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

Classificação com inferência para dados de alta dimensão

Eduardo Cavalli Lacerda

Porto Alegre, 3 de Maio de 2022.

CIP - Catalogação na Publicação

Lacerda, Eduardo
Classificação com inferência para dados de alta dimensão / Eduardo Lacerda. -- 2022.
65 f.
Orientador: Marcio Valk.

Coorientador: Gabriela Cybis.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística, Programa de Pós-Graduação em Estatística, Porto Alegre, BR-RS, 2022.

1. Método de classificação. 2. Inferência. 3. High Dimensional Low Sample Size (HDLSS). 4. U-estatística. I. Valk, Marcio, orient. II. Cybis, Gabriela, 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 Eduardo Cavalli Lacerda 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): Marcio Valk

Prof. Dr. Marcio Valk

Co-orientador(a):

Prof. Dra. Gabriela Cybis

Comissão Examinadora:

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

Prof. Dr. Luiz Emílio Allem (PPGMAp-UFRGS)

Prof. Dra. Márcia Helena Barbian (PPGEst- UFRGS)

Data de Apresentação: 1 de Junho de 2022

"Temos que continuar aprendendo. Temos que estar abertos.
E temos que estar prontos para espalhar nosso conhecimento
a fim de chegar a uma compreensão mais elevada da realidade."
(Thich Nhat Hanh)

Agradecimentos

Agradeço primeiramente a Deus pela vida, por tudo que Ele me proporcionou e por dar-me força nos momentos de decisão. Aos meus pais, Hilbo Barbosa Lacerda e Nilza Cavalli Lacerda, pela base, educação, incentivo e apoio que sempre souberam me dar em todos os momentos de minha vida. A vocês que, muitas vezes, renunciaram aos seus sonhos para que eu pudesse realizar o meu. Ao meu orientador, Prof. Dr. Marcio Valk, exemplo de pessoa e profissional, pelos ensinamentos, correções, incentivo, paciência e por toda dedicação nesta dissertação. Muito obrigado por tudo! A minha coorientadora, Prof. Dra. Gabriela Cybis, pelas importantes e valiosas sugestões que contribuíram para o resultado final do trabalho.

Resumo

Neste trabalho propomos um método de classificação com inferência para dois ou mais grupos no contexto de alta dimensionalidade e baixo tamanho amostral. Nesse contexto, o método de classificação proposto é comparado com uma metodologia recentemente proposta, através de simulações e aplicação a dados reais. Além disso, um teste de hipóteses é proposto e as propriedades assintóticas da estatística de teste são obtidas, no entanto a estimação da variância se dá a partir de um procedimento de reamostragem. Resultados das simulações mostram que o classificador é competitivo com a metodologia existente e a possibilidade de identificar se a classificação em um determinado grupo é estatisticamente significativa possibilita controlar o erro do tipo I, mostrando-se uma importante ferramenta em problemas de classificação.

Palavras-chave: Inferência; High Dimensional Low Sample Size (HDLSS); Método de classificação; U-estatística.

Abstract

In this work we propose a classification method with inference for two or more groups in the high dimensional low sample size context. The classification method is compared with a recently proposed methodology, through simulations and application to a real dataset. Furthermore, a hypothesis test is proposed and the asymptotic properties of the test statistics are obtained, however the estimation of the variance is given from a procedure resampling process. Simulation results show that the classifier is competitive with the existing methodology and the possibility of identifying whether the classification in a certain group is statistically significant makes it possible to control the type I error, proving to be an important tool in classification problems.

Key words: Inference; High Dimensional Low Sample Size (HDLSS); Classification method; U-statistics.

Índice

Capítulo 1

Introdução

Neste trabalho propomos um método de classificação com inferência em dados de alta dimensionalidade com baixo tamanho amostral, conhecidos na literatura como *High Dimension Low Sample Size* (HDLSS) Hall et al. (2005), em um contexto em que se têm dois (ou mais) grupos bem definidos. O recente trabalho de Ahmad and Pavlenko (2018) apresenta um classificador para esse tipo de dados e estuda as propriedades teóricas desse estimador, como a probabilidade de erro de classificação e distribuição assintótica. O método proposto por Ahmad and Pavlenko (2018) é usado aqui como *benchmark* por se propor a solucionar o mesmo problema, tendo características comuns ao que estamos propondo, sendo não paramétrico e não exigindo normalidade dos dados. No entanto, essa metodologia trata cada nova observação a serem classificadas de forma única, atribuindo-se de maneira uniforme uma probabilidade de erro de classificação. Por esse motivo, além do classificador, propomos também um teste de hipóteses para verificar se a classificação obtida é estatisticamente significativa.

De forma geral, classificação pode ser caracterizada como o problema que consiste em identificar a qual de um conjunto de categorias conhecidas (grupos) pertence uma nova observação. Esses grupos são formados naturalmente, ou seja, são inerentes aos dados (dados genéticos de câncer A e câncer B) ou determinados (ativos financeiros com risco alto, moderado e baixo) e para cada uma das situações, diferentes características dos seus elementos são consideradas para a formação/determinação destes grupos (Rosenberg et al. (2002); Chen et al. (2015); Motlagh et al. (2019); Hennig (2015)). Nesse sentido classificação têm como intuito alocar elementos/amostras de uma população em uma, duas ou mais categorias/grupos, tomando como base um conjunto de características em cada elemento, como por exemplo, a classificação de pacientes em grupos de baixo, médio e alto risco. No entanto, diferentes medidas aplicadas as mesmas características podem levar a resultados diferentes (Euan et al. (2019)). Li and Tong (2020) notam que o problema de classificação é comum a vários campos científicos e por este motivo existem diversas notações com o mesmo significado. Por exemplo, "instâncias" são frequentemente referidas como "indivíduos" nas ciências biomédicas, "objetos" em engenharia, "observações" em estatística e *data points* em ciências de dados. Neste trabalho denotaremos amostra como sendo uma coleção de pontos amostrais, com a característica de serem multivariados e as classes/categorias são referidas aqui como grupos. Para Clarke et al. (2009) os problemas de classificação mais simples separam uma população em dois grupos, rotulados de 1 e 2. Os problemas de classificação binária quase sempre podem ser generalizados para problemas de classificação com múltiplos grupos.

Fica caracterizada assim a necessidade de encontrar uma função de decisão para discriminação

entre dados de diferentes grupos. Além disso, quando estamos trabalhando com alta dimensão e baixo tamanho amostral *(HDLSS)* temos um aumento na complexidade destes dados. Neste caso, a teoria de classificação não é adequada, principalmente devido a singularidades da matriz de variâncias e covariâncias empírica. Outra consequência da alta dimensionalidade é a chamada por Corporation and Bellman (1961) como "maldição da dimensionalidade" (*curse of dimensionality*) para significar o fato de que a dificuldade de estimação aumenta quando a dimensionalidade aumenta, pois o volume do espaço aumenta tão rápido que os dados disponíveis se tornam escassos. Para obter um resultado confiável, a quantidade de dados necessários muitas vezes cresce exponencialmente com a dimensionalidade.

No contexto de classificação podemos destacar a análise de discriminante linear (LDA), considerada uma metodologia clássica de classificação supervisionada.[1] No entanto, quando se tem uma configuração de alta dimensão a LDA não é apropriada por duas razões. Primeiro, a estimativa padrão para a matriz de covariância dentro da classe é singular e, portanto, a regra discriminante usual não pode ser aplicada. Em segundo lugar, quando a dimensão (p) dos dados é grande, o classificador resultante é difícil de interpretar, pois a regra de classificação envolve uma combinação linear de todas as características p. A maioria das propostas para resolver essas questões envolvem modificações no classificador linear da teoria classica , com foco particular na esparsidade. Uma classificação penalizada usando o LDA de Fisher é dada em Witten and Tibshirani (2011) onde os autores estenderam o problema discriminante de Fisher para a configuração de alta dimensão impondo penalidades aos vetores discriminantes. A função de penalidade é escolhida com base no problema em questão, e pode resultar em um classificador interpretável.

Para contornar as problemáticas acima Bickel and Levina (2004) discute a Regra de Independência (IR), ou regra ingênua (*naive*) de Bayes, usando apenas a diagonal da matriz de covariância empírica e a compara com a função discriminante linear de Fisher (LDF) mostrando que, na classificação de duas populações normais, essa regra de independência supera muito a regra de discriminação linear de Fisher quando o número de variáveis é grande.

Na literatura atual, métodos de classificação ou classificadores são comumente usados como ferramentas que auxiliam à tomada de decisões binárias. Essas decisões aparecem em toda parte: desde detecção de spam até identificação de biomarcadores em pesquisa médica. Por exemplo, a atual pandemia de COVID-19, os médicos precisam tomar uma decisão binária crítica: se um paciente infectado precisa de hospitalização ou não. Li and Tong (2020) destaca que existem duas estrategias poderosas que auxiliam em decisões binarias: testes de hipóteses e classificações binárias através de *machine learning*. No entanto, testes de hipóteses e classificações binárias estão enraizados em duas diferentes culturas: inferência e predição (Breiman (2001)). Em resumo, uma abordagem inferencial visa inferir uma verdade desconhecida da população a partir dos dados observados e o teste de hipótese é o meio de se obter uma resposta do tipo sim/não. Por exemplo, decidir se uma vacina é efetiva para o COVID-19 é uma questão inferencial na qual a resposta é não observável. Em contrapartida, a predição tem como objetivo prever uma propriedade não observada de uma amostra de um paciente com base nas informações disponíveis. Por exemplo, prever se o paciente irá ou não infartar baseado em informações como idade, peso, comorbidades, tabagismo, etc. Outras características dos métodos de classificação binária são a necessidade de se ter uma amostra grande e a dificuldade de se obter

[1] O aprendizado supervisionado, segundo Delua (2021), é uma abordagem de *machine learning* definida pelo uso de conjuntos de dados rotulados. Já o aprendizado não supervisionado segundo a autora, usa algoritmos de *machine learning* para analisar e agrupar conjuntos de dados sem rótulo.

resultados teóricos para estudar as suas propriedades, que geralmente limitam-se à probabilidade do erro de classificação (*misclassification error*), precisão, especificidade, entre outras. Para isso, normalmente é necessário um particionamento dos dados em conjuntos de treino, teste e validação para obtenção de estimativas dessas medidas.

Afinal, são necessárias centenas de classificadores para resolver problemas de classificação no mundo real? Essa pergunta foi feita no trabalho de Fernández-Delgado et al. (2014), no qual são avaliados 179 classificadores provenientes de 17 famílias (análise de discriminante, redes neurais, *support vector machines (SVM)*, árvores de decisão, *rule-based classifiers*, *boosting*, *bagging*, *stacking*, *random forest*, modelos lineares generalizados, vizinhos mais próximos, mínimos quadrados parciais e regressão de componentes principais, regressão logística e multinomial, *splines* entre outros métodos), com destaque para SVM e *random forest*. He et al. (2021) destaca que embora muitos classificadores tenham sido propostos com base em diferentes princípios, a maioria deles aborda o problema de classificação como um problema de otimização e não como um problema do ponto de vista inferencial. Liao and Akritas (2007) introduz um método de classificação baseado em testes de hipóteses chamado *Test-based classification* (TBC), o qual baseia-se na aplicação de um teste t para comparação de médias. O procedimento consiste em alocar a amostra a ser classificada em um dos grupos, aplicar o teste t e obter o p-valor (p_1). Em seguida, de forma análoga, alocar a amostra no outro grupo, aplicar o teste t e obter o p-valor (p_2). O menor p-valor é utilizado para determinar a classificação. No entanto este procedimento tem algumas deficiências entre elas podemos destacar a já citada questão da alta dimensionalidade, e a outra ocorre quando os dois grupos estão bem separados, podendo, neste caso, resultar em p-valores iguais a zero. Ghimire and Wang (2012) aperfeiçoa o método TBC e aplica no contexto da classificação de *pixels* de imagens. Modarres (2014) e Modarres (2018) desenvolvem um novo classificador baseado no TBC para dados discretos de alta dimensionalidade.

Considerando esse contexto e suas características, propomos um método para classificação com um teste de hipóteses para verificar a significância estatística da classificação. Como base para a nossa metodologia serão utilizados os trabalhos de Cybis et al. (2018) e Valk and Cybis (2020). A estatística do teste aqui proposto é derivada da conhecida estatística B_n, a qual pertence a uma classe de U-estatísticas de onde decorrem as propriedades, como média, variância e normalidade assintótica. Destacamos a versatilidade da nossa proposta, que além de não exigir normalidade dos dados, nem homogeneidade de variância, pode ser utilizada com matriz de dissimilaridade em vez dos dados, podendo assim utilizar diversas medidas de dissimilaridade, o que torna possível a identificação de mudanças não somente na média, mas também em outras características dos dados. Esse é um dos principais diferenciais quando comparamos com o método do Ahmad and Pavlenko (2018), que basicamente está focado em identificar diferenças nas médias.

Nosso trabalho será dividido da seguinte forma. No Capítulo 2 apresentamos um resumo da teoria de U-estatísticas, a qual começa com a definição de uma U-estatística, em seguida, discutimos algumas definições e propriedades importantes, tais como, variância, covariância e esperança de uma U-estatística. Também é observado que uma U-estatística é o estimador de menor variância dentre todos os estimadores não viesados de θ. A Seção 2.2 é dedicada à apresentação do Teorema da Decomposição de Hoeffding, logo em seguida é aplicada a Decomposição de Hoeffding para se obter a normalidade assintótica de U-estatísticas. O Capítulo 3 trata de classificadores baseados em U-estatísticas. Na Seção 3.1 apresentamos o classificador propostos por Ahmad and Pavlenko (2018) onde será denotado por AH. Nas Seções 3.2 e 3.3 começamos definindo e discutindo algumas propri-

edades das estatísticas U_n e B_n apresentadas por Cybis et al. (2018). Já na Seção 3.5 desenvolvemos nosso método de classificação denotado neste trabalho por UC baseado na estatística B_n. Também foram obtidas algumas propriedades do método a partir da Decomposição de Hoeffding. Para obter uma estimativa da variância da estatística do método (DB_n) foi utilizado um procedimento de reamostragem e encerramos o Capítulo com a regra de classificação.

No Capítulo 4 apresentamos a nossa interpretação para o problema de verificar a significância da classificação de uma amostra em um de dois grupos através de um teste de hipóteses. Propomos um teste de hipótese em duas etapas onde é usada a DB_n para tal tarefa. Na Seção 4.2 foi realizado um estudo com dados simulados a partir de distribuições p-dimensionais e foi analisado o comportamento do método de classificação. As propriedades empíricas da estatística de teste DB_n foram estudadas, e para encerrar o Capítulo abordamos brevemente a questão de múltiplos grupos. Uma breve revisão da estimação da probabilidade do erro de classificação ($misssclassification\ error$) é apresentada na Seção 4.3, onde também é apresentada uma importante medida de desempenho para comparar os métodos de classificação, chamada de taxa de erro aparente (APER) Ahmad and Pavlenko (2018).

No Capítulo 5 usamos os resultados de simulação em cenários diversos para avaliar e comparar o desempenho dos classificadores AH e UC, focando principalmente no controle do erro de classificação sob uma estrutura de dados correlacionados, de alta dimensão e também em dados de séries temporais. Com o intuito de avaliar e comparar os métodos UC e AH, no Capítulo 6, aplicamos os métodos em um conjunto de dados reais. Por fim, concluímos nosso trabalho apresentando no Capítulo 7 uma discussão a respeito dos resultados desta dissertação.

Capítulo 2

Uma breve introdução a U-Estatísticas

Este Capítulo é dedicado à apresentação dos principais resultados da teoria clássica de U-estatística, que vem sendo desenvolvida há mais de 50 anos a partir do trabalho pioneiro de Hoeffding (1948). A principal referência utilizada foi Lee (1990), embora esse assunto possa ser encontrado em Denker (1985), Fraser (1956) Capítulo 6, Lehmann (1999) Capítulo 6, entre outros. Vários estimadores não paramétricos conhecidos na literatura pertencem à classe das U-estatísticas, por exemplo, momentos de uma distribuição, média e variância, estatística do teste não paramétrico de Wilcoxon, $Testing$ $Symmetry$ e Medidas de Associação. Segundo Cox and Hinkley (1974), uma U-estatística é uma classe de estatísticas que é notadamente importante na teoria de estimação e especialmente útil para a construção de Estimadores Não Viesados de Variância Uniformemente Mínima (ENVVUM). A letra "U" vem de $unbiased$, traduzindo-se por não viesado.

Através das U-estatísticas é possível derivar estimadores ENVVUM para parâmetros estimáveis (alternativamente, funcional estatístico) para uma grande variedade de distribuições de probabilidade. A teoria da U-estatística aplica-se a classes gerais de distribuições de probabilidade e muitas estatísticas, derivadas originalmente para famílias paramétricas particulares, foram reconhecidas como U-estatísticas para distribuições gerais (Sen (1992)).

2.1 Definições e propriedades

Considere um subconjunto qualquer $\mathcal{F}$ do conjunto de funções de distribuições sobre $\mathbb{R}$ e seja $\theta = \theta(F), F \in \mathcal{F}$, um funcional definido sobre $\mathcal{F}$. Seja $F \in \mathcal{F}$ um membro desconhecido de $\mathcal{F}$ e considere uma amostra $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ vetores independentes e identicamente distribuídas (iid) com função de distribuição F. Suponha que deseja-se estimar o parâmetro $\theta = \theta(F)$ baseado na amostra.

Os problemas tratados nesse contexto são não paramétricos, o que significa que $\mathcal{F}$ será considerado um grande família de distribuições sujeitas apenas a restrições leves, como continuidade ou existência de momentos. Em geral, assume-se que $\mathcal{F}$ consiste de todos as distribuições absolutamente contínuas com suporte em $\mathbb{R}$ ou discretas (ou subclasses dessas). Nota-se que nenhuma hipótese referente ao conhecimento de F é assumida, a menos que ela é uma distribuição da família $\mathcal{F}$. A primeira noção que temos necessidade é a de um parâmetro estimável.

Definição 2.1.1. Um parâmetro θ é estimável de grau k para uma família de distribuições $\mathcal{F}$ se k é o menor tamanho da amostra para o qual existe uma função $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ tal que

$$\mathbb{E}_F[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)] = \theta, \forall F \in \mathcal{F}, \tag{2.1}$$

onde $\mathbf{X}_1, \cdots, \mathbf{X}_k$ são v.a's independentes e com distribuição comum F.

Assim, pode-se dizer que um parâmetro é estimável se existe um estimador não viesado para esse parâmetro. A função $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ que satisfaz (2.1) é chamada núcleo de θ.

Exemplo 2.1. Quando estamos trabalhando com $k = 2$ é comum a utilização da distância euclidiana como núcleo da U-estatística. Nesse caso,

$$\phi(\mathbf{X}_1, \mathbf{X}_2) = \sqrt{\sum_{i=1}^{p}(X_{1i} - X_{2i})^2}.$$

♦

Observação 2.1. Se θ_1 e θ_2 são parâmetros estimáveis de graus k_1 e k_2, respectivamente, então $\theta_1 + \theta_2$ é estimável de grau $k = \max\{k_1, k_2\}$ e $\theta_1 \cdot \theta_2$ é estimável de grau $k_1 + k_2$. Assim, um parâmetro estimável é um funcional linear. Costuma-se denominar um parâmetro estimável por funcional regular.

Sem perda de generalidade, pode-se supor que $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ é simétrica em seus argumentos, isto é, $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k) = \phi(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_k})$ para qualquer permutação $\{i_1, \ldots, i_k\}$ do conjunto $\{1, \cdots, k\}$. Se $\phi(\mathbf{X}_1, \ldots, \mathbf{X}_k)$ não for simétrica, então se pode construir a partir de $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ uma função simétrica nos argumentos:

$$\phi^*(\mathbf{X}_1, \cdots, \mathbf{X}_k) = \frac{1}{k!}\sum_{P_k}\phi(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_k}), \tag{2.2}$$

onde P_k indica a soma sobre todas as permutações no conjunto $\{1, \ldots, k\}$.

Assumindo-se que a família $\mathcal{F}$ contém as funções de distribuições de suporte finito ou contendo as funções de distribuição absolutamente contínuas é possível mostrar que existe um único estimador de θ que é não viesado e simétrico (Lee (1990)). Este estimador chamado U-estatística é definido a seguir.

Definição 2.1.2. Sejam $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ v.a's iid com função de distribuição F e seja $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ um estimador não viesado de um parâmetro estimável θ de grau $k \leq n$. Define-se a U-estatística para a amostra como sendo

$$U_n(\phi) = U(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \binom{n}{k}^{-1}\sum_{\mathcal{I}\in C_{n,k}}\phi^*(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_k}) \tag{2.3}$$

onde $\mathcal{I} = \{i_1, \cdots, i_k\}$, e o somatório é sobre todas as combinações de k inteiros escolhidos sem reposição do conjunto de inteiros de 1 a n ($C_{n,k}$) e ϕ^* é o núcleo simétrico correspondente a ϕ definido em (2.2).

Observação 2.2. A condição não viesado sobre $\phi^*(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_k})$ garante que a U-estatística seja também não viesada, pois aplicando a esperança em ambos os lados de (2.3) e usando o fato das variáveis aleatórias serem iid têm-se que $\mathbb{E}[U(\mathbf{X}_1, \cdots, \mathbf{X}_n)] = \mathbb{E}[\phi^*(\mathbf{X}_1, \cdots, \mathbf{X}_k)] = \theta$. Pelo mesmo argumento, se $\phi^*(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ for um estimador assintoticamente não viesado, então o estimador U-estatística criado a partir dele também é assintoticamente não viesado.

Exemplo 2.2. *Momentos.* Se $\mathcal{F}$ é a família de todas as distribuições na reta real com média finita, então a média $\mu = \mu(F) = \int_{\mathbb{R}} x dF(x)$, é um parâmetro estimável de grau $k = 1$, pois $\phi(X_1) = X_1$ é um estimador não viesado de μ. A correspondente U-estatística é a média amostral, $U_n = \overline{X}_n = (1/n)\sum_1^n X_i$. Similarmente, se $\mathcal{F}$ é a família de distribuições na reta real com k-ésimo momento finito, $\mu_k = \int_{\mathbb{R}} x^k dF(x)$ é um parâmetro estimável de grau 1 com U-estatística, $(1/n)\sum_1^n X_i^k$. ♦

Exemplo 2.3. Como encontrar um estimador para a o quadrado da média, $\theta(F) = \mu^2$? Como $\mathbb{E}(X_1 X_2) = \mu^2$, este também é um parâmetro estimável de grau 2. A U-estatística U_n de (2.3) correspondente a $\phi(x_1, x_2) = x_1 x_2$ é

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j = \frac{2}{n(n-1)} \sum_{i<j} X_i X_j.$$

♦

Exemplo 2.4. Se $\mathcal{F}$ é a família de distribuições na reta real com segundo momento, então a variância, $\sigma^2 = \mu_2 - \mu^2$, é um parâmetro estimável de grau 2, pois pode-se estimar μ_2 por $\mathbf{X}_1^2$ e μ^2 por $\mathbf{X}_1 X_2$:

$$\mathbb{E}\left(X_1^2 - X_1 X_2\right) = \sigma^2.$$

Embora, $\phi(x_1, x_2) = x_1^2 - x_1 x_2$, não seja simétrico em x_1 e x_2, é possível transformá-lo em um núcleo simétrico da forma

$$\phi^*(x_1, x_2) = \frac{1}{2}\left(\phi(x_1, x_2) + \phi(x_2, x_1)\right) = \frac{x_1^2 - 2x_1 x_2 + x_2^2}{2} = \frac{(x_1 - x_2)^2}{2},$$

de onde segue a U-estatística

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \sum_1^n \left(X_i - \bar{X}\right)^2,$$

que é um conhecido estimador não viesado da variância. ♦

Uma das principais vantagens da U-estatística é que ela é o estimador de menor variância dentre todos os estimadores não viesados de θ. Isso pode ser mostrado observando que a U-estatística U_n é uma função da estatística de ordem $\left(\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(n)} \right)$ que é suficiente e completa. Daí segue do Teorema 8.4.6 em Rohatgi and Md (2001) que U_n é o estimador não viesado de variância mínima do seu valor esperado que é θ. Uma prova simples pode ser obtida se assumir que $\mathcal{F}$ é a família de funções de distribuição de suporte finito ou a família de funções de distribuição absolutamente contínuas, e usando a unicidade do estimador não viesado e simétrico. Esta última prova é apresentada a seguir.

Teorema 2.1.3. Seja θ um parâmetro estimável de grau k com núcleo ϕ, definido sobre o conjunto $\mathcal{F}$ das funções de distribuição absolutamente contínuas ou das funções de distribuição de suporte finito. Então a U-estatística $U_n(\phi)$ é o estimador de variância mínima na classe de todos os estimadores não viesados de θ, baseado numa amostra de tamanho $n \geq k$.

Prova: Seja $\phi = \phi(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ um estimador não viesado de θ baseado numa amostra de tamanho n, $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, de $F \in \mathcal{F}$. Seja ϕ^* o estimador simétrico correspondente a ϕ, ou seja,

$$\phi^*(\mathbf{X}_1, \cdots, \mathbf{X}_n) = \frac{1}{n!} \sum_{P_n} \phi(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_n}),$$

onde a soma é considerada sobre todas as permutações P_n de $\{1, \ldots, n\}$. Então ϕ^* é um estimador não viesado e simétrico de θ. Logo, como foi observado anteriormente, ϕ^* coincide com U_n sobre $\mathbb{R}$. Assim, aplicando a desigualdade de Cauchy Schwartz, pode-se escrever

$$\begin{aligned} U_n^2 &= \left(\sum_{P_n} \frac{1}{n!} \phi(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_n}) \right)^2 \\ &\leq \sum_{P_n} \left(\frac{1}{n!} \right)^2 \sum_{P_n} \phi^2(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_n}) \\ &= \frac{1}{n!} \sum_{P_n} \phi^2(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_n}). \end{aligned}$$

Logo,

$$\begin{aligned} \mathbb{E}[U_n] &\leq \frac{1}{n!} \sum_{P_n} \mathbb{E}[\phi^2(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_n})] \\ &= \frac{1}{n!} \sum_{P_n} \mathbb{E}[\phi^2(\mathbf{X}_1, \cdots, \mathbf{X}_n)] \\ &= \mathbb{E}[\phi^2(\mathbf{X}_1, \cdots, \mathbf{X}_n)]. \end{aligned}$$

Como $\mathbb{E}U_n = \mathbb{E}\phi$, segue que $\operatorname{Var} U_n \leq \operatorname{Var} \phi$, o que conclui a demonstração do Teorema 2.1.3. □

A seguir serão apresentadas algumas propriedades importantes sobre a variância de uma U-estatística U_n.

Teorema 2.1.4. Seja $U_n(\phi)$ uma U-estatística com núcleo $\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right)$ de grau k. Então

$$\operatorname{Var}[U_n(\phi)] = \binom{n}{K}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\sigma_c^2, \tag{2.4}$$

onde

$$\sigma_c^2 = \operatorname{Var}\phi_c\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right) \tag{2.5}$$

e

$$\phi_c\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right) = \mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \mid \ \mathbf{X}_1 = \mathbf{X}_1, \ldots, \mathbf{X}_c = \mathbf{X}_c\right], \ c = 1, \ldots, k. \tag{2.6}$$

Para demostrar o teorema acima necessita-se de alguns resultados preliminares. O primeiro deles é uma consequência imediata das propriedades de esperança condicional.

Lema 2.1.5. Seja

$$\phi_c\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right) = \mathbb{E}\left[\phi\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_k\right) \mid \mathbf{X}_1 = \mathbf{X}_1, \ldots, \mathbf{X}_c = \mathbf{X}_c\right], \ c = 1, \ldots, k.$$

então

i)

$$\phi_c\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right) = \mathbb{E}\left[\phi_d\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_d\right) \mid \ \mathbf{X}_1 = \mathbf{X}_1, \cdots, \mathbf{X}_c = \mathbf{X}_c\right] \text{ para } 1 \leq c < d \leq k.$$

ii)

$$\mathbb{E}\left[\phi_c\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right] = \mathbb{E}\left[\phi\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right].$$

No lema a seguir será obtido uma expressão para a variância e a covariância de $\phi_c\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)$.

Lema 2.1.6. Seja $\sigma_c^2 = \operatorname{Var}\left[\phi_c\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right]$, onde $\phi_c\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)$ foi definido por (2.6). Então,

$$\zeta_c = \operatorname{Cov}\left[\phi_c\left(\ \mathbf{X}_{\alpha_{1c}}, \cdots, \mathbf{X}_{\alpha_{kc}}\right), \phi_c\left(\ \mathbf{X}_{\beta_{1c}}, \cdots, \mathbf{X}_{\beta_{kc}}\right)\right] = \sigma_c^2, \tag{2.7}$$

onde $\{\alpha_{1c}, \ldots, \alpha_{kc}\}$ e $\{\beta_{1c}, \ldots, \beta_{kc}\}$ são subconjuntos de k elementos escolhidos de $\{1, \ldots, n\}$ com c elementos em comum.

Prova: Por definição, $\sigma_c^2 = \mathbb{E}\left[\phi_c^2\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right] - \left\{\mathbb{E}\left[\phi_c\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right]\right\}^2$. Pelo Lema 2.1.5, item (ii), temos que para c arbitrário,

$$\mathbb{E}\left[\phi_c\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right] = \mathbb{E}\left[\phi\left(\ \mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right].$$

Assim,

$$\sigma_c^2 = \mathbb{E}\left[\phi_c^2\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right] - \left\{\mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right]\right\}^2.$$

Por outro lado, como $\mathbf{X}_1, \ldots, \mathbf{X}_n$ são independentes e identicamente distribuídas e $\{\alpha_{1c}, \ldots, \alpha_{kc}\}$ e $\{\beta_{1c}, \ldots, \beta_{kc}\}$ têm c elementos em comum, pode-se escrever

$$\begin{aligned}
&\operatorname{Cov}\left[\phi\left(\mathbf{X}_{\alpha_{1c}}, \cdots, \mathbf{X}_{\alpha_{kc}}\right), \phi\left(\mathbf{X}_{\beta_{1c}}, \cdots, \mathbf{X}_{\beta_{kc}}\right)\right] \\
&= \mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right)\right] \\
&- \mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right] \mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right)\right] \\
&= \mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right)\right] \\
&- \left\{\mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right]\right\}^2.
\end{aligned}$$

Comparando as expressões para

$$\operatorname{Cov}\left[\phi\left(\mathbf{X}_{\alpha_{1c}}, \cdots, \mathbf{X}_{\alpha_{kc}}\right), \phi\left(\mathbf{X}_{\beta_{1c}}, \cdots, \mathbf{X}_{\beta_{kc}}\right)\right]$$

e σ_c^2, é suficiente mostrar que

$$\mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right)\right] = \mathbb{E}\left[\phi_c^2\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right].$$

Mas, como $\mathbf{X}_1, \ldots, \mathbf{X}_n$ são i.i.d. temos

$$\begin{aligned}
&\mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right)\right] \\
&= \int \cdots \int \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right) \prod_{i=1}^{2k-c} dF(x_i) \\
&= \int \cdots \int \left[\int \cdots \int \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right) \prod_{i=c+1}^{k} dF(x_i)\right] \\
&\cdot \left[\int \cdots \int \phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_c, \mathbf{X}_{k+1}, \cdots, \mathbf{X}_{2k-c}\right) \prod_{i=k+1}^{2k-c} dF(x_i)\right] \prod_{i=1}^{c} dF(x_i) \\
&\int \cdots \int \phi_c^2\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right) \prod_{i=1}^{c} dF(x_i) \\
&= \mathbb{E}\left[\phi_c^2\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right]
\end{aligned}$$

e o resultado segue. □

Como os resultados dos Lemas 2.1.5 e 2.1.6 temos as ferramentas necessárias para demonstrar o Teorema 2.1.4.

Prova: (Do Teorema 2.1.4). Segue que

$$\begin{aligned}\operatorname{Var} U_n(\phi) &= \operatorname{Var}\left[\binom{n}{k}^{-1}\sum_{C_{n,k}}\phi\left(\mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k}\right)\right]\\ &= \operatorname{Cov}\left[\binom{n}{k}^{-1}\sum_{C_{n,k}}\phi\left(\mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k}\right),\binom{n}{k}^{-1}\sum_{C_{n,k}}\phi\left(\mathbf{X}_{\beta_1},\cdots,\mathbf{X}_{\beta_k}\right)\right]\\ &= \binom{n}{k}^{-2}\sum_{C_{n,k}}\sum_{C_{n,k}}\operatorname{Cov}\left[\phi\left(\mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k}\right),\phi\left(\mathbf{X}_{\beta_1},\cdots,\mathbf{X}_{\beta_k}\right)\right]\end{aligned}$$

A covariância é zero entre os conjuntos $\{\phi(\alpha_{1_0},\cdots,\alpha_{k_0})\}$ e $\{\phi(\beta_{1_0},\cdots,\beta_{k_0})\}$, pois não há elementos em comum e então há independência entre eles. Assim, precisa-se considerar a covariância entre os pares $\{\phi(\alpha_{1_c},\cdots,\alpha_{k_c})\}$ e $\{\phi(\beta_{1_c},\cdots,\beta_{k_c})\}$ para $c\neq 0$, $c=1,\ldots,k$. Com isso,

$$\operatorname{Var} U_n(\phi) = \binom{n}{k}^{-2}\sum_{c=1}^{k} m\operatorname{Cov}\left[\phi\left(\mathbf{X}_{\alpha_{1c}},\cdots,\mathbf{X}_{\alpha_{kc}}\right),\phi\left(\mathbf{X}_{\beta_{1c}},\cdots,\mathbf{X}_{\beta_{kc}}\right)\right].$$

Precisamos determinar o m. O primeiro membro do par $\phi\left(\mathbf{X}_{\alpha_{1c}},\cdots,\mathbf{X}_{\alpha_{kc}}\right)$ e $\phi\left(\mathbf{X}_{\beta_{1c}},\cdots,\mathbf{X}_{\beta_{kc}}\right)$ pode ser escolhido de $\binom{n}{k}$ maneiras. O segundo membro tem que ter c elementos em comum com o primeiro membro, assim existem $\binom{k}{c}$ maneiras de escolher c. Como não se quer ter mais do que c elementos em comum com o primeiro membro, os $k-c$ elementos restantes devem ser escolhidos de $\binom{n-k}{k-c}$ maneiras. Assim $m=\binom{n}{k}\cdot\binom{k}{c}\cdot\binom{n-k}{k-c}$. Então,

$$\begin{aligned}\operatorname{Var} U_n(\phi) &= \binom{n}{k}^{-2}\sum_{c=1}^{k}\binom{n}{k}\cdot\binom{n}{c}\cdot\binom{n-k}{k-c}\zeta_c\\ &= \binom{n}{k}^{-1}\sum_{c=1}^{k}\binom{k}{c}\cdot\binom{n-k}{k-c}\zeta_c.\end{aligned}$$

Agora, pelo Lema 2.1.6 temos que $\zeta_c=\sigma_c^2$ e a equação (2.4) está provado. □

Como consequência do Teorema 2.1.4 se obtém um resultado sobre o comportamento assintótico da variância de uma U-estatística, quando o tamanho da amostra cresce, e a partir daí, obtém-se a consistência em média quadrática do estimador.

Corolário 2.1.7. Seja $\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n$ uma amostra aleatória com distribuição $F\in\mathcal{F}$ e θ um parâmetro estimável de grau k. Para $n\geq k$ seja $U_n(\phi)=U\left(\mathbf{X}_1,\cdots,\mathbf{X}_n\right)$ uma U-estatística com núcleo simétrico ϕ. Se $\mathbb{E}\left[\phi^2\left(\mathbf{X}_1,\cdots,\mathbf{X}_k\right)\right]<\infty$, então

$$\lim_{n\to\infty} n\operatorname{Var} U\left(\mathbf{X}_1,\cdots,\mathbf{X}_n\right)=m^2\zeta_1 \tag{2.8}$$

onde ζ_1 é a covariância dada em (2.7), e consequentemente

$$U\left(\mathbf{X}_1,\cdots,\mathbf{X}_n\right)\xrightarrow{(2)}\theta, \tag{2.9}$$

onde $\xrightarrow{(2)}$ indica convergência em média quadrática.

Prova: Para provar (2.8), tem-se do Teorema 2.1.4 que

$$n \operatorname{Var} U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right) = \frac{\sum_{c=1}^{k} n\binom{k}{c}\binom{n-k}{k-c}\zeta_c}{\binom{n}{k}}.$$

Desenvolvendo o termo geral do somatório para cada $c = 1, \ldots, k$ se obtém

$$\begin{aligned}
\frac{n\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}}\zeta_c &= \frac{nk!(n-m)!k!(n-k)!}{n!c!(k-c)!(n-2k+c)!}\zeta_c \\
&= K_c\frac{(n-m)!(n-k)!}{(n-2k+c)!n!}\zeta_c \\
&= K_c\frac{(n-m)(n-k+1)\cdots(n-2k+c+1)}{(n-1)\cdots(n-k+1)}\zeta_c
\end{aligned} \tag{2.10}$$

onde $K_c = \frac{[k!]^2}{c![(k-c)]^2}$. Quando $c = 1$, tem-se em (2.10) o mesmo número de termos envolvendo n no numerador e no denominador. Assim,

$$\frac{(n-m)(n-k+1)\cdots(n-2k+c+1)}{(n-1)\cdots(n-k+1)} \to 1, \text{ quando } n \to \infty.$$

Quando $c > 1$, tem-se mais termos envolvendo n no denominador do que no numerador de (2.10). Segue que

$$\frac{(n-m)(n-k+1)\cdots(n-2k+c+1)}{(n-1)\cdots(n-k+1)} \to 0, \text{ quando } n \to \infty.$$

Consequentemente, no somatório sobra um único termo referente a $c = 1$, e portanto,

$$n \operatorname{Var} U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right) \xrightarrow{p} K_1\zeta_1 = k^2\zeta_1, \text{ quando } n \to \infty.$$

Resta provar (2.9). Para isso, deve-se escrever

$$\mathbb{E}\left[U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right) - \theta\right]^2 = VarU\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right),$$

pois

$$\mathbb{E}\left[U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right)\right] = \mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right].$$

Assim, como $\mathbb{E}\left[\phi\left(\mathbf{X}_1, \cdots, \mathbf{X}_k\right)\right] = \theta$ apenas resta provar que:

$$\operatorname{var} U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right) \to 0 \text{ quando } n \to \infty.$$

Mas, por (2.8) tem-se que o limite $n \operatorname{Var} U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right)$ existe e é finito. Logo,

$$\lim_{n\to\infty} \operatorname{Var} U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right) = \lim_{n\to\infty} \frac{n \operatorname{Var} U\left(\mathbf{X}_1, \cdots, \mathbf{X}_n\right)}{n} = 0.$$

□

2.2 Decomposição de Hoeffding

Nessa seção será apresentada uma representação obtida por Hoeffding (1961), de uma U-estatística de grau k como soma de U-estatísticas não correlacionadas de grau $1, 2, \ldots, k$. Essa decomposição é extremamente útil para obtenção de propriedades de estimadores que podem ser escritos como U-estatísticas. Para obter esta decomposição, define-se recursivamente os núcleos $\phi^{(1)}, \ldots, \phi^{(k)}$ correspondentes ao núcleo simétrico ϕ, da seguinte forma:

$$\phi^{(1)} = \phi^1(\ \mathbf{X}_1) = \phi_1(\ \mathbf{X}_1) - \mathbb{E}\left[\phi(\ \mathbf{X}_1, \cdots, \mathbf{X}_k)\right] \tag{2.11}$$

e para $c = 2, \ldots, k$,

$$\phi^{(c)} = \phi^c(\ \mathbf{X}_1, \cdots,\ \mathbf{X}_c) = \phi_c(\ \mathbf{X}_1, \cdots,\ \mathbf{X}_c) - \sum_{j=1}^{c-1} \sum_{C_{c,j}} \phi^j(\ \mathbf{X}_{\alpha_1}, \cdots,\ \mathbf{X}_{\alpha_j}) - \mathbb{E}\left[\phi(\ \mathbf{X}_1, \cdots, \mathbf{X}_k)\right]. \tag{2.12}$$

Observa-se que os núcleos ϕ^c definidos nas equações (2.11) e (2.12) são simétricos em virtude de $\phi(\ \mathbf{X}_1, \cdots,\ \mathbf{X}_k)$ ser simétrico. É fundamental que $\phi^{(c)}$ seja simétrico, pois será visto que $\phi^{(c)}$ vai ser o núcleo de uma U-estatística.

Teorema 2.2.1. (Teorema da Decomposição de Hoeffding) Seja $U(\ \mathbf{X}_1, \cdots, \mathbf{X}_n)$ a U-estatística correspondente a um núcleo simétrico $\phi(\ \mathbf{X}_1, \cdots, \mathbf{X}_k)$ de grau k. Para $c = 1, \ldots, k$, seja H_n^j a U-estatística de grau $j, j = 1, \ldots, k$, baseada no núcleo $\phi^{(j)}$, definido em (2.11) e em (2.12), ou seja,

$$H_n^j = \binom{n}{j} \sum_{C_{n,j}} \phi^j(\ \mathbf{X}_{\alpha_1}, \cdots,\ \mathbf{X}_{\alpha_j}). \tag{2.13}$$

Então,

$$U\left(\ \mathbf{X}_1, \cdots,\ \mathbf{X}_n\right) = \mathbb{E}\left[\phi(\ \mathbf{X}_1, \cdots, \mathbf{X}_k)\right] + \sum_{j=1}^{k} \binom{k}{j} H_n^j. \tag{2.14}$$

A prova de (2.14) será realizada em vários passos, que serão apresentados na forma de lemas. Primeiramente defina-se a soma para $j < c < k$,

$$S_j(i_1, \cdots, i_k) = \sum_{C_{k,j}} \phi^j(\ \mathbf{X}_{\alpha_1}, \cdots,\ \mathbf{X}_{\alpha_j}). \tag{2.15}$$

Lema 2.2.2. A partir da definição em (2.15) segue que

$$\sum_{C_{n,k}} S_j(i_1, \cdots, i_k) = \binom{n-j}{k-j} \sum_{C_{n,j}} \phi^j(\ \mathbf{X}_{\alpha_1}, \cdots,\ \mathbf{X}_{\alpha_k}).$$

Prova: Por (2.15) temos

$$\sum_{C_{n,k}} S_j(i_1, \cdots, i_k) = \sum_{C_{n,k}} \sum_{C_{k,j}} \phi^j(\ \mathbf{X}_{\alpha_1}, \cdots,\ \mathbf{X}_{\alpha_j}). \tag{2.16}$$

Em (2.16), para cada conjunto de k elementos formados a partir do conjunto $\{1,\cdots,n\}$, considerando todos os conjuntos de j elementos que é possível formar a partir deles. Assim, tem-se em (2.16) todos os conjuntos de j elementos que é possível obter a partir do conjunto $\{1,\cdots,n\}$, cada um deles aparece um número m de vezes em (2.16). Por exemplo, considerando-se os conjuntos $\{1,\cdots,k\}$ e $\{1,\cdots,k-1,k+1\}$ formados a partir de $\{1,\cdots,n\}$, pode-se retirar de $\{1,\cdots,k\}$ e $\{1,\cdots,k-1,k+1\}$ o mesmo conjunto $\{1,\cdots,j\}$. Dessa discussão segue,

$$\sum_{C_{n,k}} S_j(i_1,\cdots,i_k) = m \sum_{C_{n,j}} \phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_k}).$$

Será determinado m. Nota-se que para ocorrer uma situação similar ao exemplo dado, ou seja, para que conjuntos de k elementos distintos gerem o mesmo conjunto de j elementos, é necessário que eles tenham j elementos em comum, assim o número m de vezes que cada conjunto de j elementos formados a partir do conjunto $\{1,\cdots,n\}$ aparece em (2.16) é $\binom{n-j}{k-j}$. □

O lema a seguir relaciona as somas $S_j(\cdot)$ com o kernel $\phi(\cdot)$.

Lema 2.2.3. Considere a somas definidas em (2.15). Então

$$\phi(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k,) = \sum_{j=1}^{k} S_j\left(i_1,\cdots,i_k,\right) + \mathbb{E}\left[\phi(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right].$$

Prova: Na equação (2.12) temos

$$\phi^k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) = \phi_k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) - \sum_{j=1}^{k-1}\sum_{C_{k,j}} \phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_j}) - \mathbb{E}\left[\phi(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]. \quad (2.17)$$

Pela definição de $S_j(i_1,\cdots,i_k)$ em (2.15), pode-se reescrever (2.17),

$$\phi^k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) = \phi_k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) - \sum_{j=1}^{k-1} S_j\left(i_1,\cdots,i_k,\right) - \mathbb{E}\left[\phi\left(\mathbf{X}_1,\cdots,\mathbf{X}_k\right)\right] \quad (2.18)$$

e isolando $\sum_{j=1}^{k-1} S_j\left(i_1,\cdots,i_k,\right)$ em (2.18), obtém-se

$$\sum_{j=1}^{k-1} S_j\left(i_1,\cdots,i_k,\right) = \phi_k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) - \phi^k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) - \mathbb{E}\left[\phi\left(\mathbf{X}_1,\cdots,\mathbf{X}_k\right)\right]. \quad (2.19)$$

Tem-se também que

$$\sum_{j=1}^{k} S_j\left(i_1,\cdots,i_k,\right) + \mathbb{E}\left[\phi(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right] = \sum_{j=1}^{k-1} S_j\left(i_1,\cdots,i_k,\right) + S_k\left(i_1,\cdots,i_k,\right) + \mathbb{E}\left[\phi(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]. \quad (2.20)$$

Daí, substituindo (2.19) em (2.20) concluí-se que,

$$\sum_{j=1}^{k} S_j\left(i_1,\cdots,i_k,\right) + \mathbb{E}\left[\phi(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right] = \phi_k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) - \phi^k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k) - S_k\left(i_1,\cdots,i_k,\right).$$

Novamente pela definição de $S_j\,(i_1,\cdots,i_k,)$ em (2.15), obtém-se

$$S_k(i_1,\cdots,i_k)=\phi^k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k).$$

Daí

$$\sum_{j=1}^{k} S_j(i_1,\cdots,i_j)+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]=\phi_k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k).$$

Mas,

$$\phi_k(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k)=\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\,|\,\mathbf{X}_1=\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k=\ \mathbf{X}_k\right]=\phi(\ \mathbf{X}_1,\cdots,\ \mathbf{X}_k),$$

e assim o resultado está provado. □

Prova: (Do Teorema 2.2.1). Dos Lemas 2.2.2 e 2.2.3 e de (2.15) segue que

$$\begin{aligned}
U\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_n) &= \binom{n}{k}^{-1}\sum_{C_{n,k}}\phi\,(\ \mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k})\\
&= \binom{n}{k}^{-1}\sum_{C_{n,k}}\{\sum_{j=1}^{k}S_j\,(i_1,\cdots,i_k)+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]\}\\
&= \binom{n}{k}^{-1}\left[\sum_{j=1}^{k}\sum_{C_{n,k}}S_j\,(i_1,\cdots,i_k)\right]+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]\\
&= \binom{n}{k}^{-1}\binom{n-j}{k-j}\sum_{j=1}^{k}\sum_{C_{n,j}}\phi^j\,(\ \mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k})+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]\\
&= \sum_{j=1}^{k}\binom{n}{k}^{-1}\binom{n-j}{k-j}\sum_{C_{n,j}}\phi^j\,(\ \mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k})+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]
\end{aligned}$$

Agora, pode-se verificar que:

$$\binom{n}{k}^{-1}\binom{n-j}{k-j}=\binom{k}{j}\binom{n}{j}^{-1} \tag{2.21}$$

Assim,

$$\begin{aligned}
U\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_n) &= \sum_{j=1}^{k}\binom{k}{j}\binom{n}{j}^{-1}\sum_{C_{n,j}}\phi^j\,(\ \mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k})+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]\\
&= \sum_{j=1}^{k}\binom{k}{j}H_n^j+\mathbb{E}\left[\phi\,(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]
\end{aligned}$$

onde H_n^j é uma U-estatística de grau j baseado no núcleo simétrico $\phi^{(j)}$, o que finaliza a demonstração do Teorema 2.2.1. □

As componentes U-estatística que aparecem na Decomposição de Hoeffding também podem ser escritas como uma U-estatística de grau k e núcleo $\phi_j(i_1,\cdots,i_k)$, pois da definição (2.13) de H_n^j, de (2.21) e do lema 2.2.2 podemos escrever:

$$\begin{aligned}\binom{k}{j}H_n^j &= \binom{k}{j}\binom{n}{j}^{-1}\sum_{C_{n,j}}\phi^j\left(\mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k}\right)\\ &= \binom{n}{k}^{-1}\binom{n-j}{k-j}\sum_{C_{n,j}}\phi^j\left(\mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_k}\right)\\ &= \binom{n}{k}^{-1}\sum_{C_{n,k}}S_j(i_1,\cdots,i_k).\end{aligned} \tag{2.22}$$

Observa-se também que se considerar o termo R_n^c obtido do truncamento da decomposição H depois de c termos, então R_n^c é uma U-estatística de grau k com núcleo $\sum_{j=c+1}^{k} S_j(i_1,\cdots,i_k)$, pois por (2.22) segue

$$\begin{aligned}R_n^c = \sum_{j=c+1}^{k}\binom{k}{j}H_n^j &= \sum_{c+1}^{k}\binom{n}{k}^{-1}\sum_{C_{n,k}}S_j(i_1,\cdots,i_k)\\ &= \binom{n}{k}^{-1}\sum_{C_{n,k}}\sum_{j=c+1}^{k}S_j(i_1,\cdots,i_k).\end{aligned}$$

Falta provar que as componentes U-estatísticas que aparecem na Decomposição de Hoeffding são não correlacionadas, mas será preciso do seguinte lema cuja prova será omitida e pode ser vista em Lee (1990), página 28.

Lema 2.2.4. Sejam $\phi_c(\mathbf{X}_1,\cdots,\mathbf{X}_c)$ e $\phi^c(\mathbf{X}_1,\cdots,\mathbf{X}_c)$ como foram definidas em (2.6), (2.11) e (2.12), respectivamente. Então

(i) Para $c=1,\cdots,j-1$ e $j=1,\cdots,k$ temos $\phi^c(\mathbf{X}_1,\cdots,\mathbf{X}_c)=0$;

(ii) $\mathbb{E}\left[\phi^j\left(\mathbf{X}_1,\cdots,\mathbf{X}_j\right)\right]=0$.

O seguinte resultado trata das propriedades estatísticas dos termos da Decomposição de Hoeffding, como variâncias e covariâncias.

Teorema 2.2.5. Seja H_n^j como no Teorema 2.2.1.

(i) Seja $j<j'$ e sejam $\{\alpha_1,\cdots,\alpha_{j'}\}$ e $\{\beta_1,\cdots,\beta_{j'}\}$ subconjuntos de j e j' elementos respectivamente de $\{1,\cdots,n\}$. Então

$$\operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1},\cdots,\mathbf{X}_{\alpha_j}\right),\phi^{j'}\left(\mathbf{X}_{\beta_1},\cdots,\mathbf{X}_{\beta_{j'}}\right)\right]=0 \tag{2.23}$$

e, assim,

$$\operatorname{Cov}\left(H_n^j,H_n^{j'}\right)=0. \tag{2.24}$$

(ii) Sejam $\{\alpha_1, \cdots, \alpha_j\}$ e $\{\beta_1, \cdots, \beta_j\}$ subconjuntos de j elementos distintos de $\{1, \cdots, n\}$. Então

$$\operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right), \phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_j}\right)\right] = 0 \tag{2.25}$$

(iii) Além disso, tem-se

$$\operatorname{Var} H_n^j = \binom{n}{j}^{-1} \delta_j^2, \tag{2.26}$$

onde $\delta_j^2 = \operatorname{Var} \phi^j\left(\mathbf{X}_1, \cdots, \mathbf{X}_j\right)$.

Prova:

(**i**) Para provar (2.23) temos pelo Lema 2.2.4, item (ii) (2.25), que $\mathbb{E}\left[\phi^j\left(\mathbf{X}_1, \cdots, \mathbf{X}_j\right)\right] = 0$ $\forall j = 1, \cdots, k$. Assim,

$$\operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right), \phi^{j'}\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right)\right] = \mathbb{E}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \phi^{j'}\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right)\right].$$

Como $j < j'$, certamente há elementos de $\left\{\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right\}$ que não estão em $\left\{\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_{j'}}\right\}$. Daí, há pelo menos uma variável aleatória $\mathbf{X}_{j'}$ que aparece em $\phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right)$, mas não aparece em $\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right)$. Portanto, $\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right)$ é independente de $\mathbf{X}_{j'}$. Então, pode-se escrever

$$\begin{aligned}
&\operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right), \phi^{j'}\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right)\right] \\
&= \mathbb{E}\{\mathbb{E}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \phi^{j'}\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right)\right] | \mathbf{X}_{j'}\} \\
&= \mathbb{E}\{\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \mathbb{E}\left[\phi^{j'}\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right) | \mathbf{X}_{j'}\right]\} \\
&= \mathbb{E}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \phi_1^{j'}\left(\mathbf{X}_{j'}\right)\right] \\
&= 0,
\end{aligned}$$

pois pelo Lema 2.2.4, item (i), temos que $\phi_1^{j'}\left(\mathbf{X}_{j'}\right) = 0$. Daí, (2.24) segue imediatamente de (2.23), pois

$$\operatorname{Cov}\left(H_n^j, H_n^{j'}\right) = \binom{n}{j}^{-1}\binom{n}{j'}^{-1} \sum_{C_{n,j}} \sum_{C_{n,j'}} \operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right), \phi^{j'}\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_{j'}}\right)\right] = 0.$$

(**ii**) Se $\{\alpha_1, \cdots, \alpha_j\} \cap \{\beta_1, \cdots, \beta_j\}$ é vazia, então $\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right)$ e $\phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_j}\right)$ são independentes e portanto a $\operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right), \phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_j}\right)\right] = 0$. Senão, sem perda de generalidade supõe-se que $\{1, \cdots, c\}$ sejam os índices das variáveis aleatórias que não estão em $\{\alpha_1, \cdots, \alpha_j\}$. Daí

$$\begin{aligned}
&\operatorname{Cov}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right), \phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_j}\right)\right] = \\
&= \mathbb{E}\{\mathbb{E}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_j}\right)\right] | \mathbf{X}_1, \cdots, \mathbf{X}_c\} \\
&= \mathbb{E}\{\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \mathbb{E}\left[\phi^j\left(\mathbf{X}_{\beta_1}, \cdots, \mathbf{X}_{\beta_j}\right) | \mathbf{X}_1, \cdots, \mathbf{X}_c\right]\} \\
&= \mathbb{E}\left[\phi^j\left(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}\right) \phi_c^j\left(\mathbf{X}_1, \cdots, \mathbf{X}_c\right)\right] \\
&= 0,
\end{aligned}$$

pois $c < j$ e, pelo Lema 2.2.4 (i), $\phi_c^j(\mathbf{X}_1, \cdots, \mathbf{X}_c) = 0$.

(**iii**) Como H_n^j é uma U-estatística baseada no núcleo $\phi^{(j)}$, pode ser usado o teorema 2.1.4 para obter

$$\operatorname{Var} H_n^j = \binom{n}{j}^{-1} \sum_{c=1}^{j} \binom{j}{c} \binom{n-j}{j-c} \operatorname{Var} \phi_c^j(\mathbf{X}_1, \cdots, \mathbf{X}_c)$$

Mas, pelo Lema 2.2.4, $\phi_c^j(\mathbf{X}_1, \cdots, \mathbf{X}_c) = 0$ para $c = 1, \cdots, j-1$ e daí

$$\begin{aligned}
\operatorname{Var} H_n^j &= \binom{n}{j}^{-1} \operatorname{Var} \phi_j^j(\mathbf{X}_1, \cdots, \mathbf{X}_j) \\
&= \binom{n}{j}^{-1} \operatorname{Var} \phi^j(\mathbf{X}_1, \cdots, \mathbf{X}_j) \\
&= \binom{n}{j}^{-1} \delta_j^2.
\end{aligned}$$

□

Como aplicação da decomposição de Hoeffding se pode obter uma expressão alternativa para a variância de uma U-estatística em termos das esperanças das variâncias de $\phi^{(j)}$.

Corolário 2.2.6. A variância de uma U-estatística é dada por:

$$\operatorname{Var} U(\mathbf{X}_1, \cdots, \mathbf{X}_n) = \sum_{j=1}^{k} \binom{k}{j}^2 \binom{n}{j}^{-1} \delta_j^2,$$

onde $\delta_j^2 = \operatorname{Var} \phi^j(\mathbf{X}_1, \cdots, \mathbf{X}_j)$, com ϕ^j definido em (2.12).

Prova: Dos Teoremas 2.2.1 e 2.2.5 segue

$$\begin{aligned}
\operatorname{Var} U(\mathbf{X}_1, \cdots, \mathbf{X}_n) &= \operatorname{Var} \left[\sum_{j=1}^{k} \binom{k}{j} H_n^j + \mathbb{E}\left[\phi(\ \mathbf{X}_1, \cdots, \mathbf{X}_k)\right] \right] \\
&= \operatorname{Var} \sum_{j=1}^{k} \binom{k}{j} H_n^j \\
&= \sum_{j=1}^{k} \binom{k}{j} \operatorname{Var} H_n^j \\
&= \sum_{j=1}^{k} \binom{k}{j}^2 \binom{n}{j}^{-1} \delta_j^2.
\end{aligned}$$

□

No Teorema 2.1.4 foi obtido uma expressão para a variância de uma U-estatística em termos de $\sigma_c^2 = \operatorname{Var}\phi_c(\mathbf{X}_1, \cdots, \mathbf{X}_c)$. Na proposição a seguir será obtido uma relação entre σ_c^2 e δ_j^2 que aparece na expressão da variância obtida no corolário acima.

Proposição 2.2.7. Considere σ_c^2 definido em (2.5) e δ_j^2 em (2.26). Então

$$\sigma_c^2 = \sum_{j=1}^{c} \binom{c}{j} \delta_j^2.$$

Prova: Em (2.12), foi definido

$$\phi^{(c)} = \phi^c(\mathbf{X}_1, \cdots, \mathbf{X}_c) = \phi_c(\mathbf{X}_1, \cdots, \mathbf{X}_c) - \sum_{j=1}^{c-1} \sum_{C_{c,j}} \phi^j(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}) - \mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right].$$

Por outro lado,

$$\begin{aligned}
\sum_{j=1}^{c} \sum_{C_{c,j}} \phi^j(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}) + \mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right] &= \sum_{j=1}^{c-1} \sum_{C_{c,j}} \phi^j(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}) + \sum_{C_{c,c}} \phi^c(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_c}) \\
&\quad + \mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right] \\
&= \sum_{j=1}^{c-1} \sum_{C_{c,j}} \phi^j(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}) + \phi^c(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_c}) \\
&\quad + \mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right] \\
&= \phi^c(X_{\alpha_1}, \cdots, X_{\alpha_c})
\end{aligned}$$

Assim,

$$\phi^c(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_c}) = \sum_{j=1}^{c} \sum_{C_{c,j}} \phi^j(\mathbf{X}_{\alpha_1}, \cdots, \mathbf{X}_{\alpha_j}) + \mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right]. \tag{2.27}$$

Calculando a variância em ambos os lados de (2.27) e usando o Teorema 2.2.5, obtém-se

$$\begin{aligned}
\operatorname{Var}\phi^c(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_c}) &= \operatorname{Var}\left[\sum_{j=1}^{c}\sum_{C_{c,j}}\phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_j}) + \mathbb{E}\left[\phi(\ \mathbf{X}_1,\cdots,\mathbf{X}_k)\right]\right] \\
&= \operatorname{Var}\sum_{j=1}^{c}\sum_{C_{c,j}}\phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_j}) \\
&= \sum_{j=1}^{c}\operatorname{Var}\sum_{C_{c,j}}\phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_j}) \\
&= \sum_{j=1}^{c}\operatorname{Var}\left[\frac{\binom{c}{j}}{\binom{c}{j}}\sum_{C_{c,j}}\phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_j})\right] \\
&= \sum_{j=1}^{c}\binom{c}{j}^2\operatorname{Var}\left[\binom{c}{j}^{-1}\sum_{C_{c,j}}\phi^j(\ \mathbf{X}_{\alpha_1},\cdots,\ \mathbf{X}_{\alpha_j})\right]] \\
&= \sum_{j=1}^{c}\binom{c}{j}^2\operatorname{Var}H_c^j \\
&= \sum_{j=1}^{c}\binom{c}{j}^2\binom{c}{j}^{-1}\delta_j^2 \\
&= \sum_{j=1}^{c}\binom{c}{j}\delta_j^2.
\end{aligned}$$

□

Como consequência da proposição anterior obtém-se uma relação de ordem entre as variâncias $\sigma_c^2, 0 \le c \le k$.

Proposição 2.2.8. Para $0 \le c \le d \le k$

$$\frac{\sigma_c^2}{c} \le \frac{\sigma_d^2}{d}.$$

Prova: Será provado que $c\sigma_d^2 - d\sigma_c^2 \ge 0$. Pela Proposição 2.2.7, obtém-se

$$\begin{aligned}
c\sigma_d^2 - d\sigma_c^2 &= c\sum_{j=1}^{d}\binom{d}{j}\delta_j^2 - d\sum_{j=1}^{c}\binom{c}{j}\delta_j^2 \\
&= \sum_{j=1}^{c}\left[c\binom{d}{j} - d\binom{c}{j}\right]\delta_j^2 + \sum_{c+1}^{d}c\binom{d}{j}\delta_j^2.
\end{aligned}$$

O segundo termo da soma é positivo, pois para c arbitrário $\delta_c^2 \ge 0$. Para verificar que o primeiro termo também é positivo. Para isso basta verificar que $c\binom{d}{j} - d\binom{c}{j} \ge 0$ para $d \ge c \ge j \ge 1$. Mas,

$$c\binom{d}{j} - d\binom{c}{j} = c\frac{d!}{j!(d-j)!} - d\frac{c!}{j!(c-j)!}$$
$$= \frac{cd}{j!}(d-1)(d-2)\cdots(d-j+1) - \frac{cd}{j!}(c-1)(c-2)\cdots(c-j+1) \geq 0,$$

pois $d \geq j$. □

2.3 Normalidade assintótica

Aplicando a Decomposição de Hoeffding (Teorema 2.2.1), pode-se obter a normalidade assintótica de uma U-estatística.

Teorema 2.3.1. Seja $U_n(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ a U-estatística baseada em um núcleo simétrico $\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ correspondente a um parâmetro $\theta = \theta(F), F \in \mathcal{F}$, de grau k. Se $\mathbb{E}\left[\phi^2(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right] < \infty$ então $n^{\frac{1}{2}}(U_n - \mathbb{E}U_n)$ é assintoticamente normal, com média zero e variância assintótica $k^2\zeta_1 = k^2\sigma_1^2$.

Prova: Da Decomposição de Hoeffding (Teorema 2.2.1), pode-se escrever

$$\sqrt{n}(U_n - \mathbb{E}U_n) = \sqrt{n}\left[\mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right] + \sum_{j=1}^{k}\binom{k}{j}H_n^j - \mathbb{E}\left[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)\right]\right]$$
$$= \sqrt{n}\left[\binom{k}{1}H_n^1 + \sum_{j=2}^{k}\binom{k}{j}H_n^j\right]$$
$$= \sqrt{n}\left[\binom{k}{1}H_n^1 + R_n\right]$$
$$= \sqrt{n}\left[k\binom{n}{1}^{-1}\sum_{C_{n,1}}\phi^1(\mathbf{X}_i) + R_n\right]$$
$$= \sqrt{n}\left[\frac{k}{n}\sum_{i=1}^{n}\phi^1(\mathbf{X}_i) + R_n\right]$$

onde $R_n = \sum_{j=2}^{k} \binom{k}{j} H_n^j$. Do Teorema 2.2.5 seguem as seguintes igualdades,

$$\begin{aligned}
\operatorname{Var} \sqrt{n} R_n = n \operatorname{Var} R_n &= n \operatorname{Var} \sum_{j=2}^{k} \binom{k}{j} H_n^j \\
&= n \sum_{j=2}^{k} \operatorname{Var} \binom{k}{j} H_n^j \\
&= \sum_{j=2}^{k} n \binom{k}{j}^2 \operatorname{Var} H_n^j \\
&= \sum_{j=2}^{k} n \binom{k}{j}^2 \binom{n}{j}^{-1} \delta_j^2.
\end{aligned}$$

Então $\operatorname{Var} \sqrt{n} R_n = O(n^{-1})$ e portanto tem que o comportamento assintótico de $\sqrt{n}(U_n - \mathbb{E}U_n)$ é o mesmo de $\frac{k}{\sqrt{n}} \sum_{i=1}^{n} \phi(\mathbf{X}_i)$. Agora,

$\mathbb{E}\phi^1(\mathbf{X}_1) = \mathbb{E}\phi_1(\mathbf{X}_1) - \mathbb{E}[\phi(\mathbf{X}_1, \cdots, \mathbf{X}_k)] = 0$ e

$\operatorname{Var} \phi^1(\mathbf{X}_1) = \phi_1(\mathbf{X}_1) = \sigma_1^2 = \zeta_1 < \infty$, então como $\phi^1(\mathbf{X}_1), i \geq 1$ são $i.i.d$, segue do Teorema do Limite Central

$$\frac{\sum_{i=1}^{n} \phi^1(\mathbf{X}_i)}{\sqrt{n}} \xrightarrow{d} N(0,1), n \longrightarrow \infty,$$

o que conclui a prova do Teorema 2.3.1. □

Capítulo 3

Métodos de classificação baseado em U-estatísticas

Nesse capítulo apresentamos o classificador AH de Ahmad and Pavlenko (2018), introduzimos a metodologia baseada na estatística B_n, proposta por Sen (2006), estudada por Pinheiro et al. (2009a) e estendida por Cybis et al. (2018); Valk and Cybis (2020). Baseado nessa estatística B_n, apresentamos o nosso classificador UC e estudamos suas propriedades estatísticas.

3.1 *O Classificador* AH

No trabalho de Ahmad and Pavlenko (2018) é proposto um classificador para dois ou mais grupos para dados de alta dimensão e as distribuições podem ser não normais. O classificador é construído por uma combinação linear de duas componentes, a componente U e a componente P. A componente U é uma combinação linear de U-estatísticas que são médias de formas bilineares de vetores distintos de duas amostras independentes. Já a componente P é uma função da projeção da componente U na observação $\mathbf{X}^*$ a ser classificada.

Considere $\mathscr{G} \geq 2$ grupos (populações) cujos elementos (dados) são de alta dimensão, em que para cada $g \in \{1, \cdots, \mathscr{G}\}$ os elementos $\mathbf{X}_1^{(g)}, \cdots, \mathbf{X}_{n_g}^{(g)}$ são vetores aleatórios iid do g-ésimo grupo de tamanho n_g, com distribuição desconhecida $\boldsymbol{\mathcal{F}}_g$. Assumindo p características (dimensão) da amostra e para todo $i \in \{1, \cdots, n_g\}, \mathbf{X}_i^{(g)} = (\, \mathbf{X}_{i1}^{(g)}, \cdots, \, \mathbf{X}_{ip}^{(g)})^{\mathsf{T}}$, com g pertencente a g-ésima população (desconhecida), com matriz de covariâncias $\mathrm{Cov}(\mathbf{X}_i^{(g)}) = \boldsymbol{\Sigma}_g$, positiva definida, e vetor de médias $\mathbb{E}(\mathbf{X}_i^{(g)}) = \boldsymbol{\mu}_g$. O estimador não viesado de $\boldsymbol{\mu}_g$ é dado por

$$\overline{\mathbf{X}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{X}_i^{(g)}.$$

Considerando dois grupos para cada $g \in \{1, 2\}$, seja $\mathbf{X}_i^{(g)} = (x_{i1}^{(g)}, \cdots, x_{ip}^{(g)})^\top \sim \boldsymbol{\mathcal{F}}_g$ como definido acima e $\boldsymbol{\pi}_g$ denota a g-ésima população desconhecida. Considere também $\mathcal{R}_g = \{\mathbf{X}^* : \mathbf{X}^* \in \boldsymbol{\pi}_g\}$ a região de dados observados da g-ésima população, em que $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathcal{X}$, $\mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset$ com $\mathcal{X}$ os espaço dos $\mathbf{X}^*$ observados e $\emptyset$ o conjunto vazio. Além disso, seja $\boldsymbol{\theta}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}$ o conjunto de parâmetros de $\boldsymbol{\mathcal{F}}_g$.

O classificador propostos por Ahmad and Pavlenko (2018) consiste em uma modificação da função *custo* associada ao erro de classificação incorreta para 2 grupos dada por

$$C(\mathbf{X}^*) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{X}^* - (\bar{\mathbf{X}}_1^\mathsf{T}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2^\mathsf{T}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{X}}_2)/2,$$

onde $\mathbf{X}^*$ é o ponto a ser classificado e $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ são iguais e conhecidas. Além disso, os custos e as probabilidades de incidência a priori são iguais. Na prática $\boldsymbol{\Sigma}$ é desconhecido, por conta disso, é usada a matriz de variância-covariância estimada comum ou agregada dada por

$$\hat{\boldsymbol{\Sigma}}_{pooled} = \sum_{g=1}^{2}(n_g - 1)\hat{\boldsymbol{\Sigma}}_i / \sum_{i=1}^{2}(n_g - 1).$$

Assim, obtém-se o estimador de C($\mathbf{X}^*$)

$$\hat{C}(\mathbf{X}^*) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\mathsf{T}\hat{\boldsymbol{\Sigma}}_{pooled}^{-1}\mathbf{X}^* - (\bar{\mathbf{X}}_1^\mathsf{T}\hat{\boldsymbol{\Sigma}}_{pooled}^{-1}\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2^\mathsf{T}\hat{\boldsymbol{\Sigma}}_{pooled}^{-1}\bar{\mathbf{X}}_2)/2. \tag{3.1}$$

No contexto HDLSS, a matriz $\hat{\Sigma}_{pooled}$ é singular, portanto, $\hat{C}(\mathbf{X}^*)$ não pode ser usado, então Ahmad and Pavlenko (2018) propôs primeiramente retirar a $\hat{\boldsymbol{\Sigma}}_{pooled}$ do classificador (3.1) e consideraram

$$\tilde{C}(\mathbf{X}^*) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\mathsf{T}\mathbf{X}^* - (\bar{\mathbf{X}}_1^\mathsf{T}\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2^\mathsf{T}\bar{\mathbf{X}}_2)/2. \tag{3.2}$$

Utilizando-se das propriedades apresentadas no capítulo 2 do Searle (1971) a saber $E(\bar{\mathbf{X}}_g^\mathsf{T}\bar{\mathbf{X}}_g)) = B_g + \|\boldsymbol{\mu}_g\|^2$, onde $\|\mathbf{a}\|^2 = \mathbf{a}^\mathsf{T}\mathbf{a}$ é a norma euclidiana do vetor $\mathbf{a}$ e $B_g = tr(\boldsymbol{\Sigma}_g)/n_g$ para todo $g \in \{1, 2\}$. Assume-se que $\mathbf{X}^* \in \boldsymbol{\pi}_g$, teremos

$$E\{\tilde{C}(\mathbf{x}^*)|\mathbf{X}^* \in \boldsymbol{\pi}_g\} = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/2 - B,$$

com $B = (B_1 - B_2)/2$. Como Ahmad and Pavlenko (2018) removeu $\hat{\Sigma}_{pooled}$ do classificador (3.1), $\tilde{C}(\mathbf{X}^*)$ se tornou viesado com termo de viés B composto pelos traços das matrizes de variâncias e covariâncias desconhecidas. Se $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, então $B = (n_2 - n_1)tr(\boldsymbol{\Sigma})/(2n_1n_2)$ e o classificador tem viés positivo(negativo) quando $n_2 > n_1 (n_2 < n_1)$, e não viesado para $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ e $n_2 = n_1$. Para a correção do viés e para melhorar a precisão foi considerado o segundo componente de $\tilde{C}(\mathbf{X}^*)$ em (3.2) com

$$E\{(\bar{\mathbf{X}}_1^\mathsf{T}\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2^\mathsf{T}\bar{\mathbf{X}}_2)/2\} = B + (\|\mu_1\|^2 - \|\mu_2\|^2)/2.$$

Escrevendo

$$\bar{\mathbf{X}}_g^\mathsf{T}\bar{\mathbf{X}}_g = \frac{1}{n_g^2}\sum_{k=1}^{n_g}\sum_{r=1}^{n_i} A_{gir} = \frac{1}{n_g^2}\sum_{i=1}^{n_g} A_{gi} + \frac{1}{n_g^2}\sum_{i=1}^{n_g}\sum_{r=1k\neq r}^{n_g} A_{gi} = Q_{0g} + Q_{1g},$$

onde $A_{gi} = \mathbf{X}_i^{(g)\mathsf{T}}\mathbf{X}_i^{(g)}$, $A_{gir} = \mathbf{X}_i^{(g)\mathsf{T}}\mathbf{X}_r^{(g)}, i \neq r$. Além disso

$$E(Q_{0g}) = B_g + R_g \text{ e } E(Q_{1g}) = \|\boldsymbol{\mu}_g\|^2 - R_g,$$

com $R_g = \|\boldsymbol{\mu}\|^2/n_g$. Denotando $Q_0 = Q_{01} - Q_{02}$, $Q_1 = Q11 - Q_{12}$, $R = R_1 - R_2$ tem-se

$$\mathbb{E}(Q_0) = 2B + R \text{ e } E(Q - 1) = (\|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2) - R.$$

Ao ajustar ambos os componentes para R obtém-se $E(Q_0) - R = 2B$, $E(Q-1)+R = (\|\mu_1\|^2 - \|\mu_2\|^2)$. Assim, levando uma versão imparcial de $\tilde{C}(\mathbf{X}^*)$ em (3.2) denotado pelos autores por $A(\mathbf{X}^*)$ e a definiram como:

$$A(\mathbf{X}^*) = (\mathbf{X}^*)^\mathsf{T}(\bar{\mathbf{X}}_g - \bar{\mathbf{X}}_{g'})/p - (U_{n_g} - U_{n_{g'}})/2, \tag{3.3}$$

em que $U_{n_g} = \sum_{i \neq r}^{n_g} A_{gir}/pQ(n_g)$ é uma U-estatística com kernel simétrico, $Q(n_g) = n_g(n_g - 1)$ e $A_{gir}/P = \mathbf{X}_i^{(g)\mathsf{T}} \mathbf{X}_r^{(g)}/P$, $i \neq r$ é uma forma bilinear de componentes independentes.

Assumindo $\mathbf{X}^* \in \boldsymbol{\pi}_1$ e que as amostras são independentes, segue que

$$\mathbb{E}\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\mathsf{T} \mathbf{X}^*\} = \|\boldsymbol{\mu}_1\|^2 - \boldsymbol{\mu}_2^\mathsf{T} \boldsymbol{\mu}_1,$$

e portanto

$$\mathbb{E}\{A(\mathbf{X}^*)|\boldsymbol{\pi}_1\} = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/(2p),$$

sem termo de viés. Além disso, com $\mathbf{X}^* \in \boldsymbol{\pi}_1$, $A(\mathbf{X}^*)$ é composta por formas bilineares, sendo duas da amostra 1, uma da amostra 2 e uma mistura. Por simetria

$$\mathbb{E}\{A(\mathbf{X}^*)|\boldsymbol{\pi}_2\} = -\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/(2p),$$

com $A(\mathbf{X}^*)$ composta de formas bilineares, sendo duas da amostra 2, uma da amostra 1 e uma mistura. Para o caso de duas amostras, Ahmad and Pavlenko (2018), define a regra de classificação da seguinte forma: uma nova amostra $\mathbf{X}^*$ pertence a $\boldsymbol{\pi}_1$ (população 1) se $A(\mathbf{X}^*) > 0$ caso contrário $\mathbf{X}^*$ pertence a $\boldsymbol{\pi}_2$ (população 2) e estendeu o caso para múltiplos grupos, para maiores detalhes consultar Ahmad and Pavlenko (2018).

Este classificador $A(\mathbf{X}^*)$ (3.3) é inteiramente composto de quantidades empíricas, livre de qualquer parâmetro desconhecido, além disso, é linear mesmo que $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. No caso clássico, a violação da homocedasticidade torna o classificador quadrático, o que não é o caso de $A(\mathbf{X}^*)$. Isso fornece uma vantagem adicional ao classificador proposto para que possa ser usado sem assumir ou testar a homocedasticidade. Outra vantagem é que $A(\mathbf{X}^*)$ pode ser usado diretamente no conjunto de dados sem quaisquer pré-requisitos de redução da dimensão através de classificação, *clustering*, análise de componentes principais (PCA) ou outros meios. Já a substituição da suposição de normalidade é feita através de certas suposições sobre os momentos do modelo multivariado subjacente e sobre os traços das matrizes de covariância. A distribuição assintótica desse estimador pode ser encontrada em Ahmad and Pavlenko (2018).

3.2 *U-estatísticas no contexto de agrupamentos com inferência em dados de alta dimensão*

De maneira geral, fazer inferência no contexto de alta dimensão e baixo tamanho amostral, o HDLSS, é uma tarefa complicada. A abordagem tradicional da estatística pressupõe que a amostra seja representativa da população e os resultados assintóticos normalmente são construídos baseados no crescimento da amostra. Isso não seria diferente quando o objetivo é fazer agrupamentos estatisticamente significativos. Os trabalhos de Cybis et al. (2018) e Valk and Cybis (2020) propõe um método para agrupar de forma hierárquica um conjunto de dados, de forma não supervisionada, ou seja, não necessitando de nenhum tipo de conhecimento prévio de qual grupo pertence determinada amostra. Essa abordagem pressupõe que os dados são homogêneos, advindos de uma única distribuição de probabilidade, e o agrupamento resultante é construído de forma que somente grupos estatisticamente significativos sejam criados. Essa metodologia baseia-se nos trabalhos de Sen (2006) e Pinheiro et al. (2009b), os quais apresentam uma estatística, B_n, que mede a distância entre grupos e dentro dos grupos e

mostram que ela é uma U-estatística, e com isso tem todas as propriedades interessantes mostrados na Seção 2.2. Esse é o ponto de partida para o desenvolvimento da nossa metodologia de classificação e por esse motivo vamos apresentar a construção da estatística B_n na próxima seção.

3.3 *Definição e propriedades estatística B_n*

Seja $\mathbf{X}_1, \ldots, \mathbf{X}_n$ uma amostra aleatória de n vetores p-dimensional cada $\mathbf{X}_i$, $i = 1, \ldots, n$. Tradicionalmente, as amostras são classificadas em $\mathscr{G}$ grupos G_g, $g = 1, \ldots, \mathscr{G}$, sendo o tamanho do g-ésimo grupo G_g é dado por n_g, onde $n = n_1 + \cdots + n_{\mathscr{G}}$. No g-ésimo grupo as observações $\mathbf{X}_1^{(g)}, \ldots, \mathbf{X}_{n_g}^{(g)}$ são consideradas independentes e identicamente distribuídos com uma distribuição F_g, p-dimensional e com vetor de média finita $\boldsymbol{\mu}_g$ e a matriz de dispersão $\boldsymbol{\Sigma}_g$ positiva definida (não necessariamente normal multivariada). Seguindo a abordagem de Sen (2006) e Pinheiro et al. (2009b), definimos parâmetro funcional $\theta(F_g, F_{g'})$ como

$$\theta(F_g, F_{g'}) = \int\int \phi(x_1, x_2) dF_g(x_1) dF_{g'}(x_2), \quad x_1, x_2 \in \mathbb{R}^L, \tag{3.4}$$

onde $g \neq g'$ e $\phi(\cdot, \cdot)$ é um kernel simétrico de ordem 2. Se assumirmos que $\theta(\cdot, \cdot)$ em (3.4) é uma função linear convexa nos componentes marginais, então temos

$$\theta(F_g, F_{g'}) \geq \frac{1}{2}\{\theta(F_g, F_g) + \theta(F_{g'}, F_{g'})\}, \tag{3.5}$$

para toda distribuição F_g e $F_{g'}$. A igualdade em (3.5) ocorre somente quando temos $\mu_g = \mu_{g'}$ (assumindo Homogeneidade).

Segue-se da teoria U-estatística que o estimador não viesado para o parâmetro funcional dentro do grupo $\theta(F_g, F_g)$ é a U-estatística Hoeffding (1948), com kernel $\phi(\cdot, \cdot)$, dada por

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g)}). \tag{3.6}$$

Analogamente, o estimador não viesado para o parâmetro funcional entre grupos $\theta(F_g, F_{g'})$ é dado por

$$U_{n_g, n_{g'}}^{(g,g')} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}), \tag{3.7}$$

onde $g \neq g'$. A U-estatística geral pode ser decomposta em

$$\begin{aligned} U_n &= \sum_{g=1}^{K} \frac{n_g}{n} U_{n_g}^{(g)} + \sum_{1 \leq g < g' \leq K} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g, n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\} \\ &= W_n + B_n. \end{aligned} \tag{3.8}$$

A decomposição (3.8) nos fornece a estatística B_n, que é o principal ponto da nossa metodologia, sendo dada por

$$B_n = \sum_{1 \leq g < g' \leq K} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g, n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\}. \tag{3.9}$$

Na qual $U_{n_g}^{(g)}$ para $g \in \{1, , \ldots, K\}$ é a U-estatística associada à distância dentro do grupo, como definida em (3.6), e $U_{n_g n_{g'}}^{(g,g')}$, $g \neq g' \in \{1, \ldots, \mathscr{G}\}$, é a U-estatística associada à distância entre grupos, como definida em (3.7).

3.4 *A estatística B_n no contexto de agrupamento*

No trabalho de Cybis et al. (2018) a estatística B_n em (3.9) é utilizada para construir um teste de hipóteses para a homogeneidade de dados do tipo que foi apresentado na Seção anterior. Esse método separa os dados em dois grupos de acordo com a configuração que maximiza a U-estatística B_n e utiliza os resultados assintóticos para verificar a significância dessa partição. Caso seja significativa, a partição (grupos) é criada e o método é aplicado novamente nos subgrupos. Naturalmente, a teoria usada é considerando $\mathscr{G} = 2$ grupos. Assim a estatística B_n utilizada é da forma

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

que compara diferenças dentro e entre grupos para uma partição da amostra em dois subgrupos, onde $U_{n_1}^{(1)}$ e $U_{n_2}^{(2)}$ são U-estatísticas associadas a diferenças dentro do grupo, conforme definido em (3.6) e $U_{n_1 n_2}^{(1,2)}$ é a U-estatística associada a diferença entre grupos definida em (3.7).

Para abordar a questão de mensurar o grau de separação de dois grupos é analisado se os valores de B_n são grandes, indicando grande separação dos grupos. Também foi proposto nesse trabalho um teste para verificar a significância da classificação de uma observação da amostra em um dos dois grupos. Para isso foi sugerido uma abordagem comparativa baseada em $B_n^{(1)}$ e $B_n^{(2)}$, onde $B_n^{(1)}$ é a estatística B_n quando uma observação $\mathbf{X}^*$ da amostra é classificada no grupo $\boldsymbol{\pi}_1$ e $B_n^{(2)}$ é definido da mesma forma. Observado que se $\mathbf{X}^*$ não estiver bem classificado em $\boldsymbol{\pi}_2$ espera-se que a estatística $B_n^{(2)}$ seja menor do que B_n calculado sem incluir a nova observação, pois isso aumenta as distâncias dentro do grupo $\boldsymbol{\pi}_2$. Assim, se $B_n^{(1)}$ for maior que $B_n^{(2)}$, classificando a nova observação no grupo $\boldsymbol{\pi}_1$ produz um melhor agrupamento no sentido de que as distâncias dentro os grupos são comparativamente menores do que as distâncias entre os grupos. Embora este procedimento forneça um critério empírico para classificação, ele não avalia a significância estatística, logo foi proposto pelos autores um teste de classificação com base na diferença $DB_n = B_n^{(1)} - B_n^{(2)}$ e $\mathbb{E}(DB_n) = \mu_{B_n^{(1)}} - \mu_{B_n^{(2)}} = \mu_{DB_n}$, onde $\mu_{B_n^{(1)}}$ e $\mu_{B_n^{(2)}}$ são, respectivamente, os valores esperados das estatísticas $B_n^{(1)}$ e $B_n^{(2)}$. A hipótese nula afirma que $\mathbf{X}^*$ pertence ao grupo $\boldsymbol{\pi}_2$ e, portanto, o arranjo de amostra que produz $B_n^{(2)}$ é melhor do que aquele que produz $B_n^{(1)}$. A hipótese alternativa afirma que $\mathbf{X}^*$ está corretamente classificado no grupo $\boldsymbol{\pi}_1$. A hipótese nula e a hipótese alternativa são dadas como segue:

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

No entanto, a distribuição de DB_n não é conhecida, portanto, Cybis et al. (2018) empregaram uma técnica baseada em reamostragem para encontrar a distribuição empírica. Essa proposta foi apresentada mas foi pouco explorada naquele artigo. Além disso, da maneira proposta é possível que a μ_{DB_n} seja negativo, basta que a amostra de fato pertença ao grupo 2. Nesse sentido, apresentamos na

próxima seção um estudo completo sobre as propriedades estatísticas dessa estatística e reformulamos o teste de hipóteses.

3.5 Um classificador baseado na U-estatística B_n

Nessa Seção vamos definir o classificador para dados de alta dimensionalidade e derivar suas propriedades estatísticas, como média, variância e distribuição assintótica. Como havíamos antecipado, o classificador é baseado na estatística B_n e, consequentemente, podemos utilizar toda a teoria de U-estatísticas para obter essas propriedades. Vamos considerar aqui o mesmo contexto apresentado na Seção 3.1, com suposições equivalentes para os dados, sendo que inicialmente consideramos dados advindos de duas distribuições distintas, ou seja, temos amostras em dois grupos $\boldsymbol{\pi}_1$ e $\boldsymbol{\pi}_2$. Tradicionalmente a estatística de teste B_n é apresentada da forma

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

Usando as definições das U-estatísticas $U_{n_1 n_2}^{(1,2)}$, $U_{n_1}^{(1)}$ e $U_{n_2}^{(2)}$, podemos reescrever a B_n da forma

$$B_n = \frac{n_1 n_2}{n(n-1)} \left[\frac{2}{n_1 n_2} \sum_{i}^{n_1} \sum_{j}^{n_2} \phi\left(\mathbf{X}_i^{(1)}, \mathbf{X}_j^{(2)} \right) - \frac{2}{n_1(n_1-1)} \sum_{1 \le i < j \le n_1} \phi\left(\mathbf{X}_i^{(1)}, \mathbf{X}_j^{(1)} \right) - \frac{2}{n_2(n_2-1)} \sum_{1 \le i < j \le n_2}^{n_2} \phi\left(\mathbf{X}_i^{(2)}, \mathbf{X}_j^{(2)} \right) \right].$$

Para um novo elemento $\mathbf{X}^*$, é importante notar que ou $\mathbf{X}^*$ pertence ao grupo 1 ($\boldsymbol{\pi}_1$) ou $\mathbf{X}^*$ pertence ao grupo 2 ($\boldsymbol{\pi}_2$). Note que ainda válida a condição $n = n_1 + n_2$, sendo que a nova observação $\mathbf{X}^*$ ainda não está contabilizada. A princípio, sem perda de generalidade, podemos alocá-lo no grupo 1. Nesse caso, a estatística B_n pode ser escrita como

$$B_n^{(1)} = \frac{(n_1+1)n_2}{n(n+1)} \left[\frac{2}{(n_1+1)n_2} \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi\left(\mathbf{X}_i^{(1)}, \mathbf{X}_j^{(2)} \right) + \sum_{j=1}^{n_2} \phi\left(\mathbf{X}^{(*)}, \mathbf{X}_j^{(2)} \right) \right) - \frac{2}{n_1(n_1+1)} \left(\sum_{1 \le i < j \le n_1} \phi\left(\mathbf{X}_i^{(1)}, \mathbf{X}_j^{(1)} \right) + \sum_{i=1}^{n_1} \phi\left(\mathbf{X}_i^{(1)}, \mathbf{X}^{(*)} \right) \right) - \frac{2}{n_2(n_2-1)} \sum_{1 \le i < j \le n_2} \phi\left(\mathbf{X}_i^{(2)}, \mathbf{X}_j^{(2)} \right) \right].$$

Analogamente, ao supor que $\mathbf{X}^*$ pertence ao grupo 2, podemos observar que a estatística B_n é expressa da forma

$$B_n^{(2)} = \frac{(n_2+1)n_1}{n(n+1)}\left[\frac{2}{(n_2+1)n_1}\left(\sum_{i}^{n_1}\sum_{j}^{n_2}\phi(\mathbf{X}_i^{(1)},\mathbf{X}_j^{(2)}) + \sum_{j=1}^{n_1}\phi(\mathbf{X}^{(*)},\mathbf{X}_j^{(1)})\right)\right.$$
$$- \frac{2}{n_2(n_2+1)}\left(\sum_{1\le i<j\le n_2}\phi(\mathbf{X}_i^{(2)},\mathbf{X}_j^{(2)}) + \sum_{i=1}^{n_2}\phi(\mathbf{X}_i^{(1)},\mathbf{X}^{(*)})\right)$$
$$\left. - \frac{2}{n_1(n_1-1)}\sum_{1\le i<j\le n_1}\phi(\mathbf{X}_i^{(1)},\mathbf{X}_j^{(1)})\right].$$

Assim como mencionado em Cybis et al. (2018), segue da concepção da estatística B_n que se o novo elemento $\mathbf{X}^*$ foi corretamente alocado no grupo 1, é esperado que o módulo da diferença entre $B_n^{(1)}$ e $B_n^{(2)}$ seja grande. Se $\mathbf{X}^*$ foi incorretamente alocado no grupo 2, então é esperado que o módulo da diferença entre $B_n^{(1)}$ e $B_n^{(2)}$ seja pequena. No entanto, sabidamente em estatística a quantificação do que pode ser considerado grande ou pequeno depende da distribuição dessa diferença. Assim, propomos uma estatística que será nossa estatística de teste e será definida por:

$$DB_n = B_n^{(1)} - B_n^{(2)}. \tag{3.10}$$

Segue do trabalho de Pinheiro et al. (2009a) que tanto a $B_n^{(1)}$ quanto a $B_n^{(2)}$ são assintoticamente Normais em n e/ou p. Lembrando que nesse caso a razão de convergência é a tradicional $\sqrt{n}$ (e/ou $\sqrt{p}$). Notadamente, a estatística DB_n é uma diferença de normais e portanto sua distribuição também será assintoticamente normal. Para formalizar esse resultado, enunciamos o Teorema 3.6.1.

A partir das definições de $B_n^{(1)}$ e $B_n^{(2)}$ podemos escrever DB_n como

$$\begin{aligned} DB_n &= B_n^{(1)} - B_n^{(2)} \\ &= \frac{2}{n(n+1)}\left[\sum_{j=1}^{n_2}\phi\left(\mathbf{X}^{(*)},\mathbf{X}_j^{(2)}\right) - \sum_{j=1}^{n_1}\phi\left(\mathbf{X}_i^{(*)},\mathbf{X}_j^{(1)}\right)\right. \\ &- \left(\frac{n_2}{n_1} - \frac{(n_2+1)}{(n_1-1)}\right)\sum_{1\le i<j\le n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}_j^{(1)}\right) - \left(\frac{n_2}{n_1}\right)\sum_{j=1}^{n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}^{(*)}\right) \\ &\left. - \left(\frac{(n_1+1)}{(n_2-1)} - \frac{n_1}{n_2}\right)\sum_{1\le i<j\le n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}_j^{(2)}\right) + \left(\frac{n_1}{n_2}\right)\sum_{i=1}^{n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}^{(*)}\right)\right]. \end{aligned}$$

Podemos usar as seguintes simetrias

$$\sum_{i=1}^{n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}^{(*)}\right) = \sum_{j=1}^{n_1}\phi\left(\mathbf{X}^{(*)},\mathbf{X}_j^{(1)}\right);$$

$$\sum_{i=1}^{n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}^{(*)}\right)=\sum_{j=1}^{n_2}\phi\left(\mathbf{X}^{(*)},\mathbf{X}_j^{(2)}\right)$$

para reescrever a DB_n como

$$\begin{aligned}DB_n=&\frac{2}{n(n+1)}\left[\left(1+\frac{n_1}{n_2}\right)\sum_{i=1}^{n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}^{(*)}\right)-\left(1+\frac{n_2}{n_1}\right)\sum_{i=1}^{n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}^{(*)}\right)\right.\\&\left.-\left(\frac{n_2}{n_1}-\frac{(n_2+1)}{(n_1-1)}\right)\sum_{1\le i<j\le n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}_j^{(1)}\right)-\left(\frac{(n_1+1)}{(n_2-1)}-\frac{n_1}{n_2}\right)\sum_{1\le i<j\le n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}_j^{(2)}\right)\right]\\=&\frac{2}{n(n+1)}\left[\left(\frac{n_1+n_2}{n_2}\right)\sum_{i=1}^{n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}^{(*)}\right)-\left(\frac{n_1+n_2}{n_1}\right)\sum_{i=1}^{n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}^{(*)}\right)\right.\\&\left.+\left(\frac{n_1+n_2}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}_j^{(1)}\right)-\left(\frac{n_1+n_2}{(n_2-1)n_2}\right)\sum_{1\le i<j\le n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}_j^{(2)}\right)\right]\end{aligned}$$

Além disso, temos que $n_1+n_2=n$, e, portanto a estatística DB_n pode ser apresentada da forma

$$\begin{aligned}DB_n=&\frac{2}{(n+1)}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}^{(*)}\right)-\frac{1}{n_1}\sum_{i=1}^{n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}^{(*)}\right)\right.\\&+\left(\frac{1}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\phi\left(\mathbf{X}_i^{(1)},\mathbf{X}_j^{(1)}\right)\\&\left.-\left(\frac{1}{(n_2-1)n_2}\right)\sum_{1\le i<j\le n_2}\phi\left(\mathbf{X}_i^{(2)},\mathbf{X}_j^{(2)}\right)\right]\end{aligned}\tag{3.11}$$

De forma geral, podemos escrever a DB_n em (3.11) como

$$DB_n=\frac{1}{(n+1)}\left(2U_{n_2}^{(*2)}-2U_{n_1}^{(*1)}+U_{n_1}^{(1)}-U_{n_2}^{(2)}\right)$$

em que $U_{n_g}^{(*g)}=\frac{1}{n_g}\sum_{i=1}^{n_g}\phi\left(\mathbf{X}_i^{(g)},\mathbf{X}^{(*)}\right)$, $g=1,2$ são as U-estatísticas obtidas ao considerar um grupo versus o novo elemento $\mathbf{X}^*$.

3.6 *Propriedades da U-estatística de teste DB_n*

Na Seção anterior construímos a estatística de teste DB_n e mostramos que a mesma é uma combinação linear de U-estatísticas. Assim, um caminho natural para estudar as propriedades da DB_n é considerar o arcabouço da teoria de U-estatísticas. Notamos que o Kernel $\phi(\cdot,\cdot)$ deve satisfazer as suposições tradicionais (ver Pinheiro et al. (2009a)). Com isso, podemos definir o seguinte parâmetro

$$\theta^{(g,g')}=\mathbb{E}\left[\phi\left(\mathrm{X}_i^{(g)},\mathrm{X}_j^{(g')}\right)\right],\ \text{para } g,g'=1,2,$$

que é a média não condicional do Kernel $\phi(\cdot,\cdot)$. A Decomposição de Hoeffding da U-estatística DB_n depende da possibilidade de escrever o Kernel $\phi(\cdot,\cdot)$ como combinação linear de componentes ortogonais da forma

$$\phi\left(\mathbf{x}_i^{(g)}, \mathbf{X}_j^{(g')}\right) = \theta^{(g,g')} + \psi_1^{(g,g')}(\mathbf{X}_i^{(g)}) + \psi_1^{(g,g')}(\mathbf{X}_j^{(g')}) + \psi_2^{(g,g')}(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}), \tag{3.12}$$

em que $\psi_1^{(g,g')}(\mathbf{X}_i^{(g)})$, $\psi_1^{(g,g')}(\mathbf{X}_i^{(g')})$ e $\psi_2^{(g,g')}(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')})$ representam, respectivamente, os termos de primeira e segunda ordem da Decomposição de Hoeffding, definidos por

$$\psi_1^{(g,g')}\left(\mathbf{X}_i^{(g)}\right) = \phi_1^{(g,g')}\left(\mathbf{X}_i^{(g)}\right) + \theta^{(g,g')},$$

$$\psi_1^{(g,g')}\left(\mathbf{X}_i^{(g')}\right) = \phi_1^{(g,g')}\left(\mathbf{X}_i^{(g')}\right) + \theta^{(g,g')}$$

e

$$\psi_2^{(g,g')}\left(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}\right) = \phi_2^{(g,g')}\left(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}\right) - \phi_1^{(g,g')}\left(\mathbf{X}_i^{(g)}\right) - \phi_1^{(g,g')}\left(\mathbf{X}_j^{(g')}\right) + \theta^{(g,g')}$$

em que

$$\phi_1^{(g,g')}\left(\mathrm{X}_i^{(g)}\right) = \mathbb{E}\left[\phi\left(\mathrm{X}_i^{(g)}, \mathrm{X}_j^{(g')}\right) \mid \mathrm{X}_i^{(g)} = \mathrm{X}_i^{(g)}\right];$$
$$\phi_1^{(g,g')}\left(\mathrm{X}_j^{(g')}\right) = \mathbb{E}\left[\phi\left(\mathrm{X}_i^{(g)}, \mathrm{X}_j^{(g')}\right) \mid \mathrm{X}_j^{(g')} = \mathrm{X}_j^{(g')}\right]$$

e

$$\phi_2^{(g,g')}\left(\mathrm{X}_i^{(g)}, \mathrm{X}_j^{(g')}\right) = \mathbb{E}\left[\phi\left(\mathrm{X}_i^{(g)}, \mathrm{X}_j^{(g')}\right) \mid \mathrm{X}_i^{(g)} = \mathrm{X}_i^{(g)}, \mathrm{X}_j^{(g')} = \mathrm{x}_j^{(g')}\right].$$

Notamos que pelas propriedade da Decomposição de Hoeffding

$$\mathbb{E}\left[\psi_1^{(g,g')}\left(\mathbf{X}_i^{(g)}\right)\right] = 0;$$
$$\mathbb{E}\left[\psi_1^{(g,g')}\left(\mathbf{X}_j^{(g')}\right)\right] = 0;$$
$$\mathbb{E}\left[\psi_2^{(g,g')}\left(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}\right)\right] = 0$$

e

$$\mathrm{Cov}\left[\psi_1^{(g,g')}\left(\mathbf{X}_i^{(g)}\right), \psi_1^{(g,g')}\left(\mathbf{X}_j^{(g')}\right)\right] = 0;$$
$$\mathrm{Cov}\left[\psi_1^{(g,g')}\left(\mathbf{X}_i^{(g)}\right), \psi_2^{(g,g')}\left(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}\right)\right] = 0;$$

Considerando agora o novo elemento $\mathbf{X}^*$ a Decomposição de Hoeffding da U-estatística de teste DB_n pode ser apresentada da forma

$$\begin{aligned} DB_n =& \frac{2}{(n+1)}\Bigg[\left(\frac{1}{n_2}\right)\sum_{i=1}^{n_2}\left(\theta^{(2,1)} + \psi_1^{(2,1)}(x_i^{(2)}) + \psi_1^{(2,1)}\ \mathbf{X}^{(*)}) + \psi_2^{(2,1)}(x_i^{(2)},\ \mathbf{X}^{(*)})\right) \\ &- \left(\frac{1}{n_1}\right)\sum_{i=1}^{n_1}\left(\theta^{(1,1)} + \psi_1^{(1,1)}(X_i^{(1)}) + \psi_1^{(1,1)}(\ \mathbf{X}^{(*)}) + \psi_2^{(1,1)}(x_i^{(1)},\ \mathbf{X}^{(*)})\right) \\ &+ \left(\frac{1}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\left(\theta^{(1,1)} + \psi_1^{(1,1)}(\ \mathbf{X}_i^{(1)}) + \psi_1^{(1,1)}(\ \mathbf{X}_j^{(1)}) + \psi_2^{(1,1)}(\ \mathbf{X}_i^{(1)},\ \mathbf{X}_j^{(1)})\right) \\ &- \left(\frac{1}{n_2(n_2-1)}\right)\sum_{1\le i<j\le n_2}\left(\theta^{(2,2)} + \psi_1^{(2,2)}(\ \mathbf{X}_i^{(2)}) + \psi_1^{(2,2)}(\ \mathbf{X}_j^{(2)}) + \psi_2^{(2,2)}(\ \mathbf{X}_i^{(2)},\ \mathbf{X}_j^{(2)})\right)\Bigg] \end{aligned}$$

onde a notação $X^{(*)}$ significa que a distribuição associada a esse novo elemento $\mathbf{X}^*$ pode ser $\boldsymbol{\pi}_1$ ou $\boldsymbol{\pi}_2$. A grande vantagem de apresentar a Decomposição de Hoeffding de uma U-estatística, é que a os componentes são ortogonais como é o caso da equação (3.12). Esse artifício garante um pouco mais de facilidades no momento de calcular a variância da U-estatística. No entanto é importante lembrar aqui que a DB_n é uma combinação linear de U-estatísticas, não tendo mais a garantia de ortogonalidade em todos os termos. Podemos reescrever essa decomposição de forma a agrupar em relação aos correspondentes somatórios

$$\begin{aligned} DB_n =& \frac{1}{(n+1)}\Bigg[2\theta^{(1,2)} - \theta^{(1,1)} - \theta^{(2,2)} \\ &+ \left(\frac{2}{n_2}\right)\sum_{i=1}^{n_2}\left(\psi_1^{(2,1)}(\ \mathbf{X}_i^{(2)}) + \psi_1^{(2,1)}(\ \mathbf{X}^{(*)}) + \psi_2^{(2,1)}(x_i^{(2)}, x^{(*)})\right) \\ &- \left(\frac{2}{n_1}\right)\sum_{i=1}^{n_1}\left(\psi_1^{(1,1)}(x_i^{(1)}) + \psi_1^{(1,1)}(\ \mathbf{X}^{(*)}) + \psi_2^{(1,1)}(x_i^{(1)},\ \mathbf{X}^{(*)})\right) \\ &+ \left(\frac{2}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\left(\psi_1^{(1,1)}(\ \mathbf{X}_i^{(1)}) + \psi_1^{(1,1)}(\ \mathbf{X}_j^{(1)}) + \psi_2^{(1,1)}(x_i^{(1)},\ \mathbf{X}_j^{(1)})\right) \\ &- \left(\frac{2}{n_2(n_2-1)}\right)\sum_{1\le i<j\le n_2}\left(\psi_1^{(2,2)}(x_i^{(2)}) + \psi_1^{(2,2)}(x_j^{(2)}) + \psi_2^{(2,2)}(x_i^{(2)},\ \mathbf{X}_j^{(2)})\right)\Bigg]. \end{aligned} \tag{3.13}$$

3.6.1 Média e Variância da DB_n

A partir da Decomposição de Hoeffding e suas propriedades, a média da estatística DB_n pode ser diretamente obtida da expressão (3.13) sendo dada por

$$\mathbb{E}(DB_n) = \frac{1}{(n+1)}\left(2\theta^{(1,2)} - \theta^{(1,1)} - \theta^{(2,2)}\right). \tag{3.14}$$

Para o cálculo da variância da DB_n é importante notar que os termos dentro de cada somatório da expressão (3.13) são ortogonais. No entanto, entre os somatórios essa propriedade não é mais valida. Porém, a dependência ocorre somente nos termos de primeira ordem, os termos de segunda ordem são ortogonais também entre os somatórios.

Sob H_0 o novo elemento $\mathbf{X}^*$ pertence ao grupo 1. Vamos utilizar essa informação para calcular a variância da DB_n sob H_0. Para essa tarefa, iremos reescrever a DB_n de forma que os termos com alguma dependência fiquem no mesmo somatório. Assim,

$$\begin{aligned} DB_n = \frac{1}{(n+1)}\Bigg[& 2\theta^{(1,2)} - \theta^{(1,1)} - \theta^{(2,2)} \\ & + \left(\frac{2}{n_2}\right)\sum_{i=1}^{n_2}\left(\psi_1^{(2,1)}(x_i^{(2)}) + \psi_1^{(2,1)}(x^{(1)}) + \psi_2^{(2,1)}(x_i^{(2)}, x^{(1)})\right) \\ & - \left(\frac{2}{n_1}\right)\sum_{i=1}^{n_1}\left(\psi_1^{(1,1)}(x_i^{(1)}) + \psi_1^{(1,1)}(x^{(1)}) + \psi_2^{(1,1)}(x_i^{(1)}, x^{(1)})\right) \\ & + \left(\frac{2}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\left(\psi_1^{(1,1)}(x_i^{(1)}) + \psi_1^{(1,1)}(x_j^{(1)}) + \psi_2^{(1,1)}(x_i^{(1)}, x_j^{(1)})\right) \\ & - \left(\frac{2}{n_2(n_2-1)}\right)\sum_{1\le i<j\le n_2}\left(\psi_1^{(2,2)}(x_i^{(2)}) + \psi_1^{(2,2)}(x_j^{(2)}) + \psi_2^{(2,2)}(x_i^{(2)}, x_j^{(2)})\right)\Bigg]. \end{aligned} \tag{3.15}$$

De forma que,

$$\begin{aligned} DB_n = \frac{1}{(n+1)}\Bigg[& 2\theta^{(1,2)} - \theta^{(1,1)} - \theta^{(2,2)} \\ & + \left(\frac{2}{n_2}\right)\sum_{i=1}^{n_2}\left(\psi_1^{(2,1)}(x^{(1)}) + \psi_2^{(2,1)}(x_i^{(2)}, x^{(1)})\right) \\ & - \left(\frac{2}{n_1}\right)\sum_{i=1}^{n_1}\left(\psi_1^{(1,1)}(x^{(1)}) + \psi_2^{(1,1)}(x_i^{(1)}, x^{(1)})\right) \\ & + \left(\frac{2}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\left(\psi_1^{(1,1)}(x_i^{(1)}) + \psi_1^{(1,1)}(x_j^{(1)}) - \frac{2}{n_1}\psi_1^{(1,1)}(x_i^{(1)}) + \psi_2^{(1,1)}(x_i^{(1)}, x_j^{(1)})\right) \\ & - \left(\frac{2}{n_2(n_2-1)}\right)\sum_{1\le i<j\le n_2}\left(\psi_1^{(2,2)}(x_i^{(2)}) + \psi_1^{(2,2)}(x_j^{(2)}) - \frac{2}{n_2}\psi_1^{(2,1)}(x_i^{(2)}) + \psi_2^{(2,2)}(x_i^{(2)}, x_j^{(2)})\right)\Bigg]. \end{aligned} \tag{3.16}$$

E, finalmente

$$
\begin{aligned}
DB_n = \frac{1}{(n+1)}\Bigg[& 2\theta^{(1,2)} - \theta^{(1,1)} - \theta^{(2,2)} \\
& + 2\psi_1^{(2,1)}(x^{(1)}) + \left(\frac{2}{n_2}\right)\sum_{i=1}^{n_2}\psi_2^{(2,1)}(x_i^{(2)}, x^{(1)}) \\
& - 2\psi_1^{(1,1)}(x^{(1)}) - \left(\frac{2}{n_1}\right)\sum_{i=1}^{n_1}\psi_2^{(1,1)}(x_i^{(1)}, x^{(1)}) \\
& + \left(\frac{2}{n_1(n_1-1)}\right)\sum_{1\le i<j\le n_1}\left(\frac{(n_1-2)}{n_1}\psi_1^{(1,1)}(x_i^{(1)}) + \psi_1^{(1,1)}(x_j^{(1)}) + \psi_2^{(1,1)}(x_i^{(1)}, x_j^{(1)})\right) \\
& - \left(\frac{2}{n_2(n_2-1)}\right)\sum_{1\le i<j\le n_2}\left(\psi_1^{(2,2)}(x_i^{(2)}) + \psi_1^{(2,2)}(x_j^{(2)}) - \frac{2}{n_2}\psi_1^{(2,1)}(x_i^{(2)}) + \psi_2^{(2,2)}(x_i^{(2)}, x_j^{(2)})\right)\Bigg].
\end{aligned} \tag{3.17}
$$

Observe que somente os termos $\psi_1^{(2,1)}(x^{(1)})$ e $\psi_1^{(1,1)}(x^{(1)})$ são ortogonais, já os termos $\psi_1^{(2,2)}(x_i^{(2)})$ e $\psi_1^{(2,1)}(x_i^{(2)})$ não são ortogonais. Definindo os seguintes parâmetros de variâncias e covariâncias

$$
\begin{aligned}
\sigma_{(1)}^{2(g,g'/g)} &= \mathrm{Var}\left[\psi_1^{(g,g')}(x_i^{(g)})\right]; \\
\sigma_{(2)}^{2(g,g')} &= \mathrm{Var}\left[\psi_2^{(g,g')}(x_i^{(g)}, x_i^{(g')})\right]; \\
\rho^* &= \mathrm{Cov}(\psi_1^{(2,1)}(x^{(1)}), \psi_1^{(1,1)}(x^{(1)})); \\
\rho_1 &= \mathrm{Cov}(\psi_1^{(2,2)}(x_i^{(2)}), \psi_1^{(2,1)}(x_i^{(2)})),
\end{aligned}
$$

podemos obter uma expressão para a variância da U-estatística DB_n.

$$
\begin{aligned}
\mathrm{Var}(DB_n) = \frac{1}{(n+1)^2}\Bigg[& 4\sigma_{(1)}^{2(2,1/1)} + \frac{4}{n_2}\sigma_{(2)}^{2(2,1)} + 4\sigma_{(1)}^{2(1,1/1)} + \frac{4}{n_1}\sigma_{(2)}^{2(1,1)} \\
& + \frac{2}{n_1(n_1-1)}\left(\frac{2n_1^2 - 4n_1 + 4}{n_1^2}\sigma_{(1)}^{2(1,1/1)} + \sigma_{(2)}^{2(1,1)}\right) \\
& + \frac{2}{n_2(n_2-1)}\left(2\sigma_{(1)}^{2(2,2/2)} + \frac{4}{n_2^2}\sigma_{(1)}^{2(2,1/2)} + \sigma_{(2)}^{2(2,2)} - 8\rho^* + \frac{4}{n_2}\rho^1\right)\Bigg]
\end{aligned}
$$

Note que $\mathrm{Var}(DB_n) = O(n^2)$ o que mostra a consistência da DB_n para estima $\mathbb{E}(DB_n)$. O seguinte teorema aborda a distribuição da estatística de teste DB_n.

Teorema 3.6.1. Sejam $\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_{n_1}^{(1)}$ e $\mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_{n_2}^{(2)}$ vetores aleatórios independentes com distribuição π_1 e π_2, respectivamente. Seja $\mathbf{X}^*$ um novo elemento com distribuição desconhecida π_j, $j = 1, 2$. Considere a estatística DB_n definida na equação (3.10). Então,

$$
\frac{DB_n - \mathbb{E}(DB_n)}{\mathrm{Var}(DB_n)} \dot{\sim} N(0,1), \text{ quando } n \text{ e/ou } p \to \infty, \tag{3.18}
$$

em que $\dot{\sim}$ denota convergência assintótica em distribuição e $\mathbb{E}(DB_n)$ é dada em (3.14).

Prova: Embora tenha suporte na teoria de U-estatísticas, a normalidade assintótica da DB_n em (3.18) pode ser obtida do simples fato que que ela é a soma de duas variáveis aleatórias que são assintoticamente normais, como pode ser visto na definição (3.10). Além disso, a média e a variância da DB_n foram apresentadas nessa seção. □

3.6.2 Estimação da variância de DB_n

A variância de DB_n depende da escolha do $Kernel$ $\phi(\cdot,\cdot)$ e das distribuições $\mathscr{F}_1$ e $\mathscr{F}_2$. Para viabilizar nossa abordagem propomos um método baseado em um procedimento de reamostragem para obter uma estimativa da variância da estatística de teste DB_n. Esse procedimento consiste em guardar uma proporção de elementos do grupo 1 para utilizar cada um deles, na reamostragem, como um novo elemento $\mathbf{X}^*$, obtendo assim vários valores para a DB_n. Chamamos de $prop_1$ a proporção reservada para compor o grupo 1 reduzido.

A vantagem desse procedimento é a diminuição do viés de estimação da variância da DB_n ao possibilitar a observação de um número grande de valores da DB_n. A desvantagem é que os valores da DB_n são obtidos a partir de um número reduzido de elementos do grupo 1, uma vez que foram reservados os elementos desse grupo para posterior utilização como $\mathbf{X}^*$. Quando somente um elemento de cada vez é retirado do grupo 1 para ser o $\mathbf{X}^*$, então temos um procedimento conhecido como $leave$-one-out e indicaremos por $prop_1 = 1$. Nesse mesmo procedimento também é obtida uma estimativa da média da DB_n sob H_0.

Um estudo de simulação foi realizado para entender o efeito da escolha de $prop_1$ no viés e no erro padrão desse estimador empírico da variância da DB_n. Também foram considerados os efeitos de tamanho de grupo (n_1), dimensão (p) e a diferença d das médias entre as distribuições $\boldsymbol{\mathcal{F}}_1$ e $\boldsymbol{\mathcal{F}}_2$. Para esse estudo foi considerada a distribuição normal, sendo que os vetores do grupo 1 são compostos por p normais univariadas independentes com média zero e variância 1. Os vetores do grupo 2 são compostos por p normais univariadas independentes com média d e variância 1. Além disso, o tamanho n_2 do G_2 foi fixado em $n_2 = 20$.

A variância teórica da DB_n é difícil de ser obtida, como foi mostrado na Subseção 3.6.1, visto isso, realizamos uma simulação de Monte Carlo para estimar essa variância e usá-la como referencial de comparação para o método baseado em reamostragem acima descrito. Nesse procedimento de Monte Carlo, geramos os grupos G_1 e G_2 repetidas e para cada repetição, um elemento $\mathbf{X}^*$ com distribuição $\mathscr{F}_1$ é também gerado para o cálculo da DB_n. Nesse estudo, para cada n_1, p e d foram realizadas 200 repetições desse procedimento, possibilitando assim uma estimativa da variância populacional da DB_n.

As Tabelas 3.1 e 3.2 apresentam os resultados dessas simulações, nota-se que com o aumento do tamanho do grupo 1 (n_1) há uma diminuição tanto no viés quanto no erro padrão do estimador empírico da variância. Já se aumentar a dimensão (p), aumenta o erro padrão do estimador empírico, e para o viés nada podemos afirmar. A medida que se aumenta a diferença d das médias entre as distribuições $\boldsymbol{\mathcal{F}}_1$ e $\boldsymbol{\mathcal{F}}_2$ têm-se um aumento no viés e no erro padrão do estimador empírico da variância de DB_n. De modo geral o $leave$-one-out e a $prop_1 = 0.9$ possuem um menor viés e um menor erro padrão do estimador empírico da variância em comparação com as demais proporções $prop_1$. Com isso

que foi observado nessa simulação, é natural adotar o procedimento $leave\text{-}one\text{-}out$ como o estimador da variância de DB_n.

Tabela 3.1: Viés empírico do estimador da variância de DB_n de acordo com a proporção de observações $prop_1$ usada no procedimento de estimação.

			$prop_1$									
d	p	n_1	0.3	0.4	0.5	0.6	0.65	0.7	0.75	0.8	0.9	1
0.25	500	10	0.24	1.66	1.27	0.92	0.65	0.58	0.48	0.40	0.35	**-0.14**
		20	1.21	0.42	0.09	0.16	0.15	0.13	0.10	0.09	**0.07**	**-0.07**
		30	0.37	0.20	0.15	0.11	0.07	0.07	0.06	0.05	0.04	**-0.01**
		50	0.12	0.16	0.08	0.04	0.03	0.02	0.03	0.03	0.02	**-0.01**
	1000	10	-1.85	1.53	1.21	1.03	1.02	0.72	0.69	0.64	0.53	**-0.50**
		20	1.18	1.07	0.48	0.50	0.47	0.34	0.25	0.26	**0.17**	-0.18
		30	0.16	0.18	0.31	0.18	0.16	0.10	0.08	0.10	0.07	**0.02**
		50	0.18	**0.03**	0.14	0.09	0.06	0.08	0.05	**0.03**	0.04	**-0.03**
	2000	10	4.93	2.93	2.51	1.76	1.56	1.14	0.90	1.15	**0.79**	-1.45
		20	-0.15	0.32	**0.02**	0.31	0.39	0.35	0.29	0.23	0.17	-0.06
		30	0.33	**-0.01**	0.14	0.05	0.12	0.04	0.04	-0.02	-0.03	-0.14
		50	1.14	0.29	0.20	0.12	0.06	0.06	0.07	0.07	0.06	**-0.05**
0.50	500	10	6.47	4.60	2.45	1.52	1.40	1.13	1.12	0.92	0.80	**-0.27**
		20	3.01	1.21	0.90	0.76	0.68	0.53	0.43	0.42	0.29	**-0.05**
		30	1.58	0.66	0.61	0.44	0.33	0.30	0.24	0.20	0.16	**-0.03**
		50	-0.07	-0.08	**-0.00**	0.01	0.01	0.01	**-0.00**	0.02	0.01	-0.01
	1000	10	22.83	9.93	6.53	3.90	3.96	2.73	2.38	2.02	1.65	**-0.32**
		20	4.21	2.83	1.13	1.00	0.87	0.65	0.63	0.54	0.45	**-0.01**
		30	2.85	1.09	0.71	0.40	0.40	0.31	0.24	0.23	0.15	**-0.04**
		50	0.60	0.38	0.30	0.13	0.16	0.13	0.10	0.09	**0.06**	**-0.06**
	2000	10	24.67	2.33	1.71	1.98	0.97	0.63	-0.32	0.45	**0.11**	-1.22
		20	16.41	7.67	4.41	3.96	3.40	2.75	2.22	2.02	1.60	**-0.77**
		30	2.26	1.52	0.68	0.57	0.42	**0.17**	0.28	0.33	0.22	0.18
		50	0.34	0.97	0.42	0.50	0.43	0.29	0.38	0.26	0.21	**-0.06**
1.00	500	10	17.12	10.78	6.30	4.42	4.03	3.23	3.12	2.64	1.98	**0.01**
		20	1.30	1.74	1.40	0.81	0.40	0.39	0.53	0.30	0.31	**-0.23**
		30	3.98	2.41	1.71	1.23	1.04	0.74	0.71	0.62	0.55	**-0.07**
		50	2.08	1.31	0.60	0.36	0.43	0.27	0.27	0.26	0.15	**0.08**
	1000	10	55.05	30.29	21.11	14.99	12.52	11.27	10.44	8.51	6.73	**-0.83**
		20	16.90	8.53	5.37	3.61	3.70	2.39	2.49	2.22	1.63	**-0.57**
		30	5.38	1.52	2.24	1.52	1.18	1.09	0.92	0.77	0.43	**-0.01**
		50	2.69	2.14	1.01	0.67	0.58	0.59	0.45	0.33	0.29	**-0.12**
	2000	10	58.62	38.88	18.57	16.27	14.43	11.56	9.75	8.65	7.02	**-0.51**
		20	22.32	20.19	12.62	6.59	6.07	5.49	4.78	4.80	3.45	**-1.01**
		30	8.89	8.93	5.11	3.21	3.39	2.79	2.11	1.58	1.44	**-0.25**
		50	6.33	2.77	1.90	1.06	0.70	0.83	0.55	0.72	0.47	**0.01**

Tabela 3.2: Erro padrão do estimador da variância de DB_n de acordo com a proporção $prop_1$ de observações usada no procedimento de estimação.

			$prop_1$									
d	p	n_1	0.3	0.4	0.5	0.6	0.65	0.7	0.75	0.8	0.9	1
0.25	500	10	20.97	7.55	3.22	2.13	1.87	1.57	1.38	1.16	**0.87**	0.93
		20	4.62	1.95	1.43	0.92	0.74	0.62	0.53	0.46	0.37	**0.32**
		30	1.80	0.89	0.50	0.35	0.30	0.26	0.22	0.17	**0.14**	0.16
		50	0.81	0.42	0.23	0.15	0.15	0.12	0.11	0.08	0.06	**0.05**
	1000	10	39.86	14.10	8.84	5.49	5.05	4.22	3.56	2.89	2.35	**2.31**
		20	8.41	3.76	2.14	1.53	1.33	1.10	1.05	0.83	**0.66**	0.67
		30	4.00	1.73	1.08	0.76	0.63	0.54	0.47	0.42	0.31	**0.27**
		50	1.68	0.78	0.44	0.30	0.27	0.21	0.21	0.17	0.12	**0.11**
	2000	10	77.62	29.75	14.97	10.02	9.12	6.85	6.60	5.27	**4.15**	4.45
		20	16.79	8.14	4.44	3.25	2.63	2.25	2.04	1.75	1.41	**1.15**
		30	8.91	4.25	2.48	1.64	1.43	1.22	1.07	0.91	0.79	**0.66**
		50	2.92	1.75	0.96	0.64	0.54	0.47	0.36	0.34	0.26	**0.22**
0.50	500	10	23.06	9.60	6.12	4.24	3.75	3.10	2.85	2.34	1.78	**1.65**
		20	7.03	3.52	2.33	1.47	1.31	1.21	1.00	0.84	**0.67**	**0.67**
		30	3.77	1.71	1.11	0.73	0.64	0.49	0.48	0.39	**0.31**	0.32
		50	1.64	0.91	0.57	0.39	0.31	0.26	0.25	0.20	0.15	**0.13**
	1000	10	45.78	21.00	11.55	7.83	7.14	5.80	5.57	4.62	3.58	**3.55**
		20	17.13	7.73	5.39	3.94	2.98	2.63	2.29	2.01	1.56	**1.39**
		30	7.72	3.61	2.55	1.64	1.37	1.19	1.09	0.89	0.66	**0.62**
		50	2.84	1.62	1.09	0.70	0.62	0.51	0.49	0.39	0.32	**0.26**
	2000	10	109.58	49.06	27.95	18.41	16.89	13.55	13.53	10.50	8.63	**7.15**
		20	30.54	16.41	10.66	6.92	5.51	4.91	4.31	3.81	3.01	**2.33**
		30	15.59	8.36	4.96	3.37	2.83	2.41	2.04	1.92	1.37	**1.12**
		50	5.80	2.82	1.99	1.38	1.06	0.88	0.84	0.71	**0.52**	0.55
1.00	500	10	51.72	27.27	17.14	12.19	11.10	9.25	8.01	6.92	5.73	**5.29**
		20	22.76	11.56	7.35	4.95	4.59	3.85	3.36	2.76	2.29	**1.82**
		30	10.78	6.05	3.52	2.58	2.23	1.69	1.56	1.33	1.09	**1.04**
		50	4.98	2.94	1.84	1.31	1.09	0.87	0.73	0.67	0.56	**0.46**
	1000	10	86.95	45.69	28.43	19.95	17.98	14.23	12.34	10.57	**8.47**	10.50
		20	39.77	21.41	14.96	10.05	7.57	7.21	6.05	5.32	4.48	**4.35**
		30	21.64	13.68	7.55	5.11	4.71	3.80	3.24	2.85	2.44	**1.91**
		50	8.70	4.58	3.13	2.38	1.92	1.43	1.50	1.21	0.99	**0.80**
	2000	10	219.84	111.64	76.48	48.18	46.15	37.78	33.73	29.03	22.50	**21.93**
		20	83.98	47.40	28.12	20.21	17.09	14.10	12.83	11.36	8.57	**8.12**
		30	45.07	21.97	14.29	9.71	7.87	7.12	6.18	5.50	4.27	**3.83**
		50	16.44	10.35	5.49	4.49	3.72	3.19	2.85	2.27	2.12	**1.66**

3.7 *O classificador* UC

Considerando $\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_{n_1}^{(1)}$ amostras aleatórias independentes do grupo 1 com distribuição p-dimensional $\mathscr{F}_1$ e $\mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_{n_2}^{(2)}$ amostras aleatórias independentes do grupo 2 com distribuição p-dimensional $\mathscr{F}_2$. Lembrando que estamos considerando $p >> n_1, n_2$, i.e. o contexto $(HDLSS)$. Seja $\mathbf{X}^*$ uma nova observação com distribuição desconhecida $\mathscr{F}_g$, $g = 1, 2$. A decisão sobre em qual grupo $\mathbf{X}^*$ deve ser classificado é tomada baseado no cálculo da estatística DB_n definida na equação (3.10). A regra de decisão consiste no fato de que se a distribuição de $\mathbf{X}^*$ é $\mathscr{F}_g$, então é esperado que $B_n^{(g')}$ seja maior do que $B_n^{(g)}$, $g \neq g' \in \{1, 2\}$. Assim,

$$\text{classifica } \mathbf{X}^* \text{ em } G_{g'} \text{ se } B_n^{(g')} > B_n^{(g)}; \text{ caso contrário classifica em } G_g.$$

Denotamos esse classificador por UC e comparamos seu desempenho com classificador AH proposto por Ahmad and Pavlenko (2018) em cenários com diferentes distribuições dos dados e com distintas estruturas de correlações. Mas, antes de apresentar os resultados desses estudos de simulações, vamos apresentar a nossa interpretação para o problema de inferência em classificação.

Capítulo 4

Inferência em classificação

O motivo para propor um teste de classificação é que mesmo conhecendo-se as propriedades de um classificador, como a probabilidade de cometer um erro de classificação, ainda é possível dizer se uma classificação é estatisticamente significativa. Isso pode ser feito utilizando as propriedades da estatística associada ao classificador e uma proposição adequada das hipóteses estatísticas.

Neste contexto se faz necessário apresentar o problema de inferência em classificação. Embora estejamos trabalhando com dados p-dimensionais, tentaremos explicar o problema a partir de dados com dimensão 1. Sendo assim, considere dados advindos de duas distribuições normais. Para o exemplo apresentado na Figura 4.1 usamos $F_1 : N(0, 1)$ e $F_2 : N(5, 1.5)$ para representar as distribuições do grupo 1 e do grupo 2.

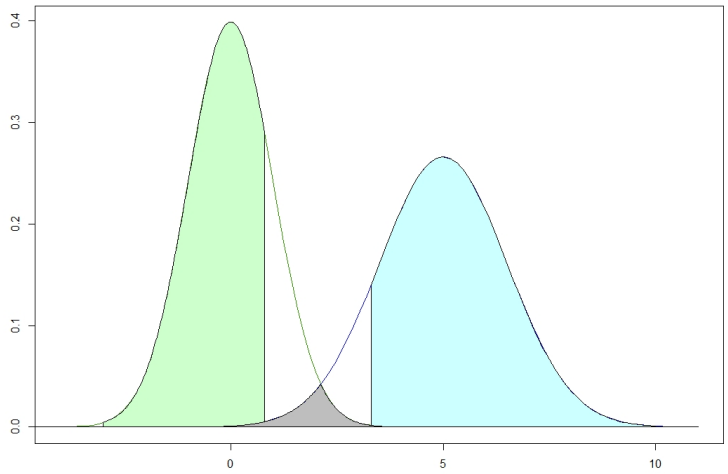

Figura 4.1: Densidade de duas distribuições normais e sua região de intersecção.

É esperado que pontos da distribuição 1 que estejam na região verde, sejam classificados como sendo da distribuição 1. Da mesma forma, pontos que estejam na região azul, deveriam ser classificados como sendo da distribuição 2. No entanto, pontos da região cinza podem ser classificados tanto na distribuição 1 quanto na distribuição 2. Naturalmente classificações feitas para pontos na região verde ou azul deveriam ser mais confiáveis do que aquelas obtidas para pontos da região cinza. A ideia é que o teste na classificação forneça essa confiabilidade estatística desejada e a distribuição da estatística de

teste é fundamental neste contexto. Por exemplo, um ponto que produza uma estatística padronizada bem "alta"deveria indicar que a classificação é estatisticamente mais confiável que uma classificação na qual o score padronizado da estatística de teste seja "baixo".

4.1 Teste de hipóteses para classificação

Como definido anteriormente, o elemento $\mathbf{X}^*$ a ser classificado deve pertencer a uma das duas populações ($\boldsymbol{\pi}_1$ ou $\boldsymbol{\pi}_2$). Vamos proceder o teste de hipótese em duas etapas. Primeiramente vamos descobrir em qual distribuição é mais verossímil para $\mathbf{X}^*$, ou seja, vamos identificar o grupo ao qual $\mathbf{X}^*$ seja mais provável de pertencer. Usaremos a DB_n para tal tarefa, classificando o $\mathbf{X}^*$ em um dos grupos. A partir da classificação, definimos o grupo G_1, o grupo ao qual foi classificado o elemento $\mathbf{X}^*$, e G_2 o grupo complementar. A hipótese teste é a seguinte:

$$H_0 : \mathbf{X}^* \in G_2 \quad \text{versus} \quad H_1 : \mathbf{X}^* \in G_1.$$

Sob H_0, $\mathbb{E}(DB_n) < 0$. Assim, rejeitar H_0 significa que temos evidências suficiente para garantir, a um nível de significância α, que o novo elemento $\mathbf{X}^*$ está corretamente classificado no grupo G_1.

No trabalho de Ahmad and Pavlenko (2018) o classificador AH apresenta a propriedade de ser assintoticamente normal, como pode ser visto na Seção 3.1. Nesse caso, também poderia ser feito um teste de significância para a classificação. No entanto, a maneira como é apresentada a teoria, focando somente no erro de classificação, o qual depende das médias dos grupos 1 e 2 e das respectivas variâncias, limita a capacidade de utilização dessas ferramentas. Desse modo, qualquer classificação dada a pontos das regiões verde, cinza ou azul, estará associado um mesmo erro de classificação.

4.2 Inferência empírica do método UC

Na Seção acima utilizamos o modelo teórico unidimensional para exemplificar a metodologia proposta. No entanto, o contexto deste trabalho remete a dados de alta dimensão (p grande), com isso, realizamos um estudo controlado com dados simulados a partir de distribuições p-dimensionais e analisamos o comportamento do método de classificação e as propriedades empíricas da estatística de teste DB_n.

Para o estudo foi considerada a distribuição normal multivariada $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, em que $p = 500$. As $n_1 = 12$ amostras do grupo 1 foram geradas a partir de uma média $\boldsymbol{\mu}_1 = \mathbf{0}$ [1] e da variância $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_{p\times p}$. Já as $n_2 = 10$ amostras do grupo 2 foram geradas a partir de uma média $\boldsymbol{\mu}_2$ com $\lfloor p/3 \rfloor$ [2] elementos iguais a 0 e os demais iguais a 0.3. Uma vez geradas as amostras, realizamos um procedimento "leave-one-out" em cada um dos grupos. Nesse procedimento as amostras são retiradas uma a uma e o método é designado a classificá-las retornando também a significância desta classificação. Esse processo resulta em $n_1 + n_2$ valores para a estatística DB_n e também $n_1 + n_2$ p-valores para o teste de classificação. Replicamos esse processo 100 vezes e com isso as analises e

[1] $\boldsymbol{\mu}_1$ é um vetor p-dimensional cujas entradas são iguais a zero

[2] $\lfloor x \rfloor$ converte um número real x no maior número inteiro menor ou igual a x

gráficos foram feitos com base em $100(n_1 + n_2)$ valores da DB_n e os correspondentes p-valores.

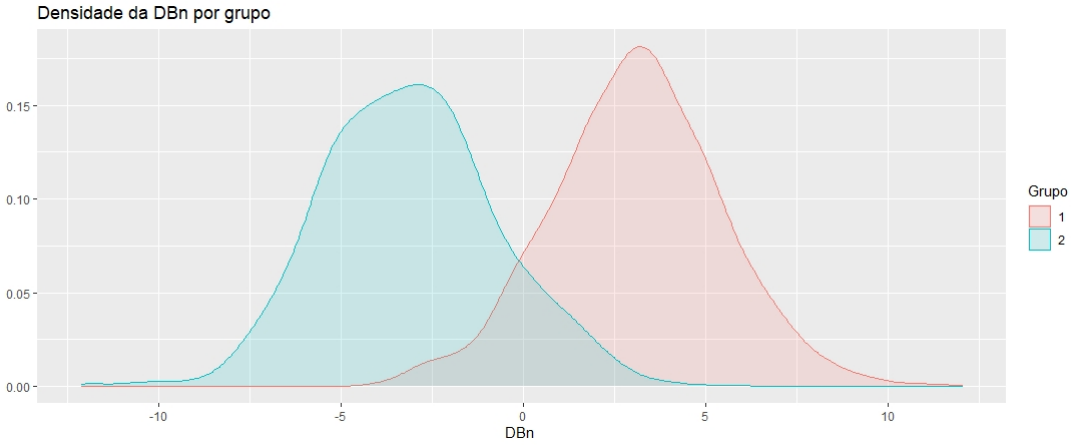

Figura 4.2: Densidade empírica da DB_n, separada por grupos de origem.

A Figura 4.2 mostra a densidade empírica da estatística DB_n separada por grupo de origem. Os valores da DB_n que resultaram na densidade em vermelho foram obtidos ao classificar amostras do grupo 1. Já a densidade em azul provém da classificação de amostras do grupo 2. Como era esperado, existe uma região de intersecção das duas densidades que corresponde aos valores da DB_n que estão próximos de 0.

A Figura 4.3 repete a Figura 4.2, com a diferença de que cada ponto marcado no eixo x corresponde a um valor da estatística DB_n obtida ao realizar uma classificação. No primeiro gráfico as cores estão de acordo com o grupo de origem do elemento que foi classificado. No segundo gráfico as cores estão de acordo com a classificação obtida pelo método UC. Note que, na região de intersecção, existe uma diferença entre o verdadeiro grupo de origem e o resultado obtido pelo método UC. Isso já era esperado e como argumentamos anteriormente a confiança na classificação obtida nessa região de intersecção não deve ser a mesma que nas regiões fora da intersecção.

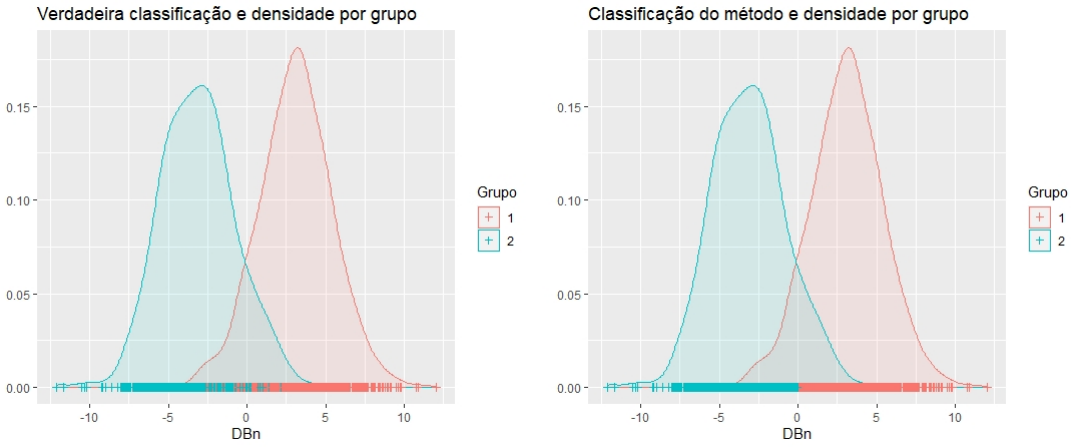

Figura 4.3: Densidade empírica da DB_n, separada por grupos de origem com a classificação verdadeira e a classificação do método UC.

Na Figura 4.4 exploramos as propriedades empíricas da estatística de teste. Além da densidade plotada para cada grupo, no eixo x os valores da DB_n, agora, são marcados (coloridos) de acordo com

sua significância estatística. No primeiro gráfico da Figura 4.4, observa-se que justamente na região

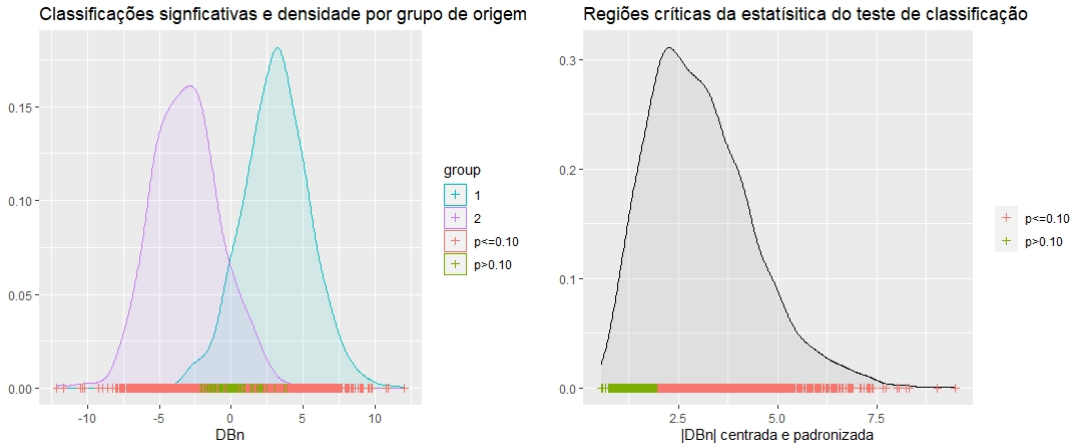

Figura 4.4: Densidade empírica da DB_n separada por grupos de origem, com indicação de classificações significativas e regiões críticas da estatística de teste do método de classificação.

de intersecção é que os resultados não deram todos significativos a um nível de significância $\alpha = 0.10$. Na direita, o gráfico mostra as regiões críticas da estatística DB_n, considerando grupos de origem, e com as classificações significativas obtidas a partir dos quantis($\alpha = 0.10$) de uma distribuição normal padrão. No segundo gráfico temos, de fato, a representação da densidade da estatística de teste sob H_1.

Como era esperado, valores absolutos grandes da estatística DB_n resultam em classificações significativas, enquanto valores menores, em módulo, não garantem uma classificação significativa. É importante destacar que esse exemplo prático corresponde exatamente ao que se esperava, conforme sugere a metodologia de inferência em classificação.

4.2.1 Considerando múltiplos grupos

Podemos também considerar múltiplos grupos. Para isso considere K grupos denotados por G_i, $i = 1, \ldots, K$. Assim como no trabalho de Ahmad and Pavlenko (2018), a regra de decisão é testar o elemento a ser classificado em todas as combinações possíveis dois a dois dos grupos, digamos $G_i \times G_j$ para $i \neq j \in \{1, \ldots, K\}$ e obter a estatística DB_n, para cada uma dessas combinações (podendo ser denotada por DB_n^{ij}). O arranjo que produzir a maior DB_n^{ij} entre todas as combinações dois a dois será utilizado para classificar esse novo elemento.

4.3 Estimação da probabilidade de erro de classificação

Para comparar o desempenho dos classificadores AH e UC precisamos definir medidas adequadas. Ahmad and Pavlenko (2018) define a probabilidade do erro de classificação (*misclassification error probability*) da seguinte forma. Para cada $g \in \{1, 2\}$, seja $\mathbf{X}_i^{(g)} = (X_{i1}^{(g)}, \cdots, X_{ip}^{(g)})^\mathsf{T} \sim \boldsymbol{\mathcal{F}}_g$ como definido anteriormente e $\boldsymbol{\pi}_g$ denota a g-ésima população desconhecida. Considere também $\mathcal{R}_g = \{\mathbf{X}^* : \mathbf{X}^* \in \boldsymbol{\pi}_g\}$ a região de dados observados da g-ésima população, em que $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathcal{X}, \mathcal{R}_1 \cap \mathcal{R}_2 =$

$\emptyset$ com $\mathcal{X}$ os espaço dos $\mathbf{X}^*$ observados e $\emptyset$ o conjunto vazio. Além disso, seja $\boldsymbol{\theta}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}$ o conjunto de parâmetros de $\boldsymbol{\mathcal{F}}_g$. Denotamos por $\pi(1 \mid 2)$ o erro de classificação de $\mathbf{X}^*$ em $\boldsymbol{\pi}_1$ quando de fato ele vem de $\boldsymbol{\pi}_2$. Formalmente,

$$\pi(g \mid g') = \mathrm{P}\left(\mathbf{X}^* \in \mathcal{R}_g \mid \mathbf{X}^* \in \boldsymbol{\pi}_{g'}\right) = \int_{\mathcal{R}_g} d\mathcal{F}_g(\mathbf{X}^* \mid \theta),$$

onde $\pi(g' \mid g) = 1 - \pi(g \mid g')$ é o complementar do erro de classificação incorreta para $g, g' \in \{1, 2\}, g \neq g'$. Dado que $\mathbb{E}\left(\overline{\mathbf{X}}_g\right) = \boldsymbol{\mu}_g$ e $\mathbb{E}\left(U_{n_g}\right) = \boldsymbol{\mu}_g^\top \boldsymbol{\mu}_g$ para todo $g \in \{1, 2\}$ e assumindo esses parâmetros conhecidos, primeiramente foi considerado o classificador $oracle$

$$A^{\text{oracle}}\ (\mathbf{X}^* \in \boldsymbol{\pi}_1) = (\mathbf{X}^*)^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/p - \left(\boldsymbol{\mu}_1^\top \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\mu}_2\right)/(2p).$$

Assumindo que $\boldsymbol{\mathcal{F}}_1$ e $\boldsymbol{\mathcal{F}}_2$ são normais multivariada, i.e., $\mathbf{X}_i^{(g)} \sim \mathcal{N}_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, com $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, então a chamada taxa de erro ótimo de A^{oracle} pode ser calculada da seguinte forma

$$\epsilon^{\text{oracle}} = \Phi\left(-\frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{2\sqrt{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}}^2}}\right),$$

onde Φ denota a função de distribuição normal padrão. O melhor desempenho possível nesta configuração de $oracle$, i.e., com $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ conhecidos, é obtido pelo classificador linear de Fisher (equivalente a regra de Bayes; ver Anderson (1958)), a saber

$$A^{\text{Fisher}}(\mathbf{X}^*) = (\mathbf{X}^*)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2,$$

com a correspondente taxa de erro de classificação dada por

$$\epsilon^{\text{Fisher}} = \Phi\left(-\frac{1}{2}\sqrt{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}}^2}\right),$$

onde $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}}^2$ é a distância de Mahalanobis. Denotando ϵ^{Fisher} como a referência, o desempenho relativo do $A^{\text{oracle}}(\mathbf{X}^*)$ pode ser avaliado teoricamente usando a seguinte expressão

$$q = \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_1^2}{\left\{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}}^2 \times \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2\right\}^{1/2}}$$

como argumento da função Φ.

Bickel and Levina (2004) propôs uma estratégia para calcular um limite para a expressão q, com base na desigualdade de Kantorovich (Bernstein (2009)). Seguindo a mesma ideia, seja $\mathbf{M}$ alguma matriz $p \times p$ simétrica positiva definida. Então para qualquer vetor $\boldsymbol{v}$

$$\frac{\|\boldsymbol{v}\|_{\mathrm{I}}^2}{\|\boldsymbol{v}\|_{\mathrm{M}}^2 \times \|\boldsymbol{v}\|_{\mathrm{M}^{-1}}^2} \geq \frac{4\lambda_{\min}(\mathbf{M}) \times \lambda_{\max}(\mathbf{M})}{\{\lambda_{\min}(\mathbf{M}) + \lambda_{\max}(\mathbf{M})\}^2},$$

onde $\lambda_{\min}(\mathbf{M})$ e $\lambda_{\max}(\mathbf{M})$ denotam o menor e o maior autovalor de $\mathbf{M}$, respectivamente. Aplicando esta desigualdade em q e denotando $\lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma}) = \kappa$ (assumindo ambos autovalores limitados e distantes de 0 e ∞), tem-se

$$q \geq 2\sqrt{\kappa}/(1 + \kappa), \tag{4.1}$$

de modo que o limite superior da probabilidade do erro de classificação de $A^{\text{oracle}}(\mathbf{X}^*)$ é

$$\epsilon^{\text{oracle}} \leq \Phi\left\{-\frac{2\sqrt{\kappa}}{1+\kappa}\Phi^{-1}\left(1-\epsilon^{\text{Fisher}}\right)\right\},$$

o qual depende essencialmente de κ, do intervalo de autovalores diferentes de zero de $\mathbf{\Sigma}$. Nota-se que, para valores moderados de κ, o aumento na taxa de classificação incorreta, induzida pela retirada da matriz de covariância ao construir $A^{\text{oracle}}(\mathbf{X}^*)$, não é grande em relação ao melhor desempenho possível, i.e., ϵ^{Fisher} ; ver Fig.4.5

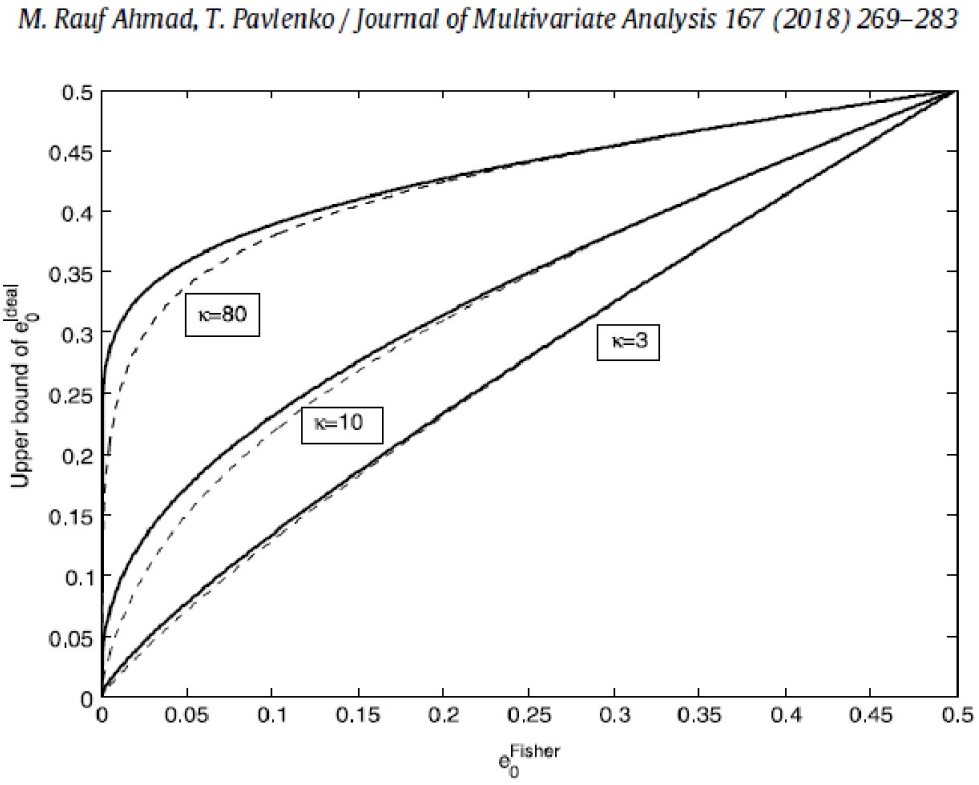

Figura 4.5: Limite superior da probabilidade do erro de classificação de $A^{(oracle)}(\mathbf{X}^*)$ em função de ϵ_0^{Fisher} o para distribuições normal (linha grossa) e t_5 (linha tracejada) com $k \in \{3, 10, 80\}$.

Além disso, o limite superior dada pela desigualdade (4.1) representa o pior cenário, de modo que se espera que os resultados empíricos sejam melhores.

Como os parâmetros são desconhecidos na prática, pode-se substituí-los pelas estimativas $\widehat{\theta_{g'}}$, conduzindo as regiões empíricas $\widehat{\mathcal{R}}_g$ a qual leva à taxa de erro real. Embora existam estimadores consistentes, como mostrado acima, a taxa real ainda não pode ser alcançada até que a forma das distribuições subjacentes sejam conhecidas. Como não se assume qualquer distribuição para $A(\mathbf{X}^*)$, recorre-se à medida mais utilizada, ou seja, a taxa de erro aparente, *apparent error rate* (APER) definido como

$$\text{APER} = \sum_{g=1}^{2} \pi_g \widehat{\pi}(g \mid g') = \frac{m_1 + m_2}{n_1 + n_2}$$

onde $\pi_g = \mathrm{P}(\mathbf{X}^* \in \mathcal{R}_g)$, $\widehat{\pi}(g \mid g') = m_g/n_g$ corresponde ao estimador de $\pi(g \mid g')$, m_g é o número de observações mal classificadas da população g na população g' e n_g é o tamanho da g-ésima amostra. Seguindo o procedimento padrão, combinamos APER com o procedimento de validação de

Lachenbruch (ou $leave\text{-}one\text{-}out$) nas amostras de treinamento e validação; ver, e.g., Dudoit et al. (2002).

Uma observação importante destacada por Johnson et al. (2014) é que o APER tende a subestimar a taxa de erro atual. Tal problema ocorre porque os dados para construir a função discriminante (classificação) são também utilizadas para avaliá-los. Ou seja, esse procedimento é consistente, mas viesado, e subestima os verdadeiros valores das probabilidades para elementos que não pertencem à amostra conjunta (novos elementos). Apesar disso, tal procedimento pode servir como uma etapa inicial de avaliação, pois se o valor de APER for muito elevado, é sinal que a regra de discriminação deve ser reformulada. O viés deste procedimento tende a zero quando os tamanhos amostrais são grandes.

Capítulo 5

Simulações

5.1 *Simulações para os classificadores* UC *e* AH

No trabalho de Ahmad and Pavlenko (2018) foram realizados alguns estudos de simulação para avaliar o desempenho do classificador AH em diferentes cenários, focando principalmente na consistência, normalidade assintótica e no controle do erro de classificação sob uma estrutura de dados correlacionados e de alta dimensão. Foram considerados os casos com dois grupos, $G = 2$, e gerados dados das distribuições normal multivariada e distribuição t-Student, ou seja, $\mathcal{F}_g$ ou é $\mathcal{N}_p\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$, $g \in \{1, 2\}$ ou é $t_v\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), v = 10$, em que $g \in \{1, 2\}$. Para cada distribuição, têm-se $\boldsymbol{\mu}_1 = \mathbf{0}$ e $\lfloor p/3 \rfloor$ elementos de $\boldsymbol{\mu}_2$ também iguais a zero, com o restante iguais a 1. Como as duas distribuições consideradas naquele estudo são simétricas, adicionamos um terceiro caso em que a distribuição é assimétrica. Para isso, consideramos um cenário em que temos uma mistura de duas normais. Consideramos as variáveis aleatórias independentes $\mathbf{X}_i^{(g)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_{xg}, \boldsymbol{\Sigma}_g\right)$ e $\mathbf{Y}_i^{(g)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_{yg}, \boldsymbol{\Sigma}_g\right)$, em que para $g = 1$, $\boldsymbol{\mu}_{x1}$ é um vetor p-dimensional com todas as entradas iguais a $-3/2$, enquanto $\boldsymbol{\mu}_{y1}$ é um vetor p-dimensional com todas as entradas iguais a $1/2$. Definimos então a variável aleatória $\mathbf{W}_i^{(g)} = 0.4\mathbf{X}_i^{(g)} + 0.6\mathbf{Y}_i^{(g)}$ para $i = 1, 2, \ldots, n_g$. Dessa forma, temos que $\mathbb{E}[\mathbf{W}_i^{(1)}] = \mathbf{0}$, e $\mathbf{W}_i^{(1)}$ têm a característica de ser assimétrica, centrada em zero. Para o grupo 2, repetimos $\boldsymbol{\mu}_{x1}$ e $\boldsymbol{\mu}_{y1}$ para as $\lfloor p/3 \rfloor$ coordenadas de $\boldsymbol{\mu}_{x2}$ e $\boldsymbol{\mu}_{y2}$ e somamos $+1$ nas demais $\lfloor 2p/3 \rfloor$ coordenadas.

Para $\boldsymbol{\Sigma}_g$, assim como em Ahmad and Pavlenko (2018), consideramos o caso em que ambas as populações têm estrutura $\mathrm{AR}(1)$, ou seja, $\mathrm{cov}\left(x_k, x_\ell\right) = \sigma^2\rho^{|k-\ell|}$ para todo $k, \ell \in \{1, \ldots, p\}$, com $\sigma^2 = 1$ para $g = 1, 2$. Para o grupo 1, foi utilizado $\rho = 0.3$ e para o grupo 2, $\rho = 0.7$, representando estruturas de baixa e alta correlação, respectivamente. Escrevendo a matriz de correlação completa para n observações consecutivas, $X_1, \ldots, X_n$, temos:

$$\begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{p-1} \\ \rho & 1 & \rho & \ldots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \ldots & 1 \end{pmatrix}.$$

Para o desempenho do classificador em amostras finitas, sob um cenário em que a dimensão é crescente, enfatizando que estamos interessados no cenário $p >> n_g$, extraímos amostras de tamanho

$n_1 = 5$, $n_2 = 7$, com $p \in \{10, 20, 50, 100, 300, 500, 700, 1000, 3000, 5000, 10000\}$. Os resultados são médias de 500 repetições executadas para cada combinação dos parâmetros mencionados acima. Além disso, foi observado o efeito de n_i grande, e também, são apresentadas as taxas de classificação incorreta para $n_1 = 10$, $n_2 = 12$. Bem como foi avaliado o classificador para diversos tamanhos de amostra, e.g., $n_1 = 5$, $n_2 = 25$ ou $n_1 = 10$, $n_2 = 50$, com resultados semelhantes, portanto, não relatados aqui.

As Figuras 5.1, 5.2 e 5.3 mostram o desempenho dos classificadores UC e AH através do erro de classificação, para diferentes dimensões p. Os gráficos a esquerda mostram as proporções de erros, onde elementos do grupo 1 foram incorretamente classificadas no grupo 2. Os gráficos da direita mostram as proporções de erros de classificação de elementos do grupo 2 que foram incorretamente classificados no grupo 1.

Na Figura 5.1 a distribuição considerada foi a normal multivariada com estrutura de correlação Autorregressiva. podemos observar um desempenho superior do método UC quando o objetivo é classificar elementos do grupo 1, enquanto o método AH tem desempenho superior quando o propósito é classificar elementos do grupo 2. Esse comportamento se repete nas Figuras 5.2 e 5.3 onde as distribuições consideradas foram a t-Student multivariada com $\nu = 10$ graus de liberdade e a mistura de normais respectivamente. Neste contexto ambos os métodos apresentaram desempenhos similares, sendo que $e(1|2)$ é menor para o método UC e $e(2|1)$ é menor para o método AH.

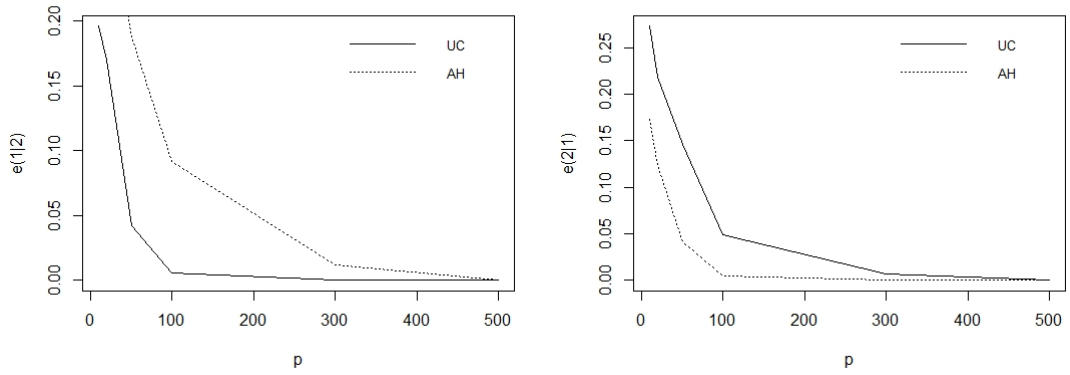

Figura 5.1: Erro de classificação para dados com distribuição Normal com $n_1 = 5$ e $n_2 = 7$.

5.2 APER em grupos desbalanceados

Como podemos observar nos resultados anteriores, os dois classificadores parecem ter desempenhos diferentes em relação aos erros de classificação $e(1/2)$ e $e(2/1)$. Além disso, observamos um efeito dos tamanhos dos grupos nos desempenhos dos métodos. Para entender melhor esse efeito, vamos considerar a medida APER descrita na Seção 4.3, a qual permite uma comparação mais direta para o entendimento desse efeito.

As Figuras 5.4 e 5.5 mostram a APER em casos em que os tamanhos de grupos são desbalanceados, sendo que em um cenário o grupo 1 é maior, e no outro cenário o grupo 2 é maior. Para o método

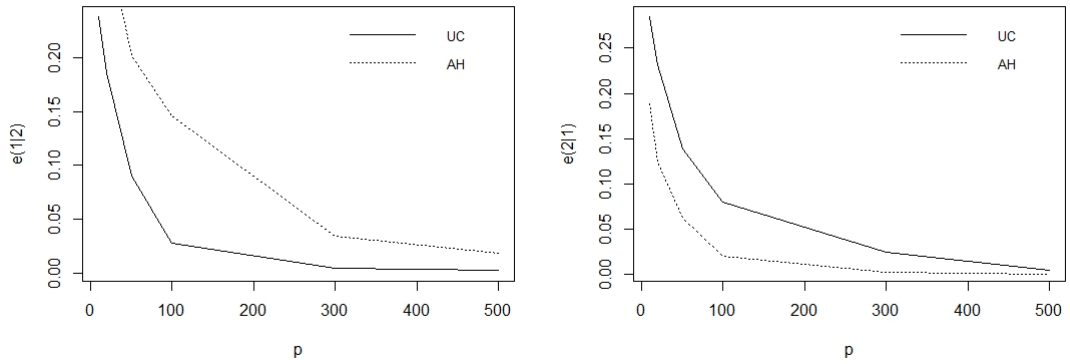

Figura 5.2: Erro de classificação para dados com distribuição t-Student com $n_1 = 5$ e $n_2 = 7$.

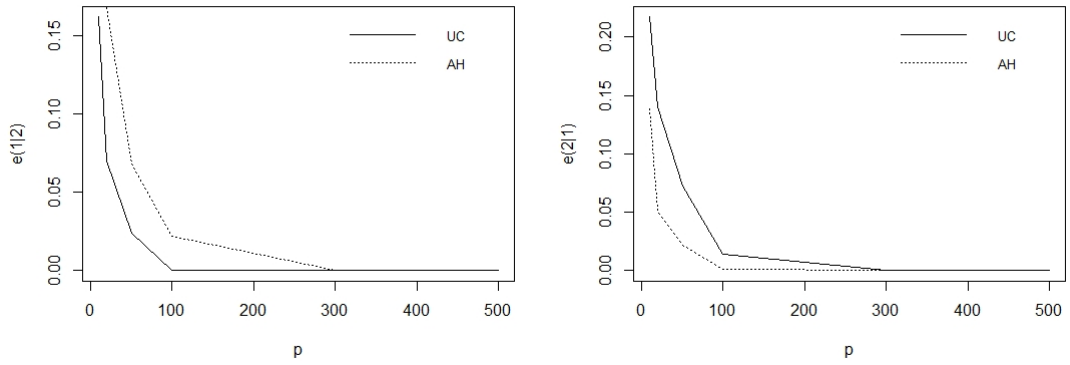

Figura 5.3: Erro de classificação para dados com mistura de normais com $n_1 = 5$ e $n_2 = 7$.

UC, a APER é menor quando o grupo 1 é maior comparado ao grupo 2, enquanto que, para o método AH, maior tamanho do grupo 1 está associado a maiores erros de classificação, quando comparada aos tamanhos de grupo 2, correspondentes.

Na Tabela 5.1 observamos o efeito dos tamanhos de grupos para o erro de classificação em ambos os métodos. Para o método UC, quando o grupo 1 é maior, menor é o APER, enquanto que, para o método AH, maior tamanho do grupo 1 corresponde a um maior erro de classificação, quando comparado ao tamanho do grupo 2. Nesse caso, o grupo 2 tem uma estrutura de correlação maior ($\rho = 0.3$ para o grupo 1 e $\rho = 0.7$ para o grupo 2) e, possivelmente, seja essa característica que afeta os classificadores de forma distinta.

O comportamento observado na Tabela 5.1 parece comum às outras distribuições (t-Student e mistura), por isso decidimos não apresentar os resultados dessa simulação.

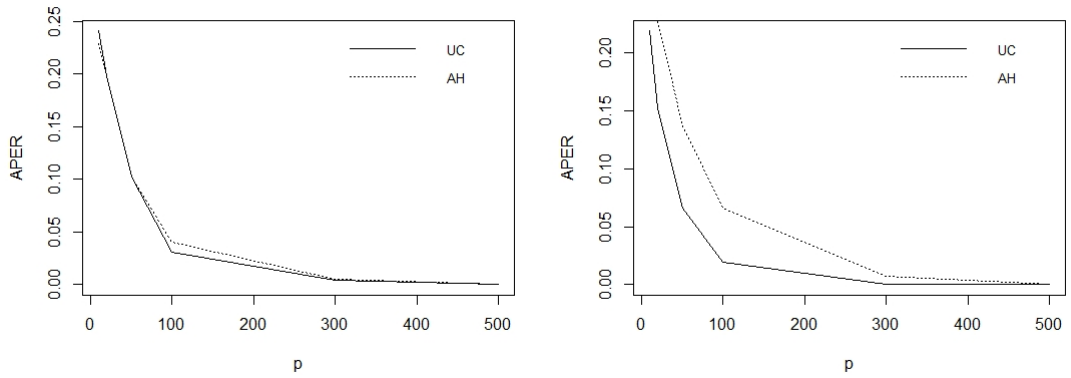

Figura 5.4: APER para dados normais, com $n_1 = 7$ e $n_2 = 5$, no primeiro gráfico e com $n_1 = 7$ e $n_2 = 20$, no segundo.

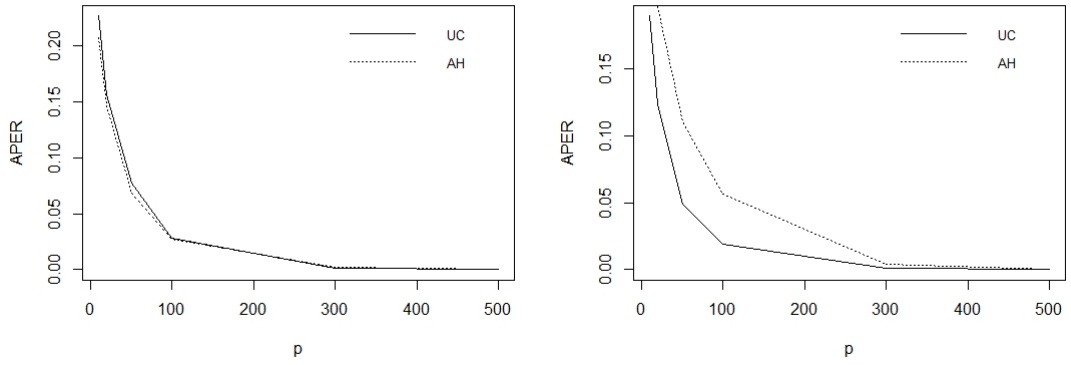

Figura 5.5: APER para dados normais, com $n_1 = 20$ e $n_2 = 40$, no primeiro gráfico e com $n_1 = 40$ e $n_2 = 20$, no segundo.

Tabela 5.1: APER dos métodos UC e AH para os tamanhos de grupos $n_1 = 20$, $n_2 = 40$ e $n_1 = 40$, $n_2 = 20$, para dados advindos da distribuição Normal.

		p					
Método	$n_1 \times n_2$	10	20	50	100	300	500
UC	20×40	0.226	0.152	0.078	0.028	0.001	0.000
	40×20	0.190	0.124	0.049	0.019	0.001	0.000
AH	20×40	0.207	0.146	0.068	0.027	0.002	0.000
	40×20	0.254	0.197	0.110	0.056	0.004	0.000

5.3 APER em grupos balanceados

Nos estudos de simulações anteriores notamos um efeito dos tamanhos dos grupos no erro de classificação. Esse efeito é observado de forma distinta para os métodos UC e AH, sendo que ao considerar

a medida $e(1/2)$, temos um erro de classificação menor para UC e quando consideramos a medida $e(2/1)$. Um efeito similar ocorre quando consideramos a medida APER em grupos desbalanceados, sendo difícil decidir qual é o melhor método nesses casos. Outra questão interessante é qual método apresenta menor APER em grupos balanceados.

A Figura 5.6 mostra os resultados obtidos para APER, considerando sempre os mesmos cenários descritos no início dessa Seção, com distribuição dos dados advindos de uma Normal.

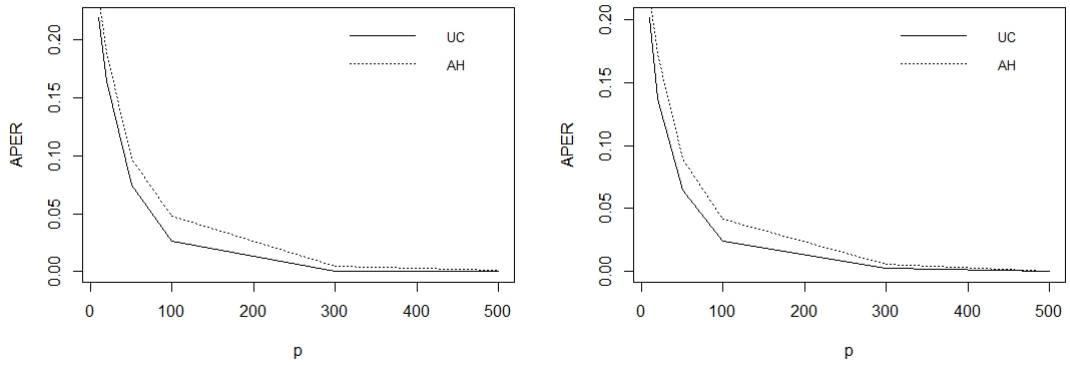

Figura 5.6: APER para dados normais e tamanhos de grupos balanceados, com $n_1 = 10$ e $n_2 = 10$, no primeiro gráfico e com $n_1 = 20$ e $n_2 = 20$, no segundo.

Podemos observar que, nesse caso em que os grupos tem tamanhos iguais, $n_1 = 10$ e $n_2 = 10$ no primeiro gráfico da Figura 5.6 e $n_1 = 20$, $n_2 = 20$ no segundo gráfico, que o método UC apresentou resultados ligeiramente melhores que o método AH.

5.4 Simulação com séries temporais

Embora na Seção anterior tenhamos explorado dados com alguma estrutura de correlação no contexto de classificação, é natural pensar em modelos de séries temporais quando o assunto é dados correlacionados. Quando se trata de séries temporais estacionárias, além de mudanças nas médias podemos estar interessados em detectar mudanças em outras características nos dados como, por exemplo, a própria estrutura de correlação. No entanto, a aplicação de metodologias clássicas como o método AH tipicamente é feita diretamente nos dados "brutos", não sendo adequada para identificação de mudanças em algumas características dos dados como, por exemplo, a correlação. Nesses casos, é comum a utilização de metodologias baseadas em medidas de dissimilaridade (ver Costa et al. (2020) para uma breve revisão sobre o assunto). Para mudanças na média, geralmente a distância euclidiana é a primeira opção, já para mudanças na estrutura de correlação a distância baseada na função de autocorrelação é comumente usada. Essa medida de dissimilaridade é descrita em Montero and Vilar (2015) da forma

$$d_{ACF}\left(\boldsymbol{X}_T, \boldsymbol{Y}_T\right) = \sqrt{\sum_{i=1}^{L}\left(\hat{\rho}_{i,X_T} - \hat{\rho}_{i,Y_T}\right)^2}, \tag{5.1}$$

em que $\hat{\boldsymbol{\rho}}_{X_T} = \left(\hat{\rho}_{1,X_T}, .., \hat{\rho}_{L,X_T}\right)^\top$ e $\hat{\rho}_{Y_T} = \left(\hat{\rho}_{1,Y_T}, .., \hat{\rho}_{L,Y_T}\right)^\top$ são vetores de autocorrelações estimadas das séries temporais X_T e Y_T respectivamente, para algum L tal que $\hat{\rho}_{i,X_T} \approx 0$ e $\hat{\rho}_{i,Y_T} \approx 0$ para $i > L$.

Todo o arcabouço teórico no qual nossa metodologia UC foi baseada permite tanto a utilização de dados brutos quanto a utilização de matrizes de dissimilaridade. No obstante, o método AH não tem essa característica. No entanto, a d_{ACF} em (5.1) nada mais é do que a distância euclidiana entre vetores de autocorrelação. Para efeitos de comparação vamos fornecer ao método AH as autocorrelações em vez dos dados, propriamente. Para ambos os métodos foram usadas as 100 primeiras autocorrelações, uma vez que nesse modelo somente as primeiras autocorrelações são diferentes de zero.

O APER foi obtido usando o mesmo procedimento das demais simulações, sendo que o grupo 2 foi gerado a partir de um modelo AR(1) com coeficiente $\phi = 0.3$. Além disso o tamanho do grupo 2 é fixo, com $n_2 = 20$. As séries do grupo 1 foram geradas a partir do modelo AR(1) com coeficiente ϕ tomando valores no conjunto Coef $= \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, sendo que também foram variadas as dimensões $p \in \{500, 1000, 2000\}$ e o tamanho do grupo 1, $n_1 \in \{10, 20, 30, 50\}$.

O modelo AR(1) com coeficiente $\phi = 0.3$ foi usado para gerar as $n_2 = 20$ séries do grupo 2. As séries a serem classificadas também foram geradas a partir desse modelo. Foram realizadas 100 replicações sendo que em cada uma dessas replicações, 30 séries temporais foram classificadas.

As séries do grupo 1 foram geradas a partir do modelo AR(1) com coeficiente ϕ tomando valores no conjunto Coef $= \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, sendo que também foram variadas as dimensões $p \in \{500, 1000, 2000\}$ e o tamanho do grupo 1, $n_1 \in \{10, 20, 30, 50\}$.

Nas Figuras 5.7, 5.8, 5.9 e na Tabela 5.2 reportamos a proporção dos erros de classificação para os diferentes cenários explorados. É importante destacar que nas linhas correspondentes ao Coef=0.3

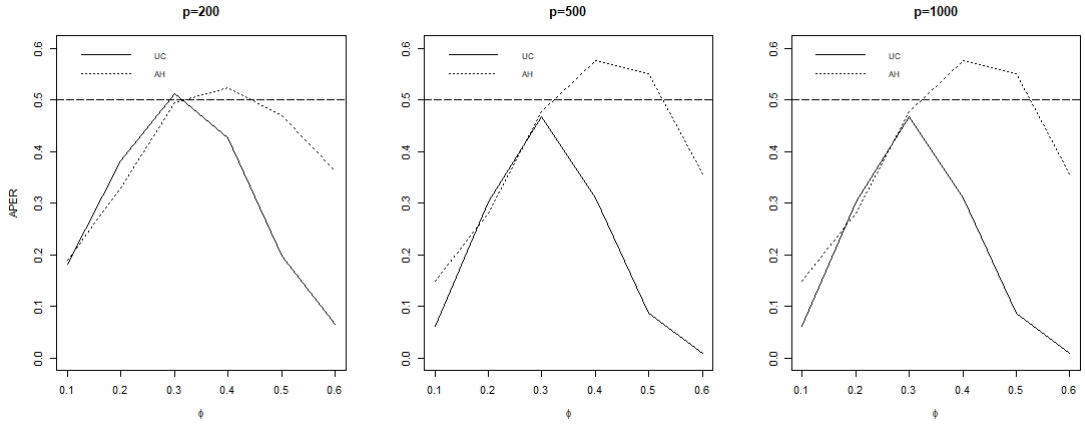

Figura 5.7: APER para dados advindo do modelo AR(1) com $n_1 = 10$ e $n_2 = 20$

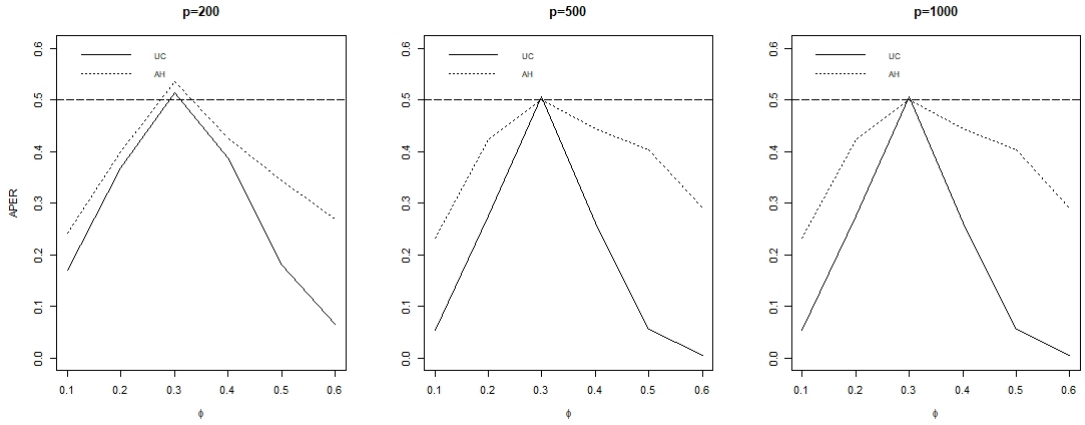

Figura 5.8: APER para dados advindo do modelo AR(1) com $n_1 = 20$ e $n_2 = 20$

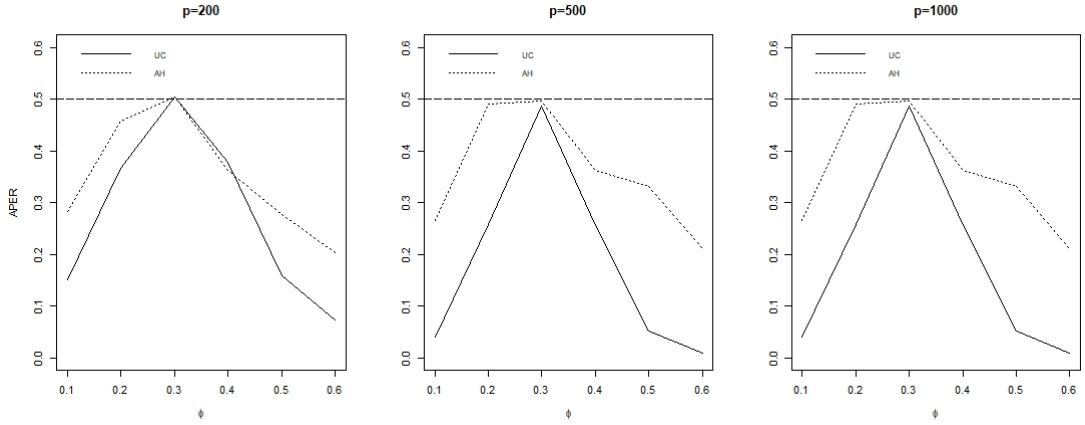

Figura 5.9: APER para dados advindo do modelo AR(1) com $n_1 = 30$ e $n_2 = 20$

temos uma situação incomum, visto que, neste caso os elementos do grupo 1 e os elementos do grupo 2 vêm de uma mesma distribuição. Esperamos que neste caso o erro de classificação seja próximo de

Tabela 5.2: Proporção do erro de classificação (APER) dos métodos UC e AH para dados do modelo AR(1) utilizando ACFs.

		10		20		30		50	
Coef	p	UC	AH	UC	AH	UC	AH	UC	AH
0.10	500	0.181	0.189	0.170	0.242	0.150	0.283	0.150	0.341
	1000	0.061	0.149	0.053	0.232	0.039	0.266	0.035	0.327
	2000	0.006	0.143	0.007	0.214	0.004	0.254	0.002	0.303
0.20	500	0.381	0.330	0.369	0.399	0.367	0.457	0.358	0.513
	1000	0.303	0.281	0.275	0.424	0.258	0.491	0.229	0.574
	2000	0.185	0.304	0.144	0.450	0.138	0.542	0.133	0.661
0.30	500	0.512	0.495	0.513	0.535	0.505	0.505	0.505	0.500
	1000	0.467	0.477	0.505	0.500	0.486	0.497	0.503	0.502
	2000	0.507	0.511	0.486	0.490	0.518	0.508	0.501	0.485
0.4	500	0.426	0.525	0.388	0.428	0.378	0.362	0.345	0.275
	1000	0.311	0.575	0.262	0.445	0.256	0.363	0.262	0.260
	2000	0.208	0.649	0.154	0.484	0.150	0.388	0.137	0.280
0.50	500	0.198	0.469	0.180	0.344	0.158	0.276	0.168	0.201
	1000	0.087	0.551	0.054	0.403	0.052	0.331	0.050	0.235
	2000	0.009	0.589	0.008	0.452	0.006	0.367	0.009	0.262
0.60	500	0.066	0.363	0.064	0.270	0.072	0.203	0.067	0.153
	1000	0.009	0.356	0.005	0.290	0.009	0.212	0.009	0.157
	2000	0.000	0.390	0.000	0.286	0.000	0.228	0.000	0.169

50%. Ambos os métodos apresentam comportamento dentro do esperado, apresentando baixo erro de classificação a medida que os grupos ficam mais separados em termos de distribuição (Coef $\neq$ 0.3) e errando proporcionalmente a classificação quando os grupos têm a mesma distribuição(Coef=0.3). No entanto, o método UC supera o método AH em termos de apresentar o menor APER, nesse cenário.

O que aconteceria se utilizássemos os próprios dados simulados do modelo AR(1)? A Tabela 5.3 apresenta uma breve simulação em que podemos notar claramente a incapacidade de classificação de ambos os métodos. Consideramos somente o caso em que os elementos do grupo 2 tem coeficiente autorregressivo 0.3, enquanto que no grupo 1 esse coeficiente Coef=0.5. Comparando-se as Tabelas 5.2 e 5.3, para as linha correspondentes ao Coef=0.5 fica evidente a necessidade da utilização de medidas de dissimilaridades nesses casos. Notadamente o método UC é superior nessas situações.

Tabela 5.3: Proporção do erro de classificação (APER) dos métodos UC e AH para o modelo AR(1) utilizando os dados "brutos".

		10		20		30		50	
Coef	p	UC	AH	UC	AH	UC	AH	UC	AH
0.50	500	0.465	0.471	0.492	0.488	0.492	0.497	0.499	0.500
	1000	0.508	0.498	0.511	0.502	0.464	0.463	0.497	0.509
	2000	0.480	0.492	0.510	0.494	0.485	0.476	0.492	0.492

5.5 Cenário com dados normais iid

Na Tabela 5.4 estão reportados as proporções dos erros de classificação $e(1/2)$ dos métodos UC e AH, considerando $n_2 = 20$ e $n_1 \in \{10, 20, 30, 50\}$. Nesse caso, em termos de classificação para grupos com a mesma distribuição ($\mu_1 = \mu_2 = 0$) os resultados foram exatamente como o esperado, ou seja, erros de classificação próximos a 50%. No entanto, nota-se que o método AH tem desempenho inferior ao UC, apresentando maior erro $e(1/2)$.

Tabela 5.4: Proporção do erro de classificação $e(1/2)$ dos métodos UC e AH, com $n_2 = 20$ e $n_1 \in \{10, 20, 30, 50\}$.

		n_1							
		10		20		30		50	
$\mu_2 - \mu_1$	p	UC	AH	UC	AH	UC	AH	UC	AH
-0.3	500	0.027	0.152	0.011	0.121	0.006	0.118	0.005	0.109
-0.3	1000	0.002	0.077	0.001	0.055	0.000	0.049	0.000	0.040
-0.3	2000	0.000	0.017	0.000	0.010	0.000	0.009	0.000	0.006
-0.2	500	0.184	0.319	0.115	0.274	0.106	0.280	0.087	0.246
-0.2	1000	0.088	0.242	0.053	0.199	0.037	0.177	0.030	0.173
-0.2	2000	0.026	0.151	0.009	0.109	0.007	0.097	0.004	0.089
-0.1	500	0.400	0.445	0.361	0.422	0.363	0.430	0.358	0.438
-0.1	1000	0.349	0.417	0.329	0.416	0.319	0.420	0.284	0.392
-0.1	2000	0.298	0.389	0.257	0.366	0.237	0.354	0.233	0.358
0.0	500	0.498	0.487	0.508	0.508	0.484	0.486	0.491	0.501
0.0	1000	0.500	0.502	0.511	0.514	0.491	0.496	0.495	0.492
0.0	2000	0.498	0.501	0.490	0.499	0.497	0.499	0.501	0.490
0.1	500	0.395	0.434	0.348	0.419	0.375	0.427	0.350	0.409
0.1	1000	0.358	0.422	0.327	0.408	0.310	0.402	0.314	0.403
0.1	2000	0.288	0.369	0.253	0.365	0.244	0.361	0.212	0.341
0.2	500	0.173	0.309	0.129	0.282	0.106	0.262	0.089	0.251
0.2	1000	0.084	0.230	0.047	0.206	0.040	0.178	0.030	0.161
0.2	2000	0.023	0.149	0.009	0.111	0.007	0.111	0.004	0.088
0.3	500	0.026	0.153	0.012	0.125	0.011	0.122	0.006	0.111
0.3	1000	0.003	0.074	0.001	0.046	0.000	0.042	0.000	0.036
0.3	2000	0.000	0.022	0.000	0.012	0.000	0.011	0.000	0.006

5.6 Inferência em classificação

A proporção de classificações significativas, ou o que podemos chamar de poder do teste de hipóteses para classificação está diretamente relacionado ao quanto as duas distribuições $\mathcal{F}_1$ e $\mathcal{F}_2$ são distintas. Quanto mais separação entre os grupos, menor será a interseção apresentada na Figura 4.2, e consequentemente, maior será o número de classificações significativas. Nas Tabelas 5.5 e 5.6 apresentamos a proporção de vezes que o teste u rejeitou a hipótese de homogeneidade dos grupos à 5% de significância ($p\text{-}homo$) e a proporção de classificações significativas à 5% de significância ($p\text{-}sig$).

Para a Tabela 5.5, o cenário descrito na Seção 5.5 foi utilizado no estudo de simulação.

Tabela 5.5: Proporção de vezes que o teste u rejeitou a homogeneidade dos grupos(p-$homo$) e proporção de classificações significativas (p-sig).

		n_1							
		10		20		30		50	
$\mu_2 - \mu_1$	p	p-$homo$	p-sig	p-$homo$	p-sig	p-$homo$	p-sig	p-$homo$	p-sig
-0.5	200	0.82	0.78	0.86	0.88	0.83	0.88	0.92	0.90
-0.5	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
-0.5	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.0	200	0.09	0.09	0.05	0.06	0.05	0.08	0.05	0.08
0.0	500	0.06	0.10	0.04	0.07	0.06	0.08	0.07	0.07
0.0	1000	0.06	0.09	0.05	0.08	0.07	0.07	0.08	0.08
0.1	200	0.08	0.07	0.10	0.06	0.07	0.11	0.07	0.07
0.1	500	0.09	0.11	0.17	0.15	0.11	0.10	0.07	0.12
0.1	1000	0.14	0.11	0.06	0.10	0.18	0.17	0.15	0.13
0.3	200	0.34	0.40	0.30	0.38	0.40	0.34	0.53	0.48
0.3	500	0.55	0.61	0.66	0.79	0.81	0.83	0.79	0.82
0.3	1000	0.87	0.79	0.88	0.84	0.93	0.94	0.96	1.00
0.5	200	0.86	0.88	0.82	0.84	0.94	0.90	0.89	0.90
0.5	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.0	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.0	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.0	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Podemos observar nos resultados da Tabela 5.5 que o poder do teste de classificação (p-sig) está diretamente relacionado à "distância"(neste caso, diferença nas médias) entre as distribuições, como era esperado. Além disso, o poder aumenta com a dimensão p de forma contundente. Já com o aumento dos grupos, embora possa se observar alguma tendência, não é possível identificar essa associação. Nas linhas associadas ao Coef=0, não temos separação dos grupos (mesma distribuição) e neste caso fica claro que tanto o p-$homo$ quanto o p-sig ficam em torno de 5% que seria o nível nominal dos testes.

Para visualizar melhor o funcionamento do teste de classificação, fizemos um estudo do comportamento da estatística DB_n, considerando dois cenários em que os dados são simulados com a mesma configuração da Tabela 5.5, porém com uma separação de grupos considerada pequena ($\mu_2 - \mu_1 = 0.1$) e uma separação de grupos maior de ($\mu_2 - \mu_1 = 0.3$). O grupo 1 tem tamanho $n_1 = 12$ e o grupo 2 tem tamanho $n_2 = 10$, com dimensão dos dados $p = 500$. Na Figura 5.10, no primeiro gráfico, a DB_n padronizada já aparece deslocada à direita, no entanto ainda um pouco centralizada em zero. Assim, ao utilizar os quantis da distribuição normal padrão, apenas 11% das observações foram classificadas de forma significativa. No segundo gráfico da Figura 5.10, a separação entre os grupos é maior, obtendo-se assim 78% das observações classificadas de forma significativa.

Realizamos também um estudo considerando dados correlacionados, com estrutura de séries temporais. Os dados foram gerados a partir de um modelo AR(1) com as mesmas configurações da

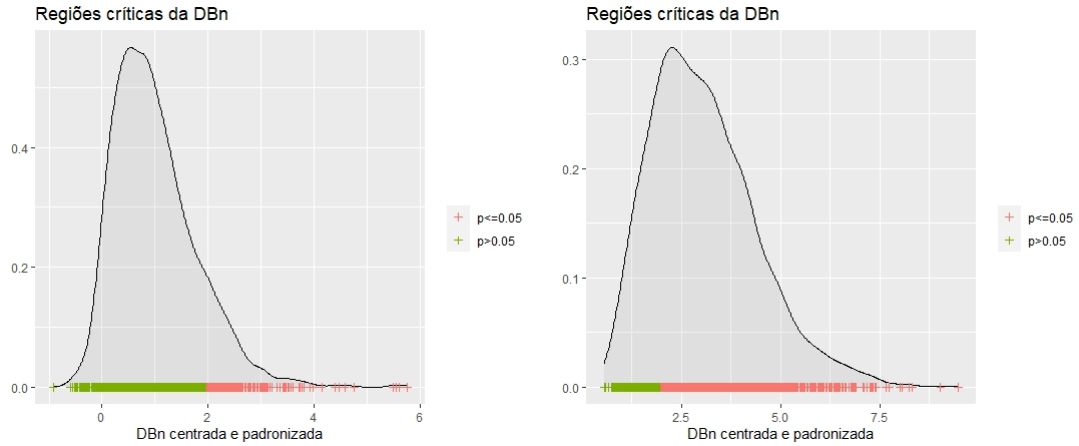

Figura 5.10: Região crítica da DB_n para uma separação pequena nos grupos, $\mu_2 - \mu_1 = 0.1$ e $\mu_2 - \mu_1 = 0.3$, com $\alpha = 0.05$.

simulação da Seção 5.4.

Tabela 5.6: Proporção de vezes que o teste u rejeitou a homogeneidade dos grupos(p-$homo$) e proporção de classificações significativas (p-sig) para dados de séries temporais, modelo AR(1) utilizando a distância baseada na ACF.

		n_1							
		10		20		30		50	
Coef	p	p-$homo$	p-sig	p-$homo$	p-sig	p-$homo$	p-sig	p-$homo$	p-sig
0.1	200	0.41	0.50	0.43	0.52	0.62	0.63	0.62	0.64
0.1	500	0.92	0.92	0.92	0.93	0.96	0.96	0.99	0.98
0.1	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	200	0.21	0.18	0.14	0.14	0.10	0.15	0.13	0.15
0.2	500	0.32	0.48	0.24	0.24	0.26	0.20	0.23	0.36
0.2	1000	0.52	0.64	0.54	0.66	0.63	0.64	0.56	0.62
0.3	200	0.09	0.09	0.06	0.10	0.05	0.06	0.08	0.07
0.3	500	0.12	0.10	0.06	0.08	0.08	0.07	0.06	0.07
0.3	1000	0.06	0.07	0.06	0.08	0.09	0.07	0.05	0.09
0.4	200	0.15	0.16	0.20	0.16	0.16	0.08	0.21	0.12
0.4	500	0.26	0.18	0.30	0.25	0.31	0.30	0.39	0.28
0.4	1000	0.46	0.42	0.60	0.57	0.67	0.74	0.71	0.64
0.5	200	0.43	0.33	0.54	0.37	0.68	0.44	0.62	0.43
0.5	500	0.85	0.78	0.94	0.92	0.94	0.86	0.93	0.93
0.5	1000	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00
0.6	200	0.91	0.82	0.93	0.84	0.94	0.88	0.93	0.92
0.6	500	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00
0.6	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Nos resultados da Tabela 5.6 podemos observar também um aumento significativo do poder do teste com o aumento da dimensão p dos dados. Já o tamanho dos grupos, não parece ser um fator tão relevante para o aumento do poder do teste de classificação, embora ainda seja possível observar

um leve melhora. Novamente, as linhas correspondentes ao Coef=0.3, os testes apresentam p-valores próximos ao nível nominal de 5%, o que é esperado visto que nesse caso os grupos apresentam mesma distribuição.

Capítulo 6

Aplicação

Apresentamos aqui uma aplicação a um conjunto de dados reais para avaliar e comparar o desempenho dos métodos UC e AH. Essa aplicação é similar à que foi considerada por Ahmad and Pavlenko (2018) e pode facilitar para efeitos de comparação das duas metodologias. Consideramos o conjunto de dados que trata de um estudo sobre *The Diffuse Large B-cell Lymphoma* (DLBCL), cujos dados podem ser obtidos no sitio *Microarray Database*: *(https://github.com/ramhiser/datamicroarray/wiki/Shipp)*, os quais foram compilados de Shipp et al. (2002). Nesse estudo, os autores coletaram 6.817 medidas de expressão genética, vindas de dois grupos independentes com um total de 77 pacientes. Em um dos grupos, haviam $(n_1 = 58)$ pacientes portadores de *Diffuse Large B-cell Lymphoma* (DLBCLs), enquanto no outro grupo, $(n_2 = 19)$ pacientes eram portadores de *follicular lymphoma*, (FL).

Para a comparação dos métodos consideramos uma abordagem de validação cruzada, sendo que em um primeiro cenário (C1), utilizamos, aleatoriamente, 1/3 de cada um dos grupos para formar o conjunto de treino e as demais observações foram alocadas para o conjunto de teste. Assim, o conjunto de treino em C1, possui $n_1 = 20$ e $n_2 = 6$. Consideramos as medidas $e(1/2)$, $e(2/1)$ e APER para comparar o desempenho das duas metodologias. Como os conjuntos de treino e teste são formados de forma aleatória, replicamos esse processo 100 vezes e reportamos as médias das medidas $e(1/2)$, $e(2/1)$ e APER. Além disso, um segundo cenário (C2) foi considerado, sendo que 2/3 de cada grupo foram usados para formar o conjunto de treino. Neste caso, o conjunto de treino em C2 possui $n_1 = 12$ e $n_2 = 40$. Esse cenário é muito similar ao que foi apresentado em Ahmad and Pavlenko (2018), com a diferença de que replicamos 100 vezes, em vez de 3, como no referido trabalho.

Tabela 6.1: Desempenho dos métodos AH e UC considerando a média de 100 replicas das medidas $e(1/2)$, $e(2/1)$ e APER para os cenários em que $n_1 = 20$, $n_2 = 6$ e $n_1 = 40$, $n_2 = 12$.

	(n_1, n_2)		(n_1, n_2)	
	$(20, 6)$		$(40, 12)$	
	UC	AH	UC	AH
$e(1/2)$	0.24	0.18	0.23	0.17
$e(2/1)$	0.15	0.19	0.18	0.23
APER	0.22	0.19	0.21	0.18

Notamos que o método AH apresenta um APER de aproximadamente 19%. Essa tendência foi observada nas simulações apresentadas na Tabela 5.4. O método de classificação UC tem uma taxa de erro de classificação APER de aproximadamente 21%. No entanto, 56% das classificações realizadas

pelo método UC tiveram p-valor< 0.05 no teste de classificação. Isso mostra que embora os métodos consigam classificar de forma satisfatória, o poder para identificar classificações significativas é um pouco menor, indicando que se trata de um problema um tanto complexo.

Capítulo 7

Conclusão e discussão

Nesse trabalho propomos um método de classificação com características adequadas para trabalhar em dados de alta dimensionalidade e um teste de hipóteses para verificar a significância da classificação.

O desenvolvimento dessa metodologia consiste na aplicação da já consolidada abordagem utilizando a Estatística B_n de Valk and Cybis (2020) de forma repetida. Assumindo em um primeiro momento que a observação a ser classificada pertence ao grupo 1, calcula a B_n, depois que a observação a ser classificada pertence ao grupo 2, calcula novamente a B_n, então, a diferença dessas B_n's, chamada de DB_n é considerada. Valores grandes e positivos da DB_n indicam que a classificação adequada é no grupo 1, enquanto valores negativos da DB_n indicam uma classificação no grupo 2. Mostramos que a DB_n é de fato uma U-estatística e utilizamos algumas propriedades dessa teoria para calcular a média e a variância da DB_n, mostrando que esta estatística é um estimador consistente da $\mathbb{E}[DB_n]$. Também mostramos que a DB_n é assintoticamente normal.

O método de classificação com inferência UC, decorrente da aplicação da DB_n é comparado com o método AH de Ahmad and Pavlenko (2018) através de estudos de simulação e aplicação a dados reais. Em um primeiro momento repetimos os cenários estudados em Ahmad and Pavlenko (2018), os quais consideram distribuições multivariadas com diferentes estruturas de correlações, obtendo resultados muito similares aos que foram reportados no trabalho original, podendo observar um desempenho competitivo dos métodos nessa situação. Além disso, acrescentamos um cenário em que os dados provém de uma mistura de distribuições, garantindo assimetria. Novamente, os métodos apresentaram desempenhos similares.

A metodologia aqui proposta herda as características das abordagens baseadas na estatística B_n (ver Valk and Cybis (2020)), permitindo a utilização de medidas de dissimilaridade. Com isso foi realizado um estudo de simulação considerando dados advindo de conhecidos modelos de séries temporais, especificamente o modelo AR(1). Embora o método AH não tenha sido desenvolvido com este propósito, comparamos o seu desempenho com UC, uma vez que a medida de similaridade utilizada é a distância euclideana das autocovariâncias estimadas, ou seja, utilizamos as autocovariâncias no lugar dos dados no método AH. Nesse caso observamos um comportamento atípico no erro de classificação do método AH, indicando que possivelmente não seja adequado para esse contexto.

O teste de classificação é uma ferramente importante para situações em que os grupos não tem uma separação bem pronunciada, ou seja, embora se saiba com certeza que os dados vêm de duas distribuições distintas por natureza, isso não fica bem caracterizado nos procedimentos estatísticos

(teste u). Nesse caso, o teste pode ajudar a identificar quais observações podem ser classificadas com maior segurança devido ao fato de existir um p-valor associado á cada classificação.

Referências Bibliográficas

Ahmad, M.R., Pavlenko, T., 2018. A u-classifier for high-dimensional data under non-normality. Journal of Multivariate Analysis 167, 269–283.

Anderson, T.W., 1958. An introduction to multivariate statistical analysis. Technical Report.

Bernstein, D.S., 2009. Matrix mathematics. Princeton university press.

Bickel, P.J., Levina, E., 2004. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. Bernoulli 10, 989–1010.

Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16, 199–231.

Chen, G.K., Chi, E.C., Ranola, J.M.O., Lange, K., 2015. Convex clustering: an attractive alternative to herarchical clustering. PLoS Computational Biology 11, e1004228.

Clarke, B., Fokoue, E., Zhang, H.H., et al., 2009. Principles and theory for data mining and machine learning. Springer.

Corporation, R., Bellman, R., 1961. Adoptive control processes: A guided tour. University Press.

Costa, Y.M., Bertolini, D., Britto, A.S., Cavalcanti, G.D., Oliveira, L.E., 2020. The dissimilarity approach: a review. Artificial Intelligence Review 53, 2783–2808.

Cox, D., Hinkley, D., 1974. Theoretical statistics chapman and hall, london. See Also .

Cybis, G.B., Valk, M., Lopes, S.R., 2018. Clustering and classification problems in genetics through u-statistics. Journal of Statistical Computation and Simulation 88, 1882–1902.

Delua, J., 2021. Supervised vs. unsupervised learning: What's the difference. Artificial intelligence Retrieved 5, 2021.

Denker, M., 1985. Asymptotic distribution theory in nonparametric statistics. Springer.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association 97, 77–87.

Euan, C., Sun, Y., Ombao, H., et al., 2019. Coherence-based time series clustering for statistical inference and visualization of brain connectivity. Annals of Applied Statistics 13, 990–1015.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research 15, 3133–3181.

Fraser, D.A.S., 1956. Nonparametric methods in statistics. .

Ghimire, S., Wang, H., 2012. Classification of image pixels based on minimum distance and hypothesis testing. Computational Statistics & Data Analysis 56, 2273–2287.

Hall, P., Marron, J.S., Neeman, A., 2005. Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 427–444.

He, Z., Sheng, C., Liu, Y., Zou, Q., 2021. Instance-based classification through hypothesis testing. IEEE Access 9, 17485–17494.

Hennig, C., 2015. What are the true clusters? Pattern Recognition Letters 64, 53–62.

Hoeffding, W., 1948. A class of statistics with asymptotically normal distributions. Annals of Mathematical Statiatics 19, 293–325.

Hoeffding, W., 1961. The strong law of large numbers for u-statistics. Technical Report. North Carolina State University. Dept. of Statistics.

Johnson, R.A., Wichern, D.W., et al., 2014. Applied multivariate statistical analysis. volume 6. Pearson London, UK:.

Lee, A.J., 1990. U-statistics, volume 110 of statistics: Textbooks and monographs.

Lehmann, E.L., 1999. Elements of large-sample theory. Springer.

Li, J.J., Tong, X., 2020. Statistical hypothesis testing versus machine learning binary classification: Distinctions and guidelines. Patterns 1, 100115.

Liao, S.M., Akritas, M., 2007. Test-based classification: A linkage between classification and statistical testing. Statistics & probability letters 77, 1269–1281.

Modarres, R., 2014. On the interpoint distances of bernoulli vectors. Statistics & Probability Letters 84, 215–222.

Modarres, R., 2018. Multinomial interpoint distances. Statistical Papers 59, 341–360.

Montero, P., Vilar, J.A., 2015. Tsclust: An r package for time series clustering. Journal of Statistical Software 62, 1–43.

Motlagh, O., Berry, A., O'Neil, L., 2019. Clustering of residential electricity customers using load time series. Applied energy 237, 11–24.

Pinheiro, A., Sen, P.K., Pinheiro, H.P., 2009a. Decomposability of high-dimensional diversity measures: Quasi-u-statistics, martingales and nonstandard asymptotics. Journal of Multivariate Analysis .

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

Rohatgi, V., Md, A., 2001. Ehsanes sales. An Introduction to Probability and Statistics .

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. Genetic structure of human populations. Science 298, 2381–2385.

Searle, S., 1971. Linear models. Technical Report.

Sen, P., 1992. Introduction to hoeffding (1948) a class of statistics with asymptotically normal distribution, in: Breakthroughs in statistics. Springer, pp. 299–307.

Sen, P.K., 2006. Robust statistical inference for high-dimensional data models with application to genomics. Austrian journal of statistics 35, 197–214.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R., 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8, 68–74.

Valk, M., Cybis, G.B., 2020. U-statistical inference for hierarchical clustering. Journal of Computational and Graphical Statistics , 1–11.

Witten, D.M., Tibshirani, R., 2011. Penalized classification using fisher's linear discriminant. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 753–772.