foram configurados a partir de uma amostra aleatória de tamanho $n = 50$, com dimensão $L = 1.000$ com as componentes do vetor simuladas de modo independente. Para cada cenário foram realizadas $R = 100$ replicações. No grupo G_1 os vetores de dados foram formados por $\ell^* \in \{1, 10, 30, 50, 100, 500\}$ entradas com distribuição $N(\mu, 1)$, onde $\mu \in \{0, 0.5, 0.75, 1.00, 1.50, 2.00\}$ e as demais ℓ^- entradas com distribuição $N(0, 1)$, já no grupo G_2 todas entradas tiveram distribuição $N(0, 1)$.

Buscando entender o comportamento do teste U otimizado, optou-se por também aplicar o teste U padrão aos cenários descritos e assim comparar as performances. A Tabela 4.1 apresenta os percentuais de separação de grupos ao nível α de 5%, considerando a aplicação do teste U otimizado, apresentado na Seção 3.2 e o teste U padrão, apresentado na Seção 2.2. Note que temos sob $H_0 : \mathbb{E}(B_n) = 0$, e os grupos são homogêneos, versus $H_1 : \mathbb{E}(B_n) > 0$, em que os grupos são distintos.

Para essa configuração de dados, o erro do tipo I, $\mathbb{P}(\text{rejeitar } H_0 | H_0 \text{ é verdadeira})$, estimado foi menor do que 1% (correspondendo a zero replicações) para ambos os métodos. Note que para ambos os métodos à medida que ℓ^* e μ aumentam o poder vai para 1, mas nos cenários em que 100% de rejeição não foi atingido, o teste U otimizado se mostrou mais poderoso que o teste U padrão. Mesmo para $\ell^* = 1$ o teste U otimizado consegue capturar a diferença entre grupos, quando esses grupos são bem separados, enquanto o teste U padrão não captura.

Tabela 4.1: Comparativo do percentual de testes significativos entre os testes U otimizado e U padrão

		Teste U					
	Método	$\mu = 0$	$\mu = 0.5$	$\mu = 0.75$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\ell^* = 1$	Otimizado	0%	0%	0%	1%	13%	63%
	Padrão	0%	0%	0%	0%	0%	0%
$\ell^* = 10$	Otimizado	0%	0%	3%	46%	100%	100%
	Padrão	0%	0%	0%	0%	9%	99%
$\ell^* = 30$	Otimizado	0%	0%	73%	100%	100%	100%
	Padrão	0%	0%	0%	58%	100%	100%
$\ell^* = 50$	Otimizado	0%	7%	100%	100%	100%	100%
	Padrão	0%	0%	46%	100%	100%	100%
$\ell^* = 100$	Otimizado	0%	80%	100%	100%	100%	100%
	Padrão	0%	13%	100%	100%	100%	100%
$\ell^* = 500$	Otimizado	0%	100%	100%	100%	100%	100%
	Padrão	0%	100%	100%	100%	100%	100%

4.3 *Distribuição DB_n otimizado*

Nesta seção serão apresentados os resultados da avaliação do comportamento da distribuição do classificador DB_n otimizado a partir dos cenários propostos. Foi simulada uma amostra aleatória de tamanho $n = 50$, com dimensão $L = 1.000$ com as componentes do vetor simuladas de modo independente. Para cada cenário foram realizadas $R = 100$ replicações. No grupo G_1 os vetores de dados foram formados por $\ell^* \in \{1, 10, 30, 50, 100, 500\}$ entradas com distribuição $N(\mu, 1)$, onde $\mu \in \{0, 0.5, 0.75, 1.00, 1.50, 2.00\}$ e as demais ℓ^- entradas com distribuição $N(0, 1)$, já o grupo G_2 todas entradas com distribuição $N(0, 1)$. Um novo elemento, $\mathbf{X}^*$, foi gerado a partir de G_1, com ℓ^* entradas de distribuição $N(\mu, 1)$.

Conforme exposto na Seção 3.1, os métodos otimizados empregam as próprias observações para a determinação dos pesos da distância ponderada. Nesse contexto, não dispomos de resultado assintótico para B_n, nem está claro se e sob quais condições isso poderia valer. Assim, torna-se mais relevante o estudo empírico da distribuição de DB_n.

Por meio da Figura 4.7 podemos analisar o comportamento da estatística DB_n para os diferentes cenários propostos.

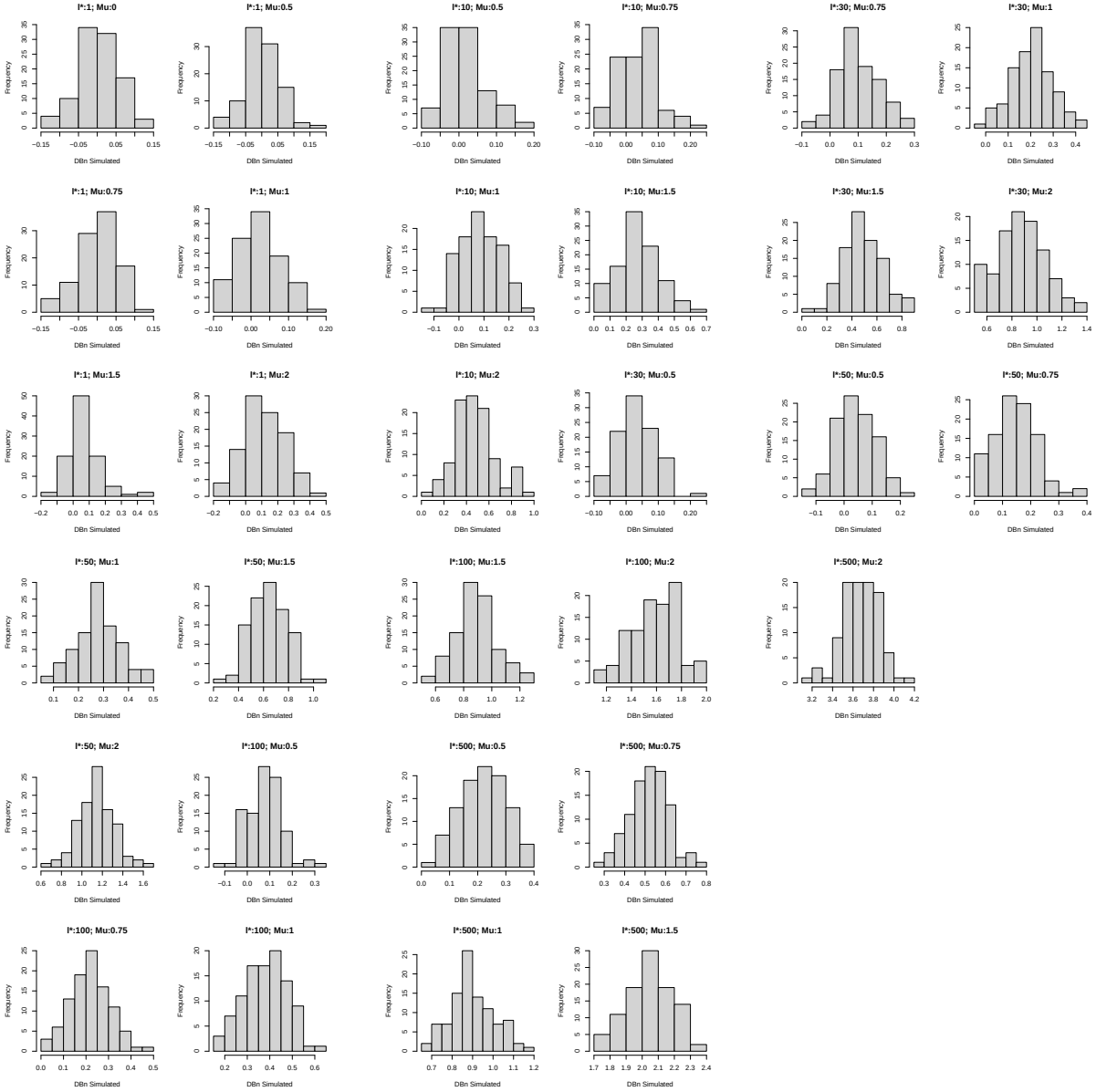

Figura 4.7: Distribuição da estatística DB_n otimizada

Para auxiliar na validação da normalidade da distribuição da estatística DB_n também foi aplicado o teste de Shapiro-Wilk para normalidade, onde a hipótese nula afirma que uma amostra tem distribuição normal e a hipótese alternativa é de não normalidade. Apenas para três cenários o teste não foi significativo ao nível de 5% ($\mu = 0,75$ e $\ell^* = 1$; $\mu = 1,5$ e $\ell^* = 1$; $\mu = 0,5$ e $\ell^* = 10$), todos os demais cenários apresentaram $p - valor > \alpha$. Portanto, temos evidências de que a estatística DB_n advém de uma distribuição normal.

Esse resultado fornece indícios de que sob certas condições a normalidade assintótica pode valer para a DB_n otimizada. Provar tal resultado, se de fato válido, e encontrar as condições para ele está fora do escopo dessa dissertação, e é um tópico de estudos futuros. Enquanto não dispomos de um resultado teórico, fazemos a escolha de utilizar diretamente os quantis da distribuição empírica ao invés de utilizar a normalidade assintótica para a obtenção de $p - valor$ nos testes otimizados.

Tabela 4.2: Percentual de classificações significativas dos testes de classificação otimizado e padrão para o real grupo de pertencimento G_1, onde H_0: $\mathbf{X}^* \in G_2$

Classificações significativas $\mathbf{X}^* \in G_1$							
	Método	$\mu = 0$	$\mu = 0.5$	$\mu = 0.75$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\ell^* = 1$	Otimizado	1%	8%	2%	11%	28%	48%
	Padrão	2%	4%	3%	5%	4%	9%
$\ell^* = 10$	Otimizado	8%	11%	21%	57%	98%	100%
	Padrão	2%	7%	7%	17%	48%	89%
$\ell^* = 30$	Otimizado	3%	23%	67%	97%	100%	100%
	Padrão	1%	6%	30%	68%	100%	100%
$\ell^* = 50$	Otimizado	5%	34%	94%	100%	100%	100%
	Padrão	2%	19%	68%	96%	100%	100%
$\ell^* = 100$	Otimizado	3%	63%	99%	100%	100%	100%
	Padrão	6%	55%	100%	100%	100%	100%
$\ell^* = 500$	Otimizado	7%	99%	100%	100%	100%	100%
	Padrão	4%	100%	100%	100%	100%	100%

4.4 *Teste de classificação otimizado*

Nesta seção serão apresentados os resultados do teste de classificação otimizado. Para avaliar o desempenho do teste de classificação consideramos os cenários configurados a partir de uma amostra aleatória de tamanho $n = 50$, com dimensão $L = 1.000$ com as componentes do vetor simuladas de modo independente. Para cada cenário foram realizadas $R = 100$ replicações. No grupo G_1 os vetores de dados foram formados por $\ell^* \in \{1, 10, 30, 50, 100, 500\}$ entradas, com distribuição $N(\mu, 1)$, onde $\mu \in \{0, 0.5, 0.75, 1.00, 1.50, 2.00\}$ e as demais ℓ^- entradas com distribuição $N(0, 1)$. Já no grupo G_2, os vetores de dados foram formados com todas entradas com distribuição $N(0, 1)$. Um novo elemento, $\mathbf{X}^*$, foi gerado a partir de G_1, com ℓ^* entradas de distribuição $N(\mu, 1)$.

As Tabelas 4.2 e 4.3 apresentam os resultados do teste de classificação, mensurando os percentuais de classificações significativas, ao de 5% nível de significância [1]. A Tabela 4.2, apresenta apenas os percentuais de classificações significativas para os $\mathbf{X}^*$ classificados em G_1, com H_0 :$\mathbf{X}^* \in G_2$. Já a Tabela 4.3, mostra os resultados da classificação significativa para $\mathbf{X}^*$ classificados em G_2, com H_0 :$\mathbf{X}^* \in G_1$.

Na Tabela 4.2, estamos considerando o teste com H_0 :$\mathbf{X}^* \in G_2$, quando a nova observação é simulada em G_1, portanto a tabela avalia poder dos testes. Podemos observar uma performance superior do método proposto quando comparado ao método padrão em todos cenários observados, notando que em todos os casos considerados as diferenças médias entre grupos estão concentradas em uma fração das componentes do vetor de dados. Esses resultados estão em linha com o ganho nas simulações do classificador na Seção 4.1. Podemos observar que para $\ell^* = 1$, à medida que a diferença média entre grupos cresce, o percentual de classificações significativas em G_1 aumenta no método otimizado, enquanto para o teste de classificação padrão esse percentual permanece baixo.

[1] Note que os p-valores empíricos são comparados com $\alpha/2 = 0.025$ devido à correção de múltiplos testes descrita na Seção 2.4

Tabela 4.3: Percentual de classificações significativas dos testes de classificação otimizado e padrão no grupo de comparação G_2, onde H_0: $\mathbf{X}^* \in G_1$. Dados simulados com $\mathbf{X}^* \in G_1$.

Classificações significativas $\mathbf{X}^* \in G_2$							
	Método	$\mu = 0$	$\mu = 0.5$	$\mu = 0.75$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\ell^* = 1$	Otimizado	3%	8%	4%	5%	5%	8%
	Padrão	1%	5%	3%	2%	3%	4%
$\ell^* = 10$	Otimizado	3%	2%	0%	5%	5%	6%
	Padrão	2%	3%	3%	2%	4%	5%
$\ell^* = 30$	Otimizado	3%	5%	1%	6%	5%	6%
	Padrão	2%	6%	3%	5%	2%	3%
$\ell^* = 50$	Otimizado	8%	6%	3%	1%	3%	5%
	Padrão	2%	3%	2%	4%	3%	3%
$\ell^* = 100$	Otimizado	4%	4%	5%	3%	4%	6%
	Padrão	7%	6%	6%	3%	6%	4%
$\ell^* = 500$	Otimizado	6%	7%	3%	7%	4%	4%
	Padrão	1%	4%	5%	7%	3%	1%

Na Tabela 4.3, estamos sob H_0 :$\mathbf{X}^* \in G_1$, porém nesse estudo de simulação todos os $\mathbf{X}^*$ são provenientes de G_1, ou seja, estamos investigando se há classificações significativas no grupo incorreto G_2. Ambos os métodos apresentaram desempenho próximo ao esperado, em torno dos 5% (ou um pouco mais) de erro do tipo I, com uma oscilação observada provavelmente atrelada ao baixo número de replicações R.

Capítulo 5

Aplicação

Nesta seção será apresentada uma aplicação do método proposto em uma base de dados reais.

Um exemplo típico do cenário de HDLSS na área do câncer é o estudo que associa dados de microarranjos a diferentes tipos de linfoma. Nele foram avaliadas 6.817 medidas refentes a expressão gênica de 77 pacientes, em dois grupos de observação: 58 pacientes portadores do linfoma difuso de grandes células B (DLBCLs); 19 pacientes portadores de linfoma folicular (FL). Os dados foram obtidos do trabalho *"The Diffuse Large B-cell Lymphoma (DLBCL)"*, apresentado por Shipp et al. (2002) e podem ser encontrados em: https://github.com/ramhiser/ datamicroarray/wiki/Shipp.

A opção por esse conjunto de dados é a comparação do método otimizado com os resultados obtidos por Lacerda (2022). O seguinte cenário, baseado no trabalho de Lacerda (2022), foi considerado: aproximadamente 2/3 dos dados foram reamostrados de forma aleatória para compor o conjunto de treinamento (grupos G_1 e G_2 de referência) e os demais formaram o conjunto de teste. Sua classificação pelo método de interesse foi observada, e esse processo foi repetido $R = 50$ vezes. O conjunto de treinamento ficou estruturado da seguinte forma: $n_{DLBCL} = 40$ e $n_{FL} = 12$, já o conjunto de teste, $n_{DLBCL} = 18$ e $n_{FL} = 7$.

A Tabela 5.1 apresenta o resultado da classificação pelo método padrão e otimizado para os dados do estudo de linfoma e, a título de comparação, um algoritmo tradicional de classificação, o Naive Bayes. É possível observar que para o grupo de pacientes portadores de DLBCL 735 dos 900 pacientes (total do grupo teste nas 50 replicações) foram classificados corretamente, representando aproximadamente 82% de classificações corretas. O grupo dos pacientes portadores de FL, 324 dos 350 pacientes foram corretamente classificados, cerca de 93% de acerto nas classificações. Ao todo foram observadas praticamente 85% de classificações corretas para este estudo. Já aplicando o classificador padrão, proposto por Lacerda (2022), observamos 686 pacientes classificados corretamente como DLBCL, mais de 76% de classificações corretas e 85% de classificações corretas para o grupo FL, 299 dos 350, representando ao todo um percentual de 78% de classificações corretas para esse conjunto de dados.

Ainda, em relação ao desempenho de classificação do método otimizado versus o método padrão e o Naive Bayes, podemos comparar a taxa de erros de classificação em cada um dos grupos pela Tabela 5.2 a seguir. Note que $E(FL|DLBCL)$ é o percentual de pacientes classificados incorretamente como portadores de FL, sendo que são portadores de DLBCL. E $E(DLBCL|FL)$ representa o percentual de pacientes classificados erroneamente como portadores de DLBCL, sendo que são portadores de FL.

Tabela 5.1: Número de classificações em cada categoria para o método otimizado, padrão e Naive Bayes.

Predito|Real	Método Otimizado	Método Padrão	Naive Bayes
(DLBCL|DLBCL)	735	686	835
(FL|DLBCL)	165	214	65
(DLBCL|FL)	26	51	211
(FL|FL)	324	299	139

Tabela 5.2: Taxa de erro de classificação para o método otimizado, padrão e Naive Bayes.

	Classificador Otimizado	Classificador Padrão	Naive Bayes
E(FL|DLBCL)	18,33%	23,78%	7,22%
E(DLBCL|FL)	7,43%	14,57%	60,29%

É possível observar o ganho de performance do classificador otimizado frente ao classificador padrão, para o erro de classificação DLBCL como FL o classificador otimizado produziu cerca de 23% menos erro nas classificações do que o classificador padrão. Já em relação ao erro de classificar FL no grupo DLBCL o método otimizado produziu aproximadamente 50% menos erro frente ao método padrão. Em relação ao Naive Bayes, é possível observar um menor percentual de classificações incorretas em $E(FL|DLBCL)$ comparado aos métodos que utilizam U-estatísticas. Porém, quando observamos o percentual de erro para o menor grupo, este é muito maior que os erros cometidos pelos outros métodos.

Vale ressaltar aqui a presença de desbalanceamento dos grupos, acarretando em dificuldades para acertar classificações em grupos menores. Esse efeito é bastante pronunciado para o Naive Bayes, já nos métodos baseados em U-estatística o problema é mitigado.

O próximo passo é entender se as classificações realizadas foram significativas, e para isso vamos aplicar o teste de classificação otimizado e compará-lo ao teste de classificação padrão. As hipóteses dos teste estão de acordo com o apresentado na Seção 3.3, em que a hipótese alternativa se refere ao grupo em que a observação foi classificada. A Tabela 5.3 apresenta o percentual de classificações significativas para todas as classificações realizadas.

Comparando os percentuais das classificações significativas de ambos métodos, observamos um maior percentual de classificações significativas, a um nível $\alpha = 5\%$[1], no método otimizado quando

[1]Note que os p-valores empíricos são comparados com $\alpha/2 = 0.025$ devido à correção de múltiplos testes descrita na

Tabela 5.3: Percentual de classificações significativas dos métodos otimizado e padrão

Predito|Real	Método Otimizado	Método Padrão
(DLBCL|DLBCL)	90,75%	74,93%
(FL|DLBCL)	22,42%	14,02%
(DLBCL|FL)	23,08%	0,00%
(FL|FL)	47,84%	27,09%

consideramos as classificações corretas (maior sensibilidade). Já o classificador padrão, proposto por Lacerda (2022), apresenta um menor percentual de resultados significativos para as classificações incorretas (maior especificidade). O método otimizado classificou significativamente cerca de 69% das classificações realizadas, enquanto o método padrão 50%.

Note que, embora o Naive Bayes não possua teste de hipóteses para a classificação, é possível quantificar a confiança na classificação no Naive Bayes por meio da probabilidade a posteriori pp da classificação. Para comparar essas probabilidades com o resultado dos testes dos outros métodos, consideramos uma heurística que designa como alta probabilidade aquelas classificações com mais 95% de pp. Entretanto, observamos que nas nossas análises, todas as classificações tiveram valores de pp extremamente altos, de modo que 100% das classificações realizadas seriam consideradas de alta probabilidade, inclusive as erradas. Essa confiança excessiva em todas as classificações é consequência do contexto HDLSS, e destaca a importância de empregar métodos apropriados ao contexto.

Dado os resultados não tão favoráveis de inferência do nosso método frente ao original, é relevante avaliar se o método otimizado é o mais adequado para esta aplicação. Lembramos que o nosso método apresentou melhor performance quando tínhamos diferença entre grupos mais concentrada apenas em algumas coordenadas.

Uma possível abordagem para avaliar se o método otimizado é adequado para esta aplicação seria avaliar os pesos calculados para os dados, entendendo se eles estão concentrados apenas em parte das variáveis. A figura 5.1 apresenta os pesos calculados para as primeiras quatro repetições das $R = 50$ realizadas. Notamos que as figuras apresentam concentração dos pesos em poucas variáveis, indicando que estas tem maior contribuição para separação dos grupos, enquanto as demais estão com pesos zero ou muito próximo disso. Lembramos que a soma dos pesos está restrita a 1. Este é um ponto a ser melhor discutido em trabalhos futuros.

Seção 2.4

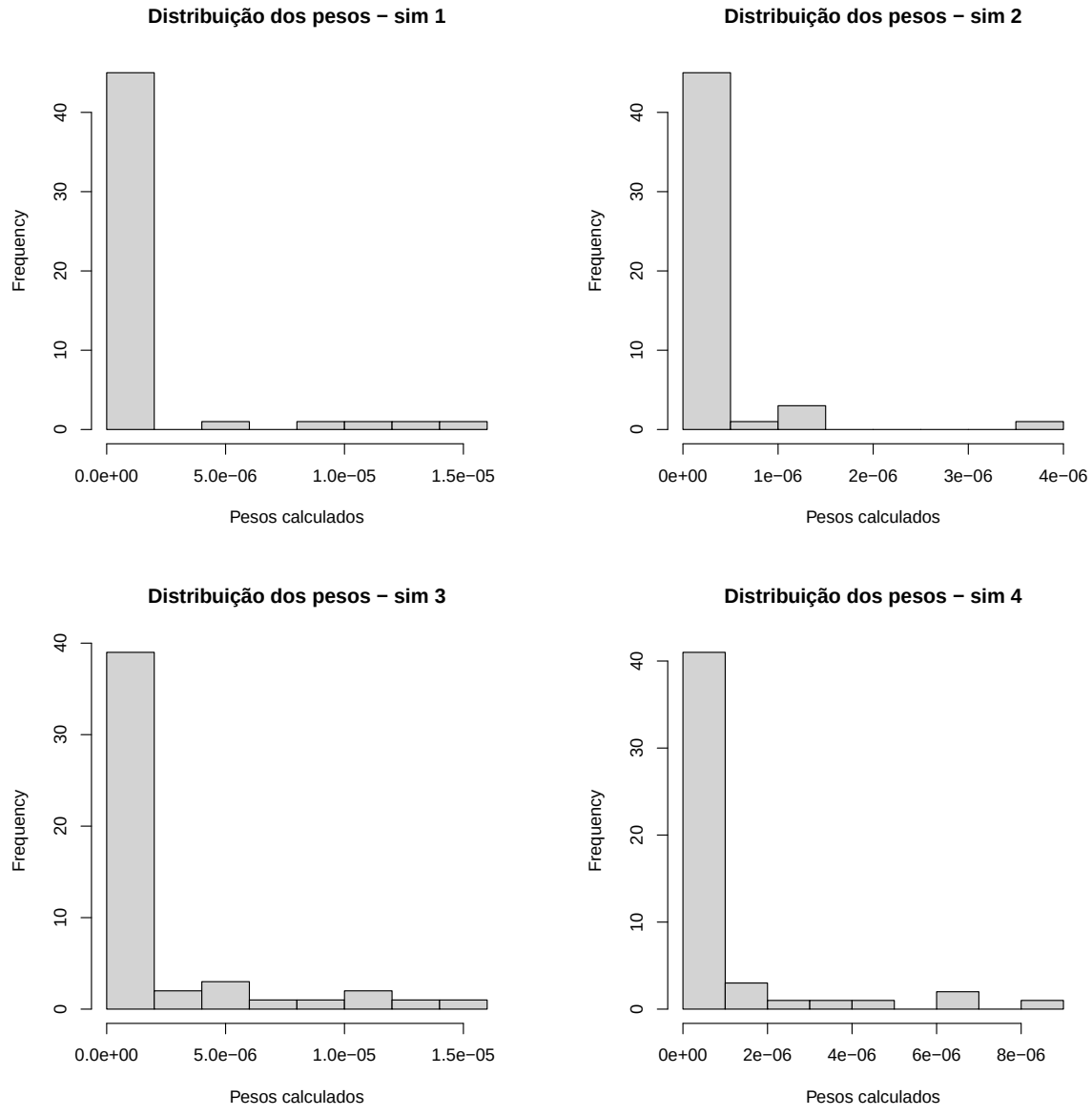

Figura 5.1: Distribuição dos pesos calculados no estudo sobre linfoma

Capítulo 6

Conclusões e trabalhos futuros

Esta dissertação se apoiou em trabalhos anteriores como o de Cybis et al. (2018) e Lacerda (2022) propondo uma modificação no cálculo das distâncias entre grupos, encontrando uma solução analítica para o problema de maximização das dissimilaridades entre grupos, isso considerando o ambiente de alta dimensão e pequeno tamanho de amostra (HDLSS). Além de realizar a classificação de um novo elemento em um de dois grupos de referência, também foi avaliado a real separação entre grupos e a significância estatística atrelada a classificação realizada.

Sabemos que em ambientes de alta dimensão e baixo tamanho de amostra métodos mais tradicionais, como os da inferência paramétrica, podem ter problemas de estimação devido à quantidade de parâmetros a serem estimados ser muito maior que a quantidade de amostras observadas. Uma alternativa é a utilização de U-estatísticas, que vem apresentando resultados promissores quanto a sua utilização em problemas de agrupamento e classificação nesse ambiente.

Ao empregar a distância otimizada construímos um classificador que é capaz de corretamente classificar conjuntos de dados com estruturas mais complexas do que aquelas encontradas pela distância euclideana convencional, principalmente quando diferentes entradas do vetor de dados tem pesos distintos para a diferenciação dos grupos. Nesse sentido, construímos um classificador potencialmente mais versátil (mesmo quando comparado com o uso de outras distâncias pré-definidas como kernel), uma vez que ele permite que os dados indiquem qual a distância apropriada para classificação. Nas comparações realizadas nesse estudo verificamos que isso resultou em menores erros de classificação, sempre que consideramos diferença entre grupos apenas em algumas componentes. Já quando todas as componentes do vetor de dados contribuem igualmente para separação dos grupos a distância euclidiana convencional performa melhor.

Nas simulações, houve um ganho no percentual de classificações corretas com a utilização do classificador otimizado, principalmente nos casos em que a diferença média entre os grupos está concentrada em algumas componentes. Esse resultado também pode ser observado no teste U otimizado, que se mostrou um teste mais poderoso que o teste U padrão nesses casos. Enquanto aqui propomos o teste U otimizado como uma etapa preliminar de validação para a classificação otimizada, o método pode ser empregado em diversos outros contextos em que queremos avaliar se dois grupos são dissimilares, além de mensurar a significância estatística dessa separação. O ganho de performance do método otimizado também foi observado na avaliação da significância estatística das classificações realizadas e deu indícios de um bom controle de erro do tipo I.

Outro ponto observado nas simulações, é que encontramos evidência para a normalidade da distribuição da estatística DB_n na maioria dos cenários de simulação. Isso é relevante dado que o método otimizado emprega as próprias observações para a determinação dos pesos da distância ponderada, o que altera a estrutura da estatística B_n, e portanto não podemos simplesmente utilizar resultados existentes de normalidade assintótica. É um ponto importante para trabalhos futuros o estudo da teoria assintótica.

É importante mencionar que as simulações realizadas consideram cenários adequados ao tipo de método proposto, nos quais esperamos que ele tenha bons resultados. Por outro lado, o custo computacional empregado dos testes otimizados é maior do que o do método original, pois a cada etapa da reamostragem os pesos associados às componentes do vetor de dados são recalculados. Vale citar que em relação os p-valores estimados a partir da técnica proposta podem variar um pouco em diferentes execuções, um exemplo disso é a variância da estatística B_n que é estimada por reamostragem.

Considerando a aplicação do método em um estudo com dados reais, notamos que o classificador otimizado apresentou menores taxas de erro que o classificador padrão. Além disso, observamos melhores resultados do método otimizado na quantidade de classificações corretas e significativas, já o método padrão se mostrou melhor em relação a quantidade de classificações incorretas que foram significativas. Ponderamos então que o aumento em sensibilidade veio às custas da especificidade.

Referências Bibliográficas

Bello, D. Z. (2021). Inferência em agrupamento considerando múltiplos grupos. Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Clémençon, S. (2014). A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42–56.

Cybis, G. B., Valk, M., and Lopes, S. R. C. (2018). Clustering and classification problems in genetics through U-statistics. *Journal of Statistical Computation and Simulation*, 88(10):1882–1902.

Hallajian, B., Motameni, H., and Akbari, E. (2022). Ensemble feature selection using distance-based supervised and unsupervised methods in binary classification. *Expert Systems with Applications*, 200:116794.

Halmos, P. R. (1946). The Theory of Unbiased Estimation. *The Annals of Mathematical Statistics*, 17(1):34 – 43.

Huang, H., Liu, Y., Yuan, M., and Marron, J. (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993.

Izbicki, R. and dos Santos, T. (2022). *Aprendizado de máquina: uma abordagem estatística*. UICLAP, São Carlos, SP.

Kalina, J. (2014). Classification methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, 34(1):10–18.

Kimes, P. K., Liu, Y., Neil Hayes, D., and Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821.

Lacerda, E. C. (2022). Classificação com inferência para dados de alta dimensão. Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Liao, S.-M. and Akritas, M. (2007). Test-based classification: A linkage between classification and statistical testing. *Statistics & Probability Letters*, 77(12):1269–1281. Silver Jubilee Issue Dedicated to Richard A. Johnson on his 70th birthday.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293.

Maitra, R., Melnykov, V., and Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, 107(497):378–392.

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469.

Modarres, R. (2022). A high dimensional dissimilarity measure. *Computational Statistics Data Analysis*, 175:107560.

Pinheiro, A., Sen, P. K., and Pinheiro, H. P. (2009). Decomposability of high-dimensional diversity measures: Quasi-U-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, 100(8):1645–1656.

Rauf Ahmad, M. and Pavlenko, T. (2018). A U-classifier for high-dimensional data under non-normality. *Journal of Multivariate Analysis*, 167:269–283.

Sen, P. K. (2006). Robust statistical inference for high-dimensional data models with application to genomics. *Austrian journal of statistics*, 35(2&3):197–214.

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *The Annals of Statistics*, 32(6):2616 – 2641.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74.

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542.

Valk, M. and Cybis, G. B. (2021). U-statistical inference for hierarchical clustering. *Journal of Computational and Graphical Statistics*, 30(1):133–143.

Valk, M. and Pinheiro, A. (2012). Time-series clustering via quasi U-statistics. *Journal of Time Series Analysis*, 33:608–619.

Wu, Y., Zhu, D., and Wang, X. (2023). Tree enhanced deep adaptive network for cancer prediction with high dimension low sample size microarray data. *Applied Soft Computing*, 136:110078.

Zhang, Z., He, J., Cao, J., and Li, S. (2023). Maximum decentral projection margin classifier for high dimension and low sample size problems. *Neural Networks*, 157:147–159.

Anexo A

Resultados complementares

Tabelas dos resultados detalhados das simulações do método proposto, discutidos na Seção 4.1 deste trabalho.

Tabela A.1: Comparativo do método otimizado versus o método utilizando a distância euclidiana dos percentuais de classificação correta para os diferentes tamanhos de amostra n, diferença média entre grupos μ e quantidade de vetores de dados ℓ^* com distribuição $N(\mu, 0)$

		$\ell^* = 1$		$\ell^* = 10$		$\ell^* = 30$		$\ell^* = 50$		$\ell^* = 100$		$\ell^* = 500$	
	μ	DO+	DE	DO+	DE	DO+	DE	DO+	DE	DO+	DE	DO+	DE
	0,05	50,40%	46,80%	50,00%	46,40%	49,60%	49,20%	46,60%	52,00%	51,20%	52,40%	52,20%	53,20%
	0,10	50,80%	52,00%	46,60%	45,60%	47,80%	47,40%	49,20%	48,00%	51,00%	51,40%	58,20%	61,80%
	0,20	49,20%	50,60%	51,80%	54,00%	54,20%	52,80%	52,60%	51,80%	55,00%	56,00%	67,00%	74,80%
	0,40	50,20%	52,60%	55,00%	53,40%	57,60%	58,40%	58,40%	59,20%	72,80%	74,80%	95,40%	99,80%
n = 30	0,80	52,40%	48,60%	67,20%	58,40%	86,00%	79,00%	92,40%	88,20%	99,00%	99,20%	100,00%	100,00%
	1,00	56,00%	52,00%	77,60%	66,60%	95,40%	88,60%	98,20%	96,80%	100,00%	100,00%	100,00%	100,00%
	1,50	61,60%	52,60%	97,40%	80,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	2,00	80,60%	59,00%	99,80%	93,40%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	2,50	87,80%	60,00%	100,00%	98,80%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	0,05	50,60%	48,80%	47,20%	43,80%	49,80%	54,20%	50,20%	51,80%	53,20%	55,20%	48,60%	53,00%
	0,10	46,20%	48,40%	50,00%	49,20%	53,40%	47,80%	48,20%	50,00%	51,40%	52,80%	60,20%	64,00%
	0,20	47,60%	49,60%	51,40%	50,00%	53,20%	55,60%	52,00%	54,20%	55,60%	59,40%	73,80%	85,20%
	0,40	48,20%	48,40%	54,00%	52,20%	64,00%	61,80%	66,80%	64,80%	75,80%	76,40%	99,20%	100,00%
n = 50	0,80	53,00%	52,40%	74,80%	63,40%	93,40%	81,40%	97,00%	93,80%	99,80%	99,60%	100,00%	100,00%
	1,00	59,40%	51,80%	87,80%	72,60%	98,20%	90,80%	99,80%	97,80%	100,00%	100,00%	100,00%	100,00%
	1,50	70,40%	55,20%	96,80%	86,40%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	2,00	83,40%	56,20%	99,60%	97,80%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	2,50	90,00%	61,00%	100,00%	99,40%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	0,05	49,40%	50,00%	52,00%	52,20%	49,80%	52,00%	47,40%	48,80%	48,60%	49,60%	52,60%	54,40%
	0,10	50,80%	48,40%	50,40%	48,60%	52,40%	51,20%	51,20%	50,20%	48,00%	46,80%	57,80%	62,60%
	0,20	49,40%	46,60%	53,20%	51,60%	51,80%	53,20%	54,00%	54,60%	58,40%	59,40%	77,40%	90,20%
	0,40	51,40%	52,00%	55,60%	51,40%	70,80%	62,60%	75,20%	71,40%	87,40%	89,00%	99,80%	100,00%
n = 100	0,80	58,00%	48,20%	84,80%	71,20%	96,40%	88,60%	99,40%	97,00%	99,80%	100,00%	100,00%	100,00%
	1,00	62,80%	52,80%	92,20%	74,40%	98,80%	95,40%	100,00%	99,60%	100,00%	100,00%	100,00%	100,00%
	1,50	77,40%	58,40%	98,60%	90,20%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	2,00	82,20%	62,00%	100,00%	98,80%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
	2,50	90,00%	67,40%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabela A.2: Pesos associados aos vetores de dados para os diferentes cenários, considerando tamanho de amostra n, diferença média entre grupos μ e quantidade de vetores de dados ℓ^* com distribuição $N(\mu, 1)$ e ℓ^- com distribuição $N(0, 1)$

	μ	$\ell^* = 1$	$\ell^- = 999$	$\ell^* = 10$	$\ell^- = 990$	$\ell^* = 30$	$\ell^- = 970$	$\ell^* = 50$	$\ell^- = 950$	$\ell^* = 100$	$\ell^- = 900$	$\ell^* = 500$	$\ell^- = 500$
	0,05	0,00880	0,01081	0,01161	0,01081	0,01132	0,01081	0,01151	0,01083	0,01112	0,01081	0,01104	0,01071
	0,10	0,01299	0,01085	0,01185	0,01087	0,01182	0,01084	0,01209	0,01081	0,01207	0,01077	0,01176	0,01046
	0,20	0,01757	0,01086	0,01615	0,01078	0,01633	0,01072	0,01617	0,01069	0,01569	0,01055	0,01411	0,00947
	0,40	0,03402	0,01077	0,03244	0,01069	0,03113	0,01042	0,03085	0,01009	0,02900	0,00951	0,02091	0,00686
n = 30	0,80	0,11043	0,01072	0,09811	0,00987	0,08501	0,00856	0,07544	0,00761	0,06141	0,00626	0,03205	0,00319
	1,00	0,15764	0,01063	0,13828	0,00915	0,11047	0,00728	0,09458	0,00617	0,07300	0,00477	0,03537	0,00233
	1,50	0,33436	0,01008	0,22689	0,00664	0,15020	0,00439	0,12040	0,00353	0,08750	0,00254	0,03998	0,00117
	2,00	0,52571	0,00904	0,27111	0,00451	0,16614	0,00274	0,13025	0,00216	0,09305	0,00153	0,04193	0,00069
	2,50	0,69734	0,00761	0,29233	0,00308	0,17289	0,00183	0,13463	0,00143	0,09561	0,00100	0,04290	0,00046
	0,05	0,01187	0,01086	0,01145	0,01084	0,01139	0,01080	0,01125	0,01082	0,01118	0,01085	0,01124	0,01062
	0,10	0,01463	0,01082	0,01239	0,01080	0,01306	0,01079	0,01295	0,01079	0,01282	0,01071	0,01227	0,01021
	0,20	0,01927	0,01079	0,02004	0,01070	0,01903	0,01068	0,01957	0,01056	0,01906	0,01027	0,01602	0,00868
	0,40	0,04731	0,01082	0,04794	0,01053	0,04462	0,01005	0,04263	0,00960	0,03892	0,00866	0,02473	0,00548
n = 50	0,80	0,17369	0,01058	0,14555	0,00897	0,11461	0,00697	0,09702	0,00595	0,07436	0,00456	0,03589	0,00220
	1,00	0,26452	0,01035	0,19374	0,00776	0,13803	0,00542	0,11287	0,00443	0,08322	0,00326	0,03856	0,00150
	1,50	0,50466	0,00916	0,26824	0,00471	0,16513	0,00288	0,12962	0,00227	0,09256	0,00162	0,04176	0,00073
	2,00	0,71861	0,00736	0,29420	0,00291	0,17365	0,00171	0,13517	0,00133	0,09593	0,00094	0,04300	0,00042
	2,50	0,85912	0,00544	0,30434	0,00191	0,17729	0,00111	0,13758	0,00086	0,09745	0,00061	0,04360	0,00027
	0,05	0,01056	0,01083	0,01201	0,01083	0,01205	0,01082	0,01222	0,01077	0,01160	0,01076	0,01161	0,01049
	0,10	0,01515	0,01083	0,01473	0,01085	0,01516	0,01073	0,01512	0,01071	0,01483	0,01054	0,01363	0,00971
	0,20	0,03217	0,01084	0,03021	0,01065	0,02863	0,01048	0,02774	0,01025	0,02631	0,00974	0,01967	0,00718
	0,40	0,08743	0,01074	0,08473	0,01009	0,07480	0,00897	0,06780	0,00812	0,05661	0,00683	0,03048	0,00365
n = 100	0,80	0,32623	0,01011	0,22000	0,00686	0,14918	0,00454	0,11917	0,00364	0,08679	0,00264	0,03977	0,00121
	1,00	0,47270	0,00936	0,26033	0,00512	0,16255	0,00318	0,12791	0,00250	0,09170	0,00179	0,04142	0,00081
	1,50	0,76448	0,00682	0,29768	0,00259	0,17505	0,00151	0,13591	0,00118	0,09635	0,00083	0,04319	0,00037
	2,00	0,90403	0,00456	0,30778	0,00150	0,17862	0,00086	0,13842	0,00067	0,09800	0,00048	0,04385	0,00021
	2,50	0,95845	0,00304	0,31134	0,00097	0,18013	0,00056	0,13957	0,00043	0,09872	0,00031	0,04416	0,00014