Universidade Federal do Amazonas - UFAM

Instituto de Ciências Exatas - ICE

Programa de Pós-Graduação em Matemática - PPGM

ANÁLISE DISCRIMINATE VIA DISTRIBUIÇÕES PREDITIVAS APROXIMADAS POR ESTIMADORES POR FUNÇÃO NÚCLEO

Autor: Diego da Silva Souza

Manaus

2012

Universidade Federal do Amazonas - UFAM

Instituto de Ciências Exatas - ICE

Programa de Pós-Graduação em Matemática - PPGM

ANÁLISE DISCRIMINATE VIA DISTRIBUIÇÕES PREDITIVAS APROXIMADAS POR ESTIMADORES POR FUNÇÃO NÚCLEO

Autor: Diego da Silva Souza

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Matemática - PPGM, da Universidade Federal do Amazonas - UFAM, como parte dos requisitos exigidos para obtenção do t'tulo de Mestre em Matemática, na área de concentração em Estatística.

Orientador: *Prof. Dr. José Raimundo Gomez Perreira*

Coorientador: *Prof. Dr. Max Sousa de Lima*

Manaus

2012

DEDICATÓRIA

À minha família Vanja, José Edson, Vivi,
Victor, Vinícius, Gringo, Livy e Bebê,
por todo amor, alegria, dedicação,
amizade e suporte.

AGRADECIMENTOS

À Deus, primeiramente, que permitiu tantas coisas boas acontecerem na minha vida sem que eu mereça, e que usou todas essas pessoas para ajudar-me nesse caminho. A Ti devo minha vida e eterna gratidão!

Aos meus pais, José Edson e Vanja, por dedicarem suas vidas e vencer tantas dificuldades para que tudo isso pudesse acontecer. Nunca poderei agradecer o suficiente a vocês e nem muito menos listar tudo o que vocês fizeram por mim. Amo muito vocês e têm minha gratidão eterna.

Aos meus irmãos Victor e Vinícius, pela amizade e companheirismo, e pelos momentos felizes que dividimos.

À minha noiva, e futura esposa, Vivi por todo carinho e dedicação, por me apoiar e ajudar nos momentos mais difíceis desde a graduação até hoje, e me dar um motivo a mais de viver.

Aos meus queridos cães Gringo, Livy e Bebê pelo amor incondicional, amizade, lambidas e mordidas.

Aos meus orientadores Prof. José Raimundo e Prof. Max Sousa, pela excelente orientação, por acreditarem no meu trabalho, pela paciência e amizade.

Aos professores Ronaldo Dias e James Dean pela participação na banca de defesa, pelas suas críticas, sugestões e correções.

Aos meus professores Celso Rômulo, José Raimundo e José Cardoso, por me acompanharem desde a graduação, pelo incentivo à pós-graduação, e por dividir seus conhecimentos e experiências.

Aos meus amigos e irmãos Nelson e Carina pela amizade, pelas longas discussões sobre probabilidade e inferência, pelas brincadeiras e besteiras que falamos, pela força e companheirismo na hora de vencer todas as lutas que passamos juntos desde a graduação.

À todos os professores do Departamento de Estatística, em especial a Prof. Amazoneida.

Aos meus amigos da área de Matemática Marcos, Lauriano, Jeferson, Silvia e Adrian pela amizade, momentos de descontração e dúvidas esclarecidas.

À CAPES pelo apoio financeiro.

RESUMO

ANÁLISE DISCRIMINATE VIA DISTRIBUIÇÕES PREDITIVAS APROXIMADAS POR ESTIMADORES POR FUNÇÃO NÚCLEO

Reconhecimento e classificação de padrões são problemas importantes em uma variedade de áreas científicas, como biologia, psicologia, medicina, visão computacional e etc. Porém este problema não é de fácil solução quando a distribuição de probabilidade dos dados é totalmente desconhecida. Neste trabalho, combinamos o método de estimação de densidades por Função Núcleo com um enfoque Bayesiano e propomos uma nova abordagem para problemas de classificação usando uma Análise Discriminante via Distribuições Preditivas Aproximadas. Estudos de simulação e aplicação em conjuntos de dados reais bastante utilizados na literatura, foram conduzidos como forma de avaliação dos métodos propostos. Os resultados mostraram que a performace dos métodos propostos são competitivos, e em alguns casos significantemente melhor, com os métodos clássicos da literatura, Análise Discriminante Linear(ADL), Análise Discriminante Quadrática(ADQ) e Análise Discriminante Naive Bayes com distribuição Normal(NNBDA).

Palavras-chave: Análise de Discriminante, Densidade Preditiva, Estimador de Núcleo, Estimação Bayesiana.

ABSTRACT

DISCRIMINANT ANALISYS VIA PREDICTIVE DISTRIBUTION APPROXIMATED BY KERNEL ESTIMATOR

Pattern Recognition and Classification problems are importantes in a variety of science fields, such as biology, psychology, medice, computer vision and etc. However, the problem is not so easy to solve when the true probability distribution of data is unknown. In this work, we combine the Kernel density estimation method with a Bayesian approach and propose a new method for classification problems using Discriminant Analisys via Approximate Predictive Distribution. Simulation studies and application in data sets widely used in literature, were conducted as an assessment of the proposed methods. The results showed that the performace of the proposed methods are competitive, and in some cases significantly better, with classical methods of literature, Linear Discriminant Analisys(LDA), Quadratic Discriminant Analisys(QDA) and Naive Bayes Discriminant Analisys with Normal distribution(NNBDA).

Keywords: Discriminant Analisys, Predictive Densities, Kernel Density Estimator, Bayesian Estimation.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Reconhecimento, descrição e classificação de padrões são problemas importantes em uma variedade de áreas científicas, tais como biologia, psicologia, medicina, marketing, visão computacional, inteligência artificial, sensoriamento remoto, etc. Tipicamente, um padrão é representado por um vetor de características $\left(x_{i1}^{(\omega_g)}, \ldots, x_{ip}^{(\omega_g)}\right)'$ e a análise discriminante consiste em classificar um determinado padrão $\mathbf{x}^{novo}$ em uma de r categorias $\omega_1, \omega_2, ..., \omega_r$ com base em suas p características $x_1, x_2, \ldots, xp$. Existem várias abordagens para esta classificação: *Redes Neurais*(Bishop (1995); Hastie *et al.* (2009)), *Métodos Fuzzy*(Bezdek & Pal, 1992), *Métodos Estatísticos*(Hastie *et al.*, 2009).

Do ponto de vista estatístico, assumi-se que o vetor de características possui uma função densidade de probabilidade f_{ω_g} típica de sua classe. Um vetor $\mathbf{x}$ pertencente à classe ω_g é visto como uma observação aleatória gerada de χ de acordo com algum modelo probabilístico f_{ω_g} condicionada à classe ω_g. A partir da modelagem de χ é construído um sistema de reconhecimento de padrões estatístico operado em dois modos: treinamento ou aprendizagem do sistema; e classificação ou teste do sistema baseado em alguma função dos dados, chamada de classificador. Matematicamente, um classificador é uma função $d : \chi \longrightarrow \Omega$, tal que $d(\mathbf{x}) = \omega_g \in \Omega \ = \omega_1, \omega_2, ..., \omega_r$ (McLachlan, 2004). Usando uma função discriminante ϕ_{ω_g} para classe ω_g (ϕ_{ω_g} pode ser f_{ω_g}), o classificador particiona o espaço χ de características em regiões de decisão e todos os vetores de características no interior de uma região de decisão são atribuídos à mesma classe. As fronteiras de decisão

geradas por χ, podem ser determinadas pelo classificador na fase de treinamento.

Os métodos usuais de classificação (Johnson & Wichern, 2007), assumem que f_{ω_g} é conhecida (ou parcialmente conhecida) o que pode não ocorrer na prática. Então, uma maneira alternativa de tratar este problema é estimar f_{ω_g} de uma forma não-paramétrica (de Lima & Atuncar, 2011). Neste caso, as suposições que são feitas sobre a estrutura probabilística geradora dos dados são fracas ou inexistentes tornando o método livre de modelos.

Por isso neste trabalho, nós apresentamos um método não-paramétrico de classificação baseado na distribuição preditiva de um novo vetor de características $\mathbf{x}^{novo}$ dado um conjunto de treinamento $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. As densidades preditivas são estimadas usando o método Kernel com enfoque Bayesiano (Bernardo (1999); de Lima & Atuncar (2011)). Por fim as densidades preditivas estimadas são usadas como função discriminante.

Em nossos estudos de simulação, empregamos algumas estruturas de dados, desde as mais triviais até casos mais complexos, como o de misturas finitas de densidades, em termos separação das classe. Nesses estudos, obtemos resultados satisfatórios onde os métodos propostos neste trabalho, mostraram-se competitivos, e em alguns casos com melhor desempenho, comparados aos métodos usuais da literatura. Também, empregamos nos estudos dos procedimentos, a aplicação em conjuntos de dados de grande uso na literatura, afim de avaliar a aplicabilidade dos métodos propostos em situações reais. Nessa aplicação, os métodos propostos também obtiveram resultados satisfatórios, mesmo aplicados em conjuntos de dados em alta dimensão.

1.1 Organização da Dissertação

Neste texto, além deste capítulo introdutório, temos mais quatro capítulos. No Capítulo 2, fazemos um revisão bibliográfica, apresentando os conceitos fundamentais de Análise Discriminante. Ainda no Capítulo 2, apresentamos os métodos de classificação clássicos na literatura: Os métodos Análise Discriminante Linear; Análise Discriminante Quadrática e o classificador Naive Bayes, em particular o método Naive Bayes Normal,

também empregado em nossas análise para a comparação de modelos. E ainda abordamos o método de classificação não paramétrico empregando estimadores por função núcleo. O capítulo segue ainda com um discussão sobre um problema frequente na aplicação em dados reais, denominado Problema da Alta Dimensionalidade(do inglês, *curse of dimensionality*), e algumas possíveis soluções para o problema, em especial a Análise de Componentes Independentes.

No Capítulo 3, apresentamos a construção de nossos métodos propostos, que empregam a Análise Discriminante via máxima densidade preditiva, e apresentamos três abordagens para a estimação dessa densidade preditiva, uma considerando o produto de funções núcleo, outra também empregando produto de funções núcleo mas utilizando a Análise de Componentes Independentes, e a última sendo uma generalização para o caso multivariado do problema. No Capítulo 4, fizemos uma análise comparativa entre os métodos propostos e os métodos clássicos, submetendo-os, primeiramente a um estudo com dados simulados de diversas estruturas de dados, e em seguida, a uma aplicação em dados reais. No Capítulo 5, temos a conclusão do trabalho e algumas ideias para trabalhos futuros que já estão sendo implementadas.

Temos ainda três anexos, onde o primeiro trata de algumas definições e resultados que servem de suporte para o texto principal. O segundo anexo, traz as distribuições e os valores dos parâmetros utilizados para gerar as observações simuladas. O último anexo, traz os códigos implementados na linguagem de programação R, dos métodos propostos no trabalho.

1.2 Recursos Computacionais

Nesse trabalho usamos a linguagem de programação R (R Development Core Team, 2012) na versão 2.14.2 de 64 bits, para implementar as propostas dessa dissertação, bem como toda análise de dados e construção de gráficos. Para os estudos de simulação usamos 10 micro-computadores do Laboratório de Estatística da Universidade Federal do Amazonas (UFAM), cedidos pelo Departamento de Estatística (DE), nas seguintes confi-

gurações: processador AMD Athlon Dual Core 4450B 2.30GHz 32bits, 2.00Gb memória RAM-DDR2, Windows Vista Bussiness sp1. Para a aplicações em dados reais usamos um notebook com processador Intel(R) CORE(TM) i7 Q740 64bits de 1.73GHz, 4.00Gb de memória RAM-DDR3 e Windows 7 Home Premium Sp1 64bits. O ambiente de programação que usamos foi o sistema tipográfico LaTeX (Lamport, 1994), sob a distribuição MikTex 2.9.

1.3 Produção Científica

- A abordagem desenvolvida empregando produto de funções núcleo (*ver Seção 3.4*), produziu o artigo científico intitulado *Bayesian predictive kernel discriminant analysis* publicado na *Pattern Recognition Letters, v. 34, p. 2079-2085, 2013.*

- Esse trabalho foi apresentado no 20° Simpósio Nacional de Probabilidade e Estatística em João Pessoa-PB(2012), sob forma de comunicação oral, com o título "Análise Discriminante via Distribuições Preditivas Aproximadas".

- Uma produção técnica, em fase de revisão, sob forma de um pacote denominado *BPKDA* para a linguagem de programação R à ser disponibilizado no repositório CRAN-R.

- Mais dois artigos estão em fase de preparação a serem submetidos ainda em 2013, um abordando o procedimento para o caso generalizado dos métodos de classificação propostos e outro abordando o software estatístico desenvolvido implementando essas abordagens.

Capítulo 2

Análise de Discriminante

2.1 Introdução

Reconhecer rostos de pessoas que há tempos que não as vemos, identificar cores ao nosso redor, reconhecer em um manada de milhares de animais as presas em potencial, são exemplos da incrível capacidade que seres humanos e outros animais têm em reconhecer seres, objetos ou ambientes em sua volta, mesmo em condições adversas. Reconhecimento é um pleno senso cognitivo e pode consistir de tarefas simples, como identificar se um ambiente está muito quente ou frio, ou tarefas não tão triviais, como analisar um eletrocardiograma e identificar uma arritmia no paciente (ver por exemplo, Jain *et al.* (2000), Marques de Sá (2001), McLachlan (2004)).

Entendemos por *objetos* algo de nosso interesse que possamos descrever e classificar. A representação conveniente desses objetos, como por exemplo imagens digitais, sinais em formas de ondas ou qualquer outra representação mensurável, que possam ser interpretadas por seres humanos, animais ou máquinas, são chamadas de *padrões* (Hastie *et al.*, 2009).

2.2 Reconhecimento de Padrões

Reconhecimento de Padrões é a disciplina cientifica cuja finalidade é a classificação de objetos em um número de categorias ou classes (Theodoridis & Koutroumbas, 2008). Desde o começo da computação, a construção e implementação de algoritmos para a emulação das habilidades humanas, ou habilidades que estão longe das nossas capacidades, para descrever e classificar objetos tem encontrado as mais intrigantes e desafiadoras tarefas (veja por exemplo, Garg *et al.* (2011)). O objetivo principal das atuais tecnologias, aplicadas em ciência e tecnologia, é o desenvolvimento de métodos capazes de emular as mais variadas habilidades humanas para descrição e classificação de objetos na construção de sistemas autônomos "inteligentes"(Marques de Sá, 2001).

Nossa sociedade evoluiu de uma fase industrial para uma pós-industrial, e a automação na produção industrial e a necessidade de informação tem se tornado de extrema importância. Assim, o *Reconhecimento de Padrões* se tornou uma parte integral da maior parte dos sistemas de máquinas "inteligentes" construídos para tomada de decisões, levando-a ao limite da pesquisa e aplicação na engenharia (Theodoridis & Koutroumbas, 2008).

2.2.1 Similaridade, Classes, Padrões e Características

Uma noção fundamental em *Reconhecimento de Padrões*, independente de qualquer tipo de aplicação, é o conceito de *similaridade*. Nós reconhecemos dois objetos como sendo similares quando eles possuem valores semelhantes em algum atributo em comum, por exemplo, os biólogos e taxonomistas estabelecem relações entre indivíduos baseados em suas características morfológicas, com a finalidade de agrupá-los em classes nas quais as espécies mais se assemelham. Frequentemente a *similaridade* está num sentido mais abstrato, não está na relação entre objetos mas sim entre um objeto em um conceito alvo ou um protótipo. Por exemplo, podemos reconhecer uma maçã pelas suas características correspondentes a uma imagem idealizada, ou protótipo, e assim diferenciá-la de outros frutos.

Classes são os estados "naturais" ou categorias de objetos associados com conceitos

ou protótipos. Neste trabalho vamos denotar por $\omega_1, \omega_2, \omega_3, \ldots, \omega_r$ como as r classes onde os objetos serão alocados e Ω o conjunto de todas as classes. *Padrões* (do inglês, *patterns*) são as representações "físicas" dos objetos. Podem ser imagens digitais, sinais ou simplesmente observações para um conjunto de variáveis mensuradas sobre os objetos. *Características* (do inglês, *features*) são os aspectos selecionados para descrição dos objetos, derivados dos *Padrões*, empregados para alocação dos objetos nas classes. Essas característica são representadas na forma de um vetor, denominado na literatura por *Vetor de Características* (do inglês, *feature vector*). Esse vetores assumem valores no *Espaço das Características*(do inglês, *feature space*), esse espaço tem propriedades regidas de acordo com a medida de similaridade definida (Marques de Sá, 2001).

2.2.2 Reconhecimento de Padrões Supervisionado e Não Supervisionados

Em geral, os problemas de reconhecimento de padrões são classificados em duas categorias: *Reconhecimento de Padrões Supervisionado*(RPS) e *Reconhecimento de Não Supervisionados*(RPNS) (Hastie *et al.*, 2009). O conceito de supervisionado está relacionado com o fato de ser conhecida a origem de um determinado objeto, com relação as classes envolvidas no problema abordado.

Considere que um determinado local, onde se admita somente a entrada de pessoas autorizadas, empregando para isso um dispositivo de identificação baseado na análise do padrão da íris das pessoas. Após uma fase de reconhecimento de cada pessoa autorizada e o armazenamento das correspondentes informações, todas as vezes que uma delas ou qualquer outra não autorizada tentar acessar o local, o dispositivo deve ser capaz de identificar a pessoa como autorizada ou não, comparando suas informações com aquelas armazenadas em seu banco de dados. Esse é o paradigma do RPS, onde seu desígnio é, primeiramente, o reconhecimento dos padrões existentes nas classes predefinidas, afim de criar uma regra, ou função, que discrimine ou classifique um novo objeto, do qual não se tem a informação da classe, em uma das classes predeterminadas. Na comunidade estatística os problemas em RPS, em geral, são denominados de *Análise de Discriminante*(A.D.)

(Johnson & Wichern, 2007).

Considere agora uma situação onde temos uma imagem de sensoriamento remoto correspondente a uma determinada região de floresta da Amazônica. Em geral essas imagens não são muito nítidas ou bem definidas, no sentido de que se possa identificar objetos de interesse, como por exemplo, uma área de desmatamento com uma possível extração ilegal de madeira. Mesmo que seja possível essa identificação por seres humanos, nestas aplicações reais o objetivo é a automação desse processo. Então, em cada imagem não é possível predeterminar os tipos de ocorrências, como rios, árvores, plantações e desmatamentos, contidas nessas imagens. Se pensarmos nas ocorrências (rios, lagos, floresta, etc) como classes e cada *pixel* (menor elemento que compõe uma imagem) como um objeto pertencente a uma dessas classes, não temos conhecimento prévio sobre a quantidade e estruturas dessas classes e nem sobre a procedência de cada objeto. Esse cenário é o de problemas em RPNS. Na comunidade estatística, em geral, recebe designação de *Análise de Agrupamentos*(do inglês, *Cluster Analysis* ou *Clustering*), e seu objetivo é a identificação da estrutura de grupos (*Clusters*), seguindo o conceito de *similaridade*, e o agrupamento desses objetos de acordo com suas características, de tal forma que os grupos sejam homogêneos internamente e heterogêneos entre si tanto quanto possível (Johnson & Wichern, 2007).

Como o enfoque desse trabalho é voltado para a *Análise de Discriminante*, nossa busca é então por uma regra que funcione como uma função discriminante para a alocação de objetos nas classes existentes. Essa regra recebe a denominação de *classificador*.

2.3 Classificadores

Os classificadores são usados na Teoria da Decisão Estatística, como uma ferramenta de RPS para alocar objetos às classes predefinidas, com base na informação contida no vetor de características correspondentes aos objetos. Em um contexto probabilístico, buscamos uma regra que minimize algum critério de risco. Em teoria, é possível encontramos uma regra ótima, mas isso requer um conhecimento completo sobre a distribuição de pro-

babilidade dos dados, o que na prática é impossível (McLachlan, 2004). Nosso objetivo nesse trabalho, como mencionado anteriormente, é determinar uma regra que se aproxime dessa solução ótima, mesmo não conhecendo a distribuição de probabilidade envolvida. Matematicamente falando, o classificador é uma função $d : \chi \to \Omega$, onde χ é Espaço das Características e $\Omega = \{\omega_1, \omega_2, ..., \omega_r\}$ é o conjunto com r classes às quais um objeto pode ser alocado. Seja $\mathbf{X}' = (X_1, \ldots, X_p)$ o vetor de características que na abordagem estatística é modelado como um vetor aleatório. Note que, a função d induz a r partições de $\mathbb{R}^p$, que denotaremos por $\{R_1, \ldots, R_r\}$, onde $R_g = \{\mathbf{x} : d(\mathbf{x}) = \omega_g\}$ e $\mathbf{x}$ um valor observado de $\mathbf{X}$. Dessa forma, um classificador pode ser visto apenas como uma função que indus uma partição de $\mathbb{R}^p$, levando a um infinidade de classificadores.

Sejam (χ, ξ, P) um espaço de probabilidade e $\mathbf{X}_1, \ldots, \mathbf{X}_n$ uma amostra aleatória de $\mathbf{X}$, modelando observações de n objetos, onde cada vetor $\mathbf{X}_i' = (X_1, \ldots, X_p)$ representa as p características do i-ésimo objeto, $i = 1, \ldots, n$. Cada $\mathbf{X}_i$ é proveniente de uma determinada classe ω_g, onde é assumido ser gerada com probabilidade π_g, $g = 1, \ldots, r$. Sendo $B \in \xi$ um *Boreliano*, a probabilidade condicional de $\mathbf{X}$ dado uma classe ω_g é denotado por $P(\mathbf{X} \in B \mid \omega_g)$, $g = 1, \ldots, r$.

Definição 2.3.1 (Teorema de Bayes) *Sejam $\boldsymbol{X}$ e $\boldsymbol{Y}$ vetores aleatórios definidos no mesmo espaço de probabilidade (χ, ξ, P), e sejam A e B Borelianos, então a regra de Bayes é dada por:*

$$P(\boldsymbol{X} \in A \mid \boldsymbol{Y} \in B) = \frac{P(\boldsymbol{Y} \in B \mid \boldsymbol{X} \in A)P(\boldsymbol{X} \in A)}{P(\boldsymbol{Y} \in B)} \tag{2.1}$$

O objetivo é empregar o *Teorema de Bayes* com a finalidade de desenvolver uma regra de classificação para em RPS. A ideia principal é classificar objetos em termos de probabilidade, ou seja, intuitivamente devemos alocar um objeto a uma classe para a qual tenha a maior probabilidade de pertencer.

2.3.1 Classificador de Máxima Probabilidade Posterior (MAP)

De maneira geral, a classificação de um objeto tem por objetivo a identificação da classe ω_g que o gerou, ou seja, o valor de uma variável aleatória não observável que identifica a classe, com base no valor observado do vetor de características $\mathbf{X}$. Na prática, não é possível obter ω_g mas sim sua estimativa $\hat{\omega}_g$ (Marques de Sá, 2001).

Seguindo o conceito intuitivo de classificador, o objetivo é identificar a classe com maior probabilidade de ter gerado a observação $\mathbf{x}$ de $\mathbf{X}$. Probabilisticamente, seja $(\mathbf{X}, Y)$ um par aleatório, onde Y é uma variável aleatória indicadora que assume valores em $\{1, \ldots, r\}$, ou seja, dessa forma temos $P(\omega_g) = P(Y = \omega_g)$. Então um classificador ideal, segundo esta consideração intuitiva, é dado por:

Definição 2.3.2 (Classificador MAP) *Um objeto com observação $\boldsymbol{x}$ será alocado em ω_g se*

$$\omega_g = \underset{\Omega}{\operatorname{argmax}}\ P(Y = \omega_g \mid \boldsymbol{x}), j = 1, ..., r, \tag{2.2}$$

onde $P(Y = \omega_g \mid \boldsymbol{x})$ é a probabilidade a posteriori da classe ω_g.

O classificador dado na Definição 2.3.2, fundamentado na maximização da distribuição posterior das classes, é conhecido como *Classificador de Máxima a Posteriori* (MAP). Agora, usando o *Teorema de Bayes*, podemos escrever essa probabilidade de um objeto ser alocado numa classe g, desde que $P(\mathbf{x}) > 0$, como:

$$P(Y = \omega_g \mid \mathbf{x}) \quad = \quad \frac{P(\mathbf{x} \mid Y = \omega_g)P(\omega_g)}{P(\mathbf{x})} \tag{2.3}$$

Note que, na expressão (2.3) a quantidade $P(\mathbf{x})$, que é a distribuição de $\mathbf{X}$ independente de classe, é constante com relação a Y. Assim, podemos reescrever a expressão (2.3) como:

$$P(Y = \omega_g \mid \mathbf{x}) \quad \propto \quad P(\mathbf{x} \mid Y = \omega_g)P(\omega_g), \tag{2.4}$$

onde o símbolo $\propto$ denota "proporcionalidade". Para simplificar a notação, vamos empregar $P(\mathbf{x} \mid Y = \omega_g) = f_{\omega_g}(\mathbf{x})$, que representa a distribuição de $\mathbf{X}$ na classe ω_g. Assim, a expressão (2.4) fica reescrita na forma de:

$$P(\omega_g \mid \mathbf{x}) \quad \propto \quad f_{\omega_g}(\mathbf{x})P(\omega_g) \tag{2.5}$$

2.3.2 Classificador de Bayes

Em alguns problemas reais, no entanto, é necessário considerarmos os possíveis erros de classificação envolvidos no problema sendo abordado. Por exemplo, imaginemos a situação onde um determinado dispositivo eletrônico é usado na detecção de aeronaves no espaço aéreo de um aeroporto. Nesse caso é razoável supor que as consequências associadas ao erro de deixar de não detectar a entrada de uma aeronave é mais grave do que aquelas associadas a um falso alerta. Assim, essa situação necessita de se atribuir diferentes custos na tomada de decisão.

Uma solução possível para esse problema é atribuir custos as decisões e construir um classificador de tal forma que o custo esperado seja minimizado. Definimos, então a *função de perda* ou *custo de má classificação*, denotada por $\lambda(i, g)$, que é a perda por alocar um objeto de ω_i em ω_g . Na prática, a função de perda é muito subjetiva e difícil de definir. De acordo com a Teoria da Decisão, é empregado o *Risco Médio* (do inglês, *Average Risk*) na construção do classificador que leva em consideração a função de perda estabelecida (Ripley, 1996).

Probabilisticamente, a regra de classificação $d(\mathbf{X})$ e a função de perda $\lambda(i, g) = \lambda(i, d(\mathbf{X}))$ são variáveis aleatórias, portanto é adequado definir um critério de escolha dos classificadores em termos de valor esperado.

Definição 2.3.3 *A Função de Risco é a perda esperada como função de ω_g, ou seja,*

$$\begin{aligned} R(d, \omega_g) &= E[\lambda(g, d(\boldsymbol{X})) \mid \omega_g] \\ &= \sum_{\substack{k=1 \\ k \neq g}}^{r} \lambda(g,k)\, P(d(\boldsymbol{X}) = \omega_k \mid \omega_g) \end{aligned}$$

Definição 2.3.4 *O Risco Médio é a perda total esperada como função das variáveis aleatórias $\boldsymbol{X}$ e ω_g, ou seja,*

$$\begin{aligned} R(d) &= E[R(d, Y)] \\ &= \sum_{g=1}^{M} R(r, g)\, P(\omega_g) \\ &= \sum_{g=1}^{M} \sum_{\substack{k=1 \\ k \neq g}}^{M} \lambda(g,k)\, P(d(\boldsymbol{X}) = \omega_k \mid \omega_g)\, P(\omega_g) \end{aligned} \tag{2.6}$$

Afim de que o *Risco Médio* seja minimizado, para uma dada função de perda $\lambda(\cdot, \cdot)$,

$$\begin{aligned} R(d) &= E[R(d, Y)] \\ &= E[\lambda(Y\,C, d(\mathbf{X}))] \\ &= E\{E[\lambda(Y, d(\mathbf{X})) \mid \mathbf{X}]\} \\ &= \int_{\mathbb{R}^r} E[\lambda(Y, d(\mathbf{x})) \mid \mathbf{X} = \mathbf{x}] dF_{\mathbf{X}}(\mathbf{x}) \end{aligned} \tag{2.7}$$

Assim, para minimizar $R(d)$ basta minimizar $E[\lambda(Y, r(\mathbf{x})) \mid \mathbf{X} = \mathbf{x}]$, logo

$$\begin{aligned} E[\lambda(Y, d(\mathbf{x}) = \omega_g) \mid \mathbf{X} = \mathbf{x}] &= \sum_{i=1}^{r} \lambda(i, g)\, P(Y = \omega_i \mid \mathbf{X} = \mathbf{x}) \\ &= \sum_{i=1}^{r} \lambda(i, g)\, \frac{f_{\omega_i}(\mathbf{x})\, P(\omega_i)}{f(\mathbf{x})} \end{aligned} \tag{2.8}$$

Logo, a expressão é minimizada para ω_g tal que $\sum_{i=1}^{r} \lambda(i, g)\, f_{\omega_i}(\mathbf{x})\, P(\omega_i)$ seja o mínimo. Assim, segue a definição:

Definição 2.3.5 (*Classificador de Bayes de Mínimo Risco*) *O classificador ótimo*

que minimiza o Risco Médio é dado por:

$$d^*(\boldsymbol{x}) = \omega_g \;\; se \;\; \sum_{i=1}^{r} \lambda(i,g)\, f_{\omega_i}(\boldsymbol{x})\, P(\omega_i) = \min_j \sum_{i=1}^{M} \lambda(i,j)\, f_{\omega_i}(\boldsymbol{x})\, P(\omega_i) \tag{2.9}$$

Se duas ou mais classes atingirem essa probabilidades mínima, o objeto é alocado em qualquer uma destas classes.

Como na prática a *função de perda* não é fácil de ser definida, em muitas aplicações empregamos a *função de perda 0-1.*

Definição 2.3.6 (*Função de Perda 0-1*) *é definida como:*

$$\lambda(i,g) = \begin{cases} 1 & se \quad i \neq g \\ 0 & se \quad i = g \end{cases} \tag{2.10}$$

Assim, com a função em (2.10), temos

$$d(\mathbf{X}) = \omega_g \;\; se \;\; \sum_{i=1}^{r} \lambda(i,g) f_{\omega_i}(\mathbf{x})\, P(\omega_i) = \min_j \sum_{i=1}^{r} \lambda(i,j)\, f_{\omega_i}(\mathbf{x})\, P(\omega_i) \tag{2.11}$$

onde

$$\begin{aligned} \sum_{i=1}^{r} \lambda(i,g) f_{\omega_i}(\mathbf{x})\, P(\omega_i) &= \sum_{\substack{i=1 \\ i \neq g}}^{r} f_{\omega_i}(\mathbf{x})\, P(\omega_i) \\ &= 1 - f_{\omega_g}(\mathbf{x})\, P(\omega_g) \end{aligned} \tag{2.12}$$

e

$$\begin{aligned} \min_j \sum_{i=1}^{r} \lambda(i,j)\, f_{\omega_i}(\mathbf{x})\, P(\omega_i) &= \min_j \sum_{\substack{i=1 \\ i \neq j}}^{r} f_{\omega_i}(\mathbf{x})\, P(\omega_i) \\ &= \min_j 1 - f_{\omega_j}(\mathbf{x})\, P(\omega_j). \end{aligned} \tag{2.13}$$

Mas minimizar $1 - f_{\omega_j}(\mathbf{x})\, P(\omega_j)$ é equivalente a maximizar $f_{\omega_j}(\mathbf{x})\, P(\omega_j)$. Portanto, o

classificador de Bayes neste caso fica na forma:

$$d^*(\mathbf{x}) = \omega_g \;\; se \;\; f_{\omega_g}(\mathbf{x})\, P(\omega_g) = \max_j f_{\omega_j}(\mathbf{x})\, P(\omega_j) \tag{2.14}$$

O *classificador de Bayes* em (2.14) traduz o conceito de que o objeto deve ser alocado na classe a que ele é mais verossímil, segundo suas próprias características, e ainda reforça essa consideração dentro do contexto de Reconhecimento de Padrões. Como a *regra de Bayes*, com a função de perda 0-1, minimiza a probabilidade de má classificação, ou seja, é ótimo, é adequado para a comparação de classificadores. Em outras palavras, consideramos um bom classificador aquele que possuir um *Risco Médio* mais próximo possível ao do *classificador de Bayes*.

Em situações reais, é impossível obter as *densidades condicionais* $f_{\omega_g}(\mathbf{x})$ e as *probabilidades a priori das classes* $P(\omega_g),\ g = 1, \ldots, r$, impedindo a construção do *classificador de Bayes*. A solução desse problema é o procedimento de estimação dessas quantidades, como a finalidade de construir um classificador que se aproxime da Regra de Bayes. Neste trabalho, vamos considerar o caso da *função de perda 0-1* para todos os métodos discriminantes propostos. Os métodos propostos serão comparados com alguns classificadores mais usuais na literatura, empregando abordagens paramétricas, como *Análise de Discriminante Linear*, *Análise de Discriminante Quadrática* e *Naive Bayes Normal*.

2.4 Métodos Discriminantes

Como na prática as distribuições condicionais e as probabilidades *a priori* não são conhecidas, então o objetivo é encontrar aproximações úteis para essas quantidades. Um procedimento é atribuir modelos de probabilidade para os dados e tais modelos têm um conjunto de parâmetros a serem estimados. Portanto, se constitui em uma abordagem paramétrica para obter estimativas para as funções f_{ω_g}. Outra abordagem consiste em não postular modelos paramétricos para as distribuições f_{ω_g} e e empregar procedimentos não paramétrico para estimação de densidades, portanto, é o paradigma não paramétrico

do problema.

2.4.1 Classificadores baseado no Modelo Normal

Métodos de classificação baseados em modelos normais predominam na prática estatística por causa de sua simplicidade e boa eficiência numa ampla variedade de problemas reais (Johnson & Wichern, 2007). Assim, a distribuição em cada classe ω_g é modelada segundo um modelo Normal p-variado com vetor de médias $\boldsymbol{\mu}_g$ e matriz de covariancias $\boldsymbol{\Sigma}_g$, logo

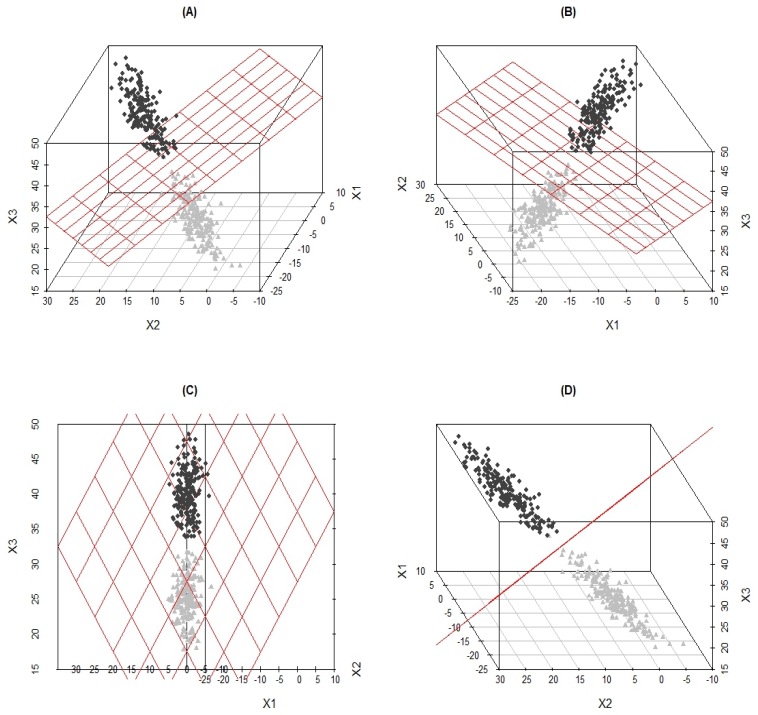
$$f_{\omega_g}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\,|\Sigma_g|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\}, g = 1,\ldots,r. \quad (2.15)$$

Em um caso especial, vamos assumir que as matrizes de covariâncias das classes são iguais, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, $g = 1,\ldots,r$. Numa comparação entre duas classes distintas g e l, é suficiente analisarmos o logaritmo da razão (do inglês, *log-ratio*) entre as probabilidades posteriores

$$
\begin{aligned}
\log \frac{P(Y=g \mid \mathbf{X}=\mathbf{x})}{P(Y=l \mid \mathbf{X}=\mathbf{x})} &= \log \frac{f_{\omega_g}(\mathbf{x})P(\omega_g)}{f_{\omega_l}(\mathbf{x})P(\omega_l)} \\
&= \log \frac{\frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\} P(\omega_g)}{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)\right\} P(\omega_l)} \\
&= \left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\} - \left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)\right\} \\
&+ \log \frac{P(\omega_g)}{P(\omega_l)} \\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_g) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_l) + \log \frac{P(\omega_g)}{P(\omega_l)} \\
&= -\frac{1}{2}\left\{\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g - \boldsymbol{\mu}_g'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_g'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g \right. \\
&\qquad \left. -\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l + \boldsymbol{\mu}_l'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_l'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l\right\} + \log \frac{P(\omega_g)}{P(\omega_l)} \\
&= \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_l) - \frac{1}{2}(\boldsymbol{\mu}_g + \boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_l) + \log \frac{P(\omega_g)}{P(\omega_l)}. \quad (2.16)
\end{aligned}
$$

A expressão (2.16) é uma função linear de $\mathbf{x}$, uma vez que a parte quadrática ($\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ $-\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$) da expressão é nula. Em virtude disso o limite da decisão entre as classes g e l é um hiperplano em $\mathbb{R}^p$, ou seja, se dividirmos $\mathbb{R}^p$ em regiões correspondentes as classes $\omega_1, \omega_2, ..., \omega_r$, então essas regiões serão separadas por hiperplanos.

Figura 2.1: Separação de duas classes por um hiperplano em 3 dimensões.



Na Figura 2.1, vemos a separação de duas classes, oriundas de populações normais com vetores de médias diferentes e matrizes de covariâncias iguais, por um plano gerado por (2.16), com diferentes perspectivas, onde a classe 1 é representada pelos pontos em formato de círculo, a classe 2 é representada pelos pontos em formato de triângulo. Nesse caso, como há somente duas classes envolvidas, vemos apenas um plano representando os limites de decisão. Se adicionarmos uma terceira classe teríamos mais dois planos gerando os limites lineares dessas regiões, e assim por diante, ou seja, para r classes teríamos $\binom{r}{2}$ planos.

Definição 2.4.1 (Análise Discriminante Linear)
Considere $\boldsymbol{X} \mid Y = \omega_g \sim Normal(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$, $g = 1, \ldots, r$. *De (2.16), o classificador de Bayes com função de perda 0-1, é da forma:*

$$d_{ADL}(\boldsymbol{x}) = \underset{\Omega}{\text{argmax}} \quad \left\{ \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g - \frac{1}{2}(\boldsymbol{\mu}_g'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_g) + logP(\omega_g) \right\}. \tag{2.17}$$

Esse procedimento é denominado de Análise Discriminante Linear (ADL).

Na prática, nós não conhecemos os parâmetros das distribuições Normais e as probabilidades *a priori* das classes, então é necessário estimar a regra definida em (2.4.1), ou seja, estimar os parâmetros envolvidos. Em geral, emprega-se os estimadores de máxima verossimilhança dados por (Hastie *et al.*, 2009):

$$\hat{P}(\omega_g) = \frac{n_g}{n}, \text{ onde } n_g \text{ é o número de observações na classe } \omega_g; \tag{2.18}$$

$$\hat{\boldsymbol{\mu}}_g = \sum_{i:Y_i=g} \frac{\mathbf{x}_i}{n_g}; \tag{2.19}$$

$$\hat{\boldsymbol{\Sigma}} = \sum_{g=1}^{r} \sum_{i:Y_i=g} \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)'}{n-r}; \tag{2.20}$$

Assim, a regra em (2.17) estimada é da forma:

$$\hat{d}_{ADL}(\mathbf{x}) = \underset{\Omega}{\text{argmax}} \quad \left\{ \mathbf{x}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_g - \frac{1}{2}(\boldsymbol{\mu}_g'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_g) + log\hat{P}(\omega_g) \right\}. \tag{2.21}$$

O outro caso particular da análise discriminante para modelo normal, é o caso mais geral, onde as matrizes de covariâncias não são consideradas iguais, assim o termo quadrático em (2.16) não é cancelado. A função *log-ratio*, entre as classes ω_g e ω_l, é dada por:

$$\begin{aligned}
\log \frac{P(Y=g \mid \mathbf{X}=\mathbf{x})}{P(Y=l \mid \mathbf{X}=\mathbf{x})} &= \log \frac{f_{\omega_g}(\mathbf{x})P(\omega_g)}{f_{\omega_l}(\mathbf{x})P(\omega_l)} \\
&= \log \frac{\frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_g^{1/2}\right|} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\} P(\omega_g)}{\frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_l\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)\right\} P(\omega_l)} \\
&= \log \frac{|\boldsymbol{\Sigma}_l|^{1/2}}{|\boldsymbol{\Sigma}_g|^{1/2}} - \frac{1}{2}\left\{(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g) - (\mathbf{x}-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)\right\} \\
&\quad + \log \frac{P(\omega_g)}{P(\omega_l)},
\end{aligned} \tag{2.22}$$

que é uma função quadrática de $\mathbf{x}$. Portanto, os limites entre as regiões de decisão são superfícies hiperquadráticas, e podendo assumir qualquer forma geral - hiperplanos, hiperesferas, hiperelipsóides, hiperparábolas, etc. (Duda *et al.*, 2000).

Definição 2.4.2 (Análise Discriminante Quadrática)
Considere $\boldsymbol{X} \mid Y = \omega_g \sim \ Normal(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, $g = 1, \ldots, r$. *De (2.22), o classificador de Bayes com função de perda 0-1, é da forma:*

$$d_{ADQ}(\boldsymbol{x}) = \underset{\Omega}{\operatorname{argmax}} \left\{ -\frac{1}{2} log\,|\boldsymbol{\Sigma}_g| - \frac{1}{2}\left\{(\boldsymbol{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_g)\right\} + logP(\omega_g) \right\}. \quad (2.23)$$

Para essa situação na prática, os estimadores de máxima verossimilhança são (Hastie *et al.*, 2009):

$$\hat{P}(\omega_g) = \frac{n_g}{n}, \text{ onde } n_g \text{ é o número de observações na classe } \omega_g; \quad (2.24)$$

$$\hat{\boldsymbol{\mu}}_g = \sum_{i:Y_i=g} \frac{\mathbf{x}_i}{n_g}; \quad (2.25)$$

$$\hat{\boldsymbol{\Sigma}}_{\omega_g} = \sum_{i:Y_i=g} \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)'}{n_g - 1}; \quad (2.26)$$

Assim, a regra em (2.23) estimada é da forma:

$$\hat{d}_{ADQ}(\mathbf{x}) = \underset{\Omega}{\operatorname{argmax}} \left\{ -\frac{1}{2} log \left|\hat{\boldsymbol{\Sigma}}_g\right| - \frac{1}{2}\left\{(\mathbf{x}-\hat{\boldsymbol{\mu}}_g)'\hat{\boldsymbol{\Sigma}}_g^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_g)\right\} + log\hat{P}(\omega_g) \right\}. \quad (2.27)$$

Uma vez que as estimativas dos parâmetros são inseridos para as correspondentes quantidades populacionais desconhecidas, não há garantia que as regras resultante em (2.21) e (2.27) irão minimizar o *Risco Médio* em uma determinada aplicação. Isso ocorre porque a regra ótima que minimiza o *Risco Médio* é derivada assumindo que as densidades Normais multivariadas são completamente especificadas, em ambas LDA e QDA. As expressões em (2.21) e (2.27) são simplesmente uma estimativa da regra ótima. Em um contexto onde podemos retirar várias amostras da mesma população, podemos então estimar o *Risco Médio* dessas regras. É razoável esperarmos que essas regras estimadas tenham uma performance melhor se o tamanho amostral for grande. De fato, as quantidades $\hat{\boldsymbol{\mu}}_g$ e $\hat{\boldsymbol{\Sigma}}$ convergem em probabilidade para $\boldsymbol{\mu}_g$ e $\boldsymbol{\Sigma}$, respectivamente, na LDA, e $\hat{\boldsymbol{\mu}}_g \xrightarrow{p} \boldsymbol{\mu}_g$ e $\hat{\boldsymbol{\Sigma}}_g \xrightarrow{p} \boldsymbol{\Sigma}_g$ na QDA (McLachlan (2004), Johnson & Wichern (2007), Izenman

(2008)).

Em uma aplicação real, um problema sério é adequação dos dados as distribuições Normais multivariadas. Se eles não atendem essa exigência, um procedimento é fazer uma transformação nos dados não normais, para se obter uma aproximação normal, e testes para verificar a igualdade das matrizes de covariâncias com o objetivo de saber se a regra linear (ADL) ou a regra quadrática (ADQ) é adequada. Em muitos casos no entanto, estas regras são empregadas com a expectativa que funcionem razoavelmente bem considerando a capacidade de modelagem da distribuição Normal. Em qualquer caso, devemos sempre verificar o desempenho de qualquer classificador empregado, assim, de forma empírica, podemos usar os dados para treinar as regras de classificação e usar esses mesmos dados para conferir seus desempenhos, e usar a regra com melhor performance para o problema em questão.

2.4.2 Classificador *Naive Bayes*

Em muitas situações práticas, nos deparamos com um problema típico que é alta dimensão dos dados. São situações em que há um número bem maior de *características* observadas em um objeto do que o número de objetos observados, isso decorre de muitos casos onde há uma grande dificuldade em obter amostras ou não existem amostras suficientes. Um procedimento com uma longa e bem sucedida história dentro do paradigma de classificação (Hand & Yu, 2001), devido a sua simplicidade, eficiência e eficácia tem sido usado como ferramenta para uma solução empírica do problema da alta dimensionalidade, denominado na literatura de *Naive Bayes* (na literatura em inglês aparece os termos: *Idiot's Bayes*, *Simple Bayes* e *Independent Bayes*) (Webb *et al.*, 2005).

Para empregar o *Classificador de Bayes* em (2.14) é necessário conhecer completamente $f_{\omega_g}(\mathbf{x})P(\omega_g)$ e como dito anteriormente em caso reais essa informação, em geral, não é conhecida e devemos então estimá-las. Vamos nos concentrar agora somente em $f_{\omega_g}(\mathbf{x})$ que é o valor da função de probabilidade no ponto $\mathbf{x}$ para a classe ω_g, mas $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ uma observação qualquer de $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$ um vetor aleatório p-variado, assim $f_{\omega_g}(\mathbf{x}) = f_{\omega_g}(x_1, x_2, \ldots, x_p)$. O procedimento do *Naive Bayes* assume

que em cada classe ω_g, $X_1, X_2, \ldots, X_p$ são independentes, assim

$$\begin{aligned} f_{\omega_g}(\mathbf{x}) &= f_{\omega_g}(x_1, x_2, \ldots, x_p) \\ &= f(x_1, x_2, \ldots, x_p \mid \omega_g) \\ &= f(x_1 \mid \omega_g) f(x_2 \mid \omega_g, x_1) \cdot \ldots \cdot f(x_p \mid \omega_g, x_1, x_2, \ldots, x_{p-1}), \\ &\quad \text{por independência,} \\ &= f(x_1 \mid \omega_g) f(x_2 \mid \omega_g) \cdot \ldots \cdot f(x_p \mid \omega_g) \\ &= f_{\omega_g}(x_1) f_{\omega_g}(x_2) \cdot \ldots \cdot f_{\omega_g}(x_p) \\ &= \prod_{j=1}^{p} f_{\omega_g}(x_j) \end{aligned} \tag{2.28}$$

Claramente o modelo em (2.28) parece não ser realístico para a maior parte dos problemas, pois a independência é uma suposição muito forte, uma vez que raramente as matrizes de covariâncias são diagonais (Hand & Yu, 2001). Além disso, na prática, podemos encontrar vários exemplos onde esse modelo seria incorreto.

Hand & Yu (2001) fazem uma extensa revisão de diversos exemplos da literatura onde a abordagem *Naive Bayes* se mostrou um melhor classificador, ou pelo menos uma boa escolha, quando comparados com outros métodos mas adequados à situação analisada (veja também Domingos & Pazzani (1997)). Também Webb *et al.* (2005) fazem uma comparação entre diversos classificadores em vários conjuntos de dados, bastante conhecidos na literatura, e obtêm resultados relevantes com relação ao *Naive Bayes*. E ainda, de modo prático, é importante ressaltar que esta abordagem apresenta menos parâmetros a serem estimados, como por exemplo, no modelo Normal onde iriamos somente estimar as médias e as variâncias, e não as covariâncias.

Domingos & Pazzani (1997) mostram em seu estudo empírico que o *Naive Bayes* ainda pode ser ótimo, com a função de perda 0-1, mesmo com a pressuposição de independência para alguns problemas onde há um alto grau de dependência entre *características*. Assim, todas essas informação na literatura nos levam a considerar que, mesmo sob uma suposição

bastante inadequada em certas ocasiões, o classificador *Naive Bayes* é uma boa escolha para algumas situações, e sem dúvida um classificador que devemos levar em consideração na comparação do método proposto nesse trabalho.

Na prática, no entanto, as distribuições marginais $f_{\omega_g}(x_j), j = 1, \ldots, p$, são desconhecidas. Assim, de modo semelhante a ADL e ADQ, vamos considerar que $f_{\omega_g}(\mathbf{x})$ é a densidade Normal multivariada, mas com $X_1, X_2, \ldots, X_p$ independentes, assim

$$f_{\omega_g}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\left|\mathbf{\Sigma}^{1/2}\right|}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\}.$$

Como os $X_i's$ são independentes, implica que $\mathbf{\Sigma} = (\Sigma_1^2, \Sigma_2^2, \ldots, \Sigma_p^2)'I$, onde I é a matriz identidade de ordem p. Logo,

$$\begin{aligned}
f_{\omega_g}(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2}\left|[(\Sigma_{1,g}^2, \ldots, \Sigma_{p,g}^2)'I]^{1/2}\right|}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'[(\Sigma_{1,g}^2, \ldots, \Sigma_{p,g}^2)'I]^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\} \\
&= \frac{1}{(2\pi)^{p/2}\left|[(\Sigma_{1,g}, \ldots, \Sigma_{p,g})'I]\right|}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'[(\frac{1}{\Sigma_{1,g}^2}, \ldots, \frac{1}{\Sigma_{p,g}^2})'I](\mathbf{x}-\boldsymbol{\mu}_g)\right\} \\
&= \frac{1}{(2\pi)^{p/2}(\Sigma_{1,g}\cdot\ldots\cdot\Sigma_{p,g})}\exp\left\{-\frac{1}{2}\left[\frac{1}{\Sigma_{1,g}^2}(x_i-\mu_{i,g})^2 + \ldots + \frac{1}{\Sigma_{p,g}^2}(x_i-\mu_{i,g})^2\right]\right\} \\
&= \frac{1}{(2\pi)^{p/2}(\Sigma_{1,g}\cdot\ldots\cdot\Sigma_{p,g})}\exp\left\{-\frac{1}{2}\left[\frac{1}{\Sigma_{1,g}^2}(x_i-\mu_{i,g})^2 + \ldots + \frac{1}{\Sigma_{p,g}^2}(x_p-\mu_{p,g})^2\right]\right\} \\
&= \frac{1}{(2\pi)^{1/2}\Sigma_{1,g}\cdot\ldots\cdot(2\pi)^{1/2}\Sigma_{p,g}}\exp\left\{-\frac{(x_i-\mu_{i,g})^2}{2\Sigma_{1,g}^2}\right\}\cdot\ldots\cdot\exp\left\{-\frac{(x_p-\mu_{p,g})^2}{2\Sigma_{p,g}^2}\right\} \\
&= \frac{1}{(2\pi)^{1/2}\Sigma_{1,g}}\exp\left\{-\frac{(x_i-\mu_{i,g})^2}{2\Sigma_{1,g}^2}\right\}\cdot\ldots\cdot\frac{1}{(2\pi)^{1/2}\Sigma_{p,g}}\exp\left\{-\frac{(x_p-\mu_{p,g})^2}{2\Sigma_{p,g}^2}\right\} \\
&= f_{\omega_g}(x_1)f_{\omega_g}(x_2)\cdot\ldots\cdot f_{\omega_g}(x_p), \qquad (2.29)
\end{aligned}$$

onde $f_{\omega_g}(x_j)$ é uma função densidade $Normal(\mu_{j,g}, \Sigma_{j,g}^2), j = 1, \ldots, p \;\; e \;\; g = 1, \ldots, r$. Esse procedimento é denominado de *Naive Bayes Normal*. Por ser bastante flexível, o *Naive Bayes* admite a imposição de vários modelos às distribuições marginais, possibilitando a criação de vários classificadores, demonstrando então sua ampla aplicabilidade. O conceito do *Naive Bayes* vai além de uma abordagem paramétrica, se estendendo também a abordagem não paramétrica, que é uma das propostas desse trabalho.

2.4.3 Classificador com estimação por Função Núcleo

Nos classificadores anteriores, nós assumimos, muitas vezes de forma subjetiva, um modelo probabilístico paramétrico para as distribuições nas classes $(f_{\omega_g}(\cdot \mid \boldsymbol{\theta}_g), g = 1, \ldots, r)$, e empregamos procedimentos para a estimação de seus respectivos parâmetros $(\boldsymbol{\theta}_g)$, o que em certas situações práticas implica em um ganho em redução computacional. Porém, nem sempre é adequado a imposição de modelos paramétricos aos dados. Uma alternativa é empregar uma abordagem *não paramétrica*. Dentre os métodos de estimação de densidade não paramétricos, o mais frequentemente empregado é o *Estimador de Densidades por Função Núcleo*(do inglês, Kernel Density Estimation) com *parâmetro de suavização* apropriado (Scott, 1992).

Sejam $\boldsymbol{\theta}$ uma matriz *pxp* não singular, denominada *matriz de largura de banda*, e $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ uma função núcleo que satisfaz as seguintes condições:

$$\mathcal{K}(\mathbf{w}) > 0 \; e \tag{2.30}$$

$$\int_{\mathbb{R}^p} \mathcal{K}(\mathbf{w}) d\mathbf{w} = 1. \tag{2.31}$$

Então, o *estimador por função núcleo multivariada* é definida por (Wand & Jones, 1993):

$$\begin{aligned} \hat{f}(\mathbf{x} \mid \boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{x}_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} |\boldsymbol{\theta}|^{-1/2} \mathcal{K}(\boldsymbol{\theta}^{-1/2}(\mathbf{x} - \mathbf{x}_i)), \end{aligned} \tag{2.32}$$

onde n é o número de observações disponíveis no processo de estimação. Na literatura, existem diversas funções núcleo disponíveis que podem ser empregadas, e que foram projetados para diversas situações (Scott, 1992). Nesse trabalho, vamos considerar apenas

uma das mais populares, a *função núcleo Normal multivariada*, definido como:

$$\mathcal{K}(\mathbf{w}) = \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp^{-\frac{1}{2}\mathbf{w}'\mathbf{w}} \tag{2.33}$$

então, a função núcleo $\mathcal{K}_{\boldsymbol{\theta}}(\mathbf{x}-\mathbf{x}_i)$ representa a função densidade da distribuição $Normal(\mathbf{x}_i, \boldsymbol{\theta})$ dada por

$$\frac{1}{(2\pi)^{p/2}} |\boldsymbol{\theta}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)'\boldsymbol{\theta}^{-1}(\mathbf{x}-\mathbf{x}_i)}. \tag{2.34}$$

Muitos trabalhos teóricos focam nos vários aspectos das propriedades de estimação relacionadas as características das funções núcleo. Agora, a qualidade das estimativas da densidade é reconhecida como sendo primariamente determinada pela escolha do *parâmetros de suavização*, e menos pela escolha da função núcleo. Na literatura, existem diversas abordagens para a determinação do *parâmetro de suavização*(ver Scott (1992), Ghosh & Ramamoorthi (2003) e de Lima & Atuncar (2011)).

Portanto, num contexto de Análise de Discriminante, podemos escrever a função de probabilidade ou densidade de probabilidade de cada classe ω_g como

$$\hat{f}_{\omega_g}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n_g} |\boldsymbol{\theta}_g|^{-1/2} \mathcal{K}(\boldsymbol{\theta}_g^{-1/2}(\mathbf{x} - \mathbf{x}_{ig})), \tag{2.35}$$

onde n_g é a quantidade de observações, $\boldsymbol{\theta}_g$ é a matriz de largura de banda e $\mathbf{x}_{ig}$ as observações na classe ω_g para $g = 1, \ldots, r$. Assim, a regra de classificação, similarmente as regras anteriores, é dada por:

$$d_{\mathcal{K}}(\mathbf{X}) = \underset{\omega_g}{\text{argmax}}\ \hat{f}_{\omega_g}(\mathbf{x})P(\omega_g). \tag{2.36}$$

2.5 O Problema da Alta Dimensionalidade

Um dos problemas recorrentes na aplicação das técnicas estatísticas no reconhecimento de padrões é a *alta dimensionalidade* no *Espaço das Características*, também conhecida na literatura como *Curse of Dimensionality* (Jain *et al.*, 2000). Não há uma definição clara de *dados de alta dimensão*. Todavia, a partir de um ponto vista estatístico, é usual focar na razão $\rho = p/n$ (p é a dimensão dos dados e n é o tamanho da amostra), para um ρ grande, temos uma alta dimensão. Um conjunto de dados multivariado "padrão", em termos de dimensionalidade, é aquele tal que $p \ll n$, e um conjunto de dados considerado de alta dimensão é aquele onde p é próximo de n ou $p \gg n$. Há também casos extremos onde $p = +\infty$ que corresponde aos chamados *dados funcionais* (Ferraty, 2010).

Muitos procedimentos que são analiticamente ou computacionalmente tratáveis num espaço de baixa dimensão, podem se tornar completamente impraticáveis em espaço de dimensão maior (Duda *et al.*, 2000). Tal dificuldade ocorre em particular com os métodos ADL e ADQ, descritos nesse trabalho. Estes métodos, que são baseados em modelo Normais apresentam dificuldades quanto a inversão da matriz de covariâncias estimada. Na ADL como as matrizes de covariâncias são consideradas iguais ($\mathbf{\Sigma}_g = \mathbf{\Sigma}$) usamos todo o conjunto de dados de todas as r classes para estimar a matriz de covariâncias e, ainda assim, numa situação onde $p \gg n$, não há garantias que esta matriz seja não singular. Na ADQ essa situação apenas piora, pois para estimar as matrizes de covariâncias usamos apenas as observações de cada classe para estimar suas respectivas matrizes de covariâncias. Apesar da ADQ ter um desempenho melhor que a ADL para conjuntos de treinamento grandes, essa situação pode ser completamente diferente para conjuntos com alta dimensionalidade (Ripley, 1996).

Várias técnicas tem sido desenvolvidas para contornar o problema de alta dimensionalidade, cujo objetivo principal é a *Redução de Dimensão*, ou seja, a busca por uma representação dos dados em uma dimensão menor que p, que possam tornar o problema mais tratável.

2.5.1 Redução de Dimensão

Há duas abordagens principais na redução de dimensão: *seleção de variáveis* (do inglês, *feature selection*) e *extração de variáveis* (do inglês, *feature extraction*) (Theodoridis & Koutroumbas, 2008). O termo *seleção de variáveis* refere-se a um procedimento que seleciona apenas as observações correspondentes a um melhor subconjuntos de variáveis(*características*), dentre as variáveis originais, que sejam capazes de discriminar eficientemente os objetos. Nesse contexto, as variáveis selecionadas preservam suas interpretações físicas originais (veja por exemplo Stingo & Vannucci (2010) e Maugis *et al.* (2011)). A *extração de variáveis* trata dos métodos que criam novas variáveis baseadas em transformações ou combinações das variáveis originais, e tais transformações podem prover uma habilidade discriminante melhor a um classificador, entretanto essas novas variáveis podem não ter um significado físico claro. Embora exista distinção entre *seleção de variáveis* e *extração de variáveis* são empregados alternadamente na literatura. Em muitas aplicações a *extração de variáveis* precede a *seleção de variáveis*, primeiramente, as variáveis são transformadas a partir dos dados e daí as variáveis transformadas com baixa habilidade discriminante, ou seja, aquela que não influencia na discriminação são descartadas (Jain *et al.*, 2000).

A discussão sobre a *redução de dimensão* não é o foco deste trabalho, no entanto, há dois procedimentos de *extração de variáveis* bastante empregados na estatística, *Análise de Componentes Principais*(PCA, do termo em inglês *Principal Component Analysis*) e *Análise de Componentes Independentes (ICA, do termo em inglês Independent Component Analysis)*, que descrevemos sumariamente.

2.5.2 Análise de Componentes Principais

Em PCA é implementada uma transformação linear das variáveis originais em variáveis não correlacionadas, denominadas de *componentes principais*, na qual as variáveis transformadas preservam a variância total das variáveis originais (Johnson & Wichern, 2007). Nas componentes principais é feita uma análise com o objetivo de selecionar al-

gumas delas e, desta forma, as observações originais são substituídas pelas observações das componentes principais selecionadas(os *scores*), de tal maneira que preserve a maior parte da variabilidade total das variáveis originais. É importante salientar que nem sempre as transformações por PCA são úteis nos problemas de análise de discriminante, uma vez que mesmo quando no espaço das variáveis originais as classes estejam bem separadas, no espaço das componentes principais a estrutura original de separação pode ser completamente obscurecida (Hastie *et al.*, 2009). Se admitirmos a normalidade das distribuições nas classes, a transformação linear preserva a normalidade e, devido a ausência de correlação entre as componentes principais, teríamos independência para as componentes principais(variáveis transformadas), então a abordagem do *Naive Bayes Normal* seria adequada para modelar as observações das componentes principais selecionadas.

2.5.3 Análise de Componentes Independentes

Análise de Componentes Independentes(ICA), do inglês *Independent Component Analisys*, é um método estatístico que por meio de transformações lineares de um vetor aleatório $\mathbf{X}$ observado obtemos um vetor aleatório $\mathbf{Y}$ cujas componentes são estocasticamente independentes. O objetivo principal da ICA é encontrar uma representação $\mathbf{Y} = \mathbf{MX}$, onde $\mathbf{M}$ não é necessariamente uma matriz quadrada, tal que as componentes de $\mathbf{Y}$ sejam independentes. Na prática, o objetivo é aproximar $\mathbf{M}$ tal que $Y_i \in \mathbf{Y}$ são "estatisticamente independentes" tanto quanto possível (Hyvarinen & Oja (2000), Stone (2004)). Na ICA, a pseudo-inversa $\mathbf{A}$ de $\mathbf{M}$ é denominada de *Matriz de Misturas*.

A restrição fundamental na ICA é que as componentes independentes $\mathbf{Y}$ devem ter distribuição mais distante o possível da distribuição Normal. Isso se deve ao fato que, por exemplo, se tivermos duas variáveis aleatórias X_1 e X_2 Normais, não correlacionadas e variância unitária, então não temos nenhuma informação sobre a matriz de misturas $\mathbf{A}$. Pelo Teorema Central do Limite, a soma de variáveis aleatórias independentes $X_1 + X_2 + \cdots + X_n$, sob certas condições, convergem em distribuição para um distribuição Normal, assim para a soma de duas variáveis variáveis independentes $X_1 + Y_1 + X_2 + Y_2 + \cdots + X_n + Y_n$ é mais próxima da distribuição Normal que qualquer uma das variáveis originais

$X_1 + X_2 + \cdots + X_n$ e $Y_1 + Y_2 + \cdots + Y_n$. Portanto, maximizando a não normalidade de **MX** então **Y** é aproximadamente independente. Dessa forma, o procedimento da ICA depende da medida de não Normalidade.

Muitos procedimentos tem sido desenvolvidos para encontrar tais transformações como a abordagem de Hyvarinen & Oja (2000) usando o método de máxima negentropia, Hastie & Tibshirani (2003) usa estimação de produto de densidades, e diversos outros procedimentos (Hyvarinen *et al.*, 2001).

Capítulo 3

Análise Discriminante via Aproximação da Densidade Preditiva Bayesiana

3.1 Introdução

Muitos classificadores na literatura, como os apresentados anteriormente, têm como um dos pressupostos a imposição de um modelo probabilístico paramétrico aos dados, e em situações práticas, essa suposição nem sempre é adequada. Tipicamente, esses modelos envolvem parâmetros desconhecidos, e, então, nos deparamos com um desafio adicional de inferir sobre os valores desses parâmetros. Alguns desses procedimentos inferenciais são extensos e nem sempre triviais. A modelagem por misturas finitas de distribuições, por exemplo, é extremamente flexível provendo uma abordagem de modelagem estatística para uma ampla variedade de fenômenos aleatórios, relaxando a suposição de distribuição aos dados (McLachlan & Peel, 2000). Por outro lado, um número maior de densidades implica numa quantidade maior de parâmetros a serem estimados, consequentemente, alguns procedimentos de estimação tornam-se mais extensos e complicados (veja por exemplos, McLachlan & Peel (2000) e Andrews *et al.* (2010)).

Nesse capítulo, vamos introduzir uma abordagem para *Análise de Discriminantes*(AD), apresentado um classificador baseado em *Máximas Densidades Preditivas*, com o objetivo de contornar o problema da suposição de modelo aos dados. As *Densidades Preditivas* são estimadas usando um método de estimação por função núcleo com enfoque Bayeasiano (Bernardo & Smith (2000); de Lima & Atuncar (2011)).

3.2 Análise de Discriminante via Máxima Densidade Preditiva

Vamos denominar o conjunto de quantidades aleatórias $\mathbf{x}^{(n_g)} = \{\mathbf{x}_1^{(\omega_g)}, \mathbf{x}_2^{(\omega_g)}, \ldots, \mathbf{x}_{n_g}^{(\omega_g)}\}$ como a *informação* relativa a experiência obtida de n_g objetos da classe ω_g, onde $g \in \{1, \ldots, r\}$, e cada $\mathbf{x}_i^{(\omega_g)} = \left(x_{i1}^{(\omega_g)}, \ldots, x_{ip}^{(\omega_g)}\right)', i = 1, 2, \ldots, n_g$. Supondo que cada $\mathbf{x}_i^{(\omega_g)}$ é proveniente de uma população ω_g com distribuição de probabilidade dada por $f_{\omega_g}(\mathbf{x} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, onde Θ é o *espaço paramétrico*. Agora, dado uma nova observação $\mathbf{x}^{novo}$, cuja população a qual é oriunda é desconhecida, podemos definir sua densidade preditiva em termos das *informações* $\left(\mathbf{x}^{(n_g)}\right)$ obtida para cada classe ω_g como:

$$f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) = \int_{\Theta} f_{\omega_g}(\mathbf{x}^{novo} \mid \boldsymbol{\theta}) \pi_{\omega_g}(\boldsymbol{\theta} \mid \mathbf{x}^{(n_g)}) d\boldsymbol{\theta}, \tag{3.1}$$

onde

$$\pi_{\omega_g}(\boldsymbol{\theta} \mid \mathbf{x}^{(n_g)}) \propto f_{\omega_g}(\mathbf{x}^{(n_g)} \mid \boldsymbol{\theta}) \pi_{\omega_g}(\boldsymbol{\theta}) \tag{3.2}$$

é a distribuição a *posteriori* de $\boldsymbol{\theta}$ baseada na informação $\mathbf{x}^{(n_g)}$, com $\pi_{\omega_g}(\boldsymbol{\theta})$ sendo a distribuição a *priori* de $\boldsymbol{\theta}$ na classe ω_g. Assim, obtemos densidades preditivas de $\mathbf{x}^{novo}$ para cada uma das classes $\omega_1, \omega_2, ..., \omega_r$ baseadas em suas respectivas *informações* $\mathbf{x}^{(n_g)}$. Portanto, similarmente aos classificadores de *Bayes* e *Máxima verossimilhança* (Johnson & Wichern, 2007), podemos construir um classificador baseado na *máxima densidade preditiva*, definido como:

Definição 3.2.1 (Classificador baseado em máxima densidade preditiva) *Devemos alocar uma nova observação* $\boldsymbol{x}^{novo}$ *na classe* $\hat{\omega}_g \in \omega_1, \omega_2, ..., \omega_r$, *tal que*

$$\hat{\omega}_g = \underset{\Omega}{\text{argmax}} f_{\hat{\omega}_g}(\boldsymbol{x}^{novo} \mid \boldsymbol{x}^{(n_g)}). \tag{3.3}$$

Diferentemente do *classificador de Bayes*, como apresentado no capítulo anterior, o classificador baseado em *máxima densidade preditiva* não leva em consideração a probabilidade a priori das classes em sua regra. Em muitos problemas reais, a *probabilidade a priori das classes* não pode ser estimada eficientemente(veja Prati *et al.* (2008)).

Na prática, a distribuição preditiva $f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)})$ de cada classe é desconhecida, portanto, devemos estimar essa função. Numa abordagem paramétrica, nós assumimos que a densidade preditiva $f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)})$ pertence a uma família paramétrica de modelos que dependem de certas condições. A eficiência dessa abordagem depende da escolha desse modelo paramétrico. Assim, se o modelo admitido for próximo ou igual ao verdadeiro, as inferências sobre o modelo proveniente serão adequadas, entretanto, se não for, produzirá resultados inadequados levando em classificações errôneas.

Uma abordagem alternativa para determinar $f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)})$ é uma abordagem *não paramétrica*, onde as suposições feitas sobre a estrutura probabilística originária dos dados são fracas ou inexistentes. Um dos métodos não paramétricos mais utilizados para a estimação da densidade preditiva é o *estimador por função núcleo*(ou simplesmente *estimador de núcleo*) (Scott, 1992).

3.3 Estimação da Densidade Preditiva Bayesiana por Função Núcleo

Considere um conjunto de observações das quantidades aleatórias $\mathbf{x}^{(n_g)} = \{\mathbf{x}_1^{(\omega_g)}, \mathbf{x}_2^{(\omega_g}, \ldots, \mathbf{x}_{n_g}^{(\omega_g)}\}$ sendo a *informação* relativa ao experimento. Agora, vamos considerar uma

partição

$$\mathbf{x}^{(n_g)} = \left\{\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_k^{(k)}, \mathbf{y}_1^{(m)}, \ldots, \mathbf{y}_m^{(m)}\right\} = \left\{\mathbf{x}^{(k)}, \mathbf{y}^{(m)}\right\}, \tag{3.4}$$

da informação $\mathbf{x}^{(n_g)}$, onde $\mathbf{x}^{(k)}$ e $\mathbf{y}^{(m)}$ são denominadas *partição de treino e partição de teste*, respectivamente, com $m = n_g - k$. Então o estimador natural para a densidade de $\mathbf{x}^{novo}$ na classe ω_g, considerando apenas a partição de treino $\mathbf{x}^{(k)}$, empregando o estimador por função núcleo (Wand & Jones, 1993), é dada por

$$f_{\omega_g}(\mathbf{x}^{novo} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) = \frac{1}{k}\sum_{i=1}^{k} \mathcal{K}(\mathbf{x}^{novo} \mid \mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g), \tag{3.5}$$

com

$$\mathcal{K}(\mathbf{x}^{novo} \mid \mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g) = |\boldsymbol{\theta}_g|^{-1/2}\, \mathcal{K}\left(\boldsymbol{\theta}^{-1/2}(\mathbf{x}^{novo} - \mathbf{x}_{ig}^{(k)})\right),$$

onde $\mathcal{K}$ é a função núcleo, como definida em (2.32), e $\boldsymbol{\theta}_g$ é a matriz de largura de banda na classe ω_g, é dado por

Agora, empregando a aproximação

$$\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) = \prod_{j=1}^{m}\left\{\frac{1}{k}\sum_{i=1}^{k} \mathcal{K}(\mathbf{y}_{jg}^{(m)} \mid \mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g)\right\}, \tag{3.6}$$

de tal forma que a densidade preditiva para a nova observação $\mathbf{x}^{novo}$ na classe ω_g seja aproximada por

$$\begin{aligned}
\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) &= \int_{\Theta} f_{\omega_g}(\mathbf{x}^{novo} \mid \boldsymbol{\theta}_g)\pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)})d\boldsymbol{\theta}_g \\
&\approx \int_{\Theta} \frac{1}{k}\sum_{i=1}^{k} \mathcal{K}(\mathbf{x}^{novo} \mid \mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g)\pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)})d\boldsymbol{\theta}_g \\
&= \frac{1}{k}\sum_{i=1}^{k}\int_{\Theta} \mathcal{K}(\mathbf{x}^{novo} \mid \mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g)\pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)})d\boldsymbol{\theta}_g,
\end{aligned} \tag{3.7}$$

representando o valor médio de k funções núcleo integradas com respeito a distribuição a posteriori de $\boldsymbol{\theta}_g$. Integrando com respeito a matriz de largura de banda $\boldsymbol{\theta}_g$, eliminamos

a necessidade de empregamos procedimentos de estimação dessa matriz, bem como sua influência no processo de classificação. Também, se consideramos diferentes Q partições de $\mathbf{x}_q^{(n_g)} = \left\{\mathbf{x}_q^{(k)}, \mathbf{y}_q^{(m)}\right\}$, podemos empregar o método de *Validação Cruzada "Q-fold"* (do inglês, Q-fold Cross-Validation) (Efron, 1983) e obtermos

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \approx \frac{1}{Q}\sum_{q=1}^{Q} \hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}_q^{(n_g)}). \tag{3.8}$$

3.4 Estimação da Densidade Preditiva Bayesiana por Produto de Funções Núcleo Normais

Nessa seção vamos considerar um procedimento que trata as componentes do vetor de características como sendo independentes. Assumimos que o modelo preditivo de $\mathbf{x}^{novo}$ pode ser aproximado por uma mistura de função núcleo, uma vez que isso não implique em desacordo com a teoria de probabilidade(West (1991); Bernardo (1999)). Assim, utilizamos o estimador preditivo bayesiano por função núcleo como um produto de funções núcleo. Na prática, o produto de funções núcleo univariado é recomendado por simplicidade e afim de evitar o problema da alta dimensionalidade (Scott, 1992).

Empregando (3.7), o produto de função núcleo é obtido fazendo $\boldsymbol{\theta}_g = diag(\theta_1, \theta_2, \ldots, \theta_p)$, neste caso teremos o seguinte estimador para as densidades preditivas:

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \boldsymbol{\theta}_g) = \frac{1}{k}\sum_{j=1}^{k}\prod_{l=1}^{p} \theta_{lg}^{-1/2}\mathcal{K}\left(\frac{(x_l^{novo} - x_{ilg}^{(k)})}{\theta_{lg}}\right). \tag{3.9}$$

Considerando, a função núcleo $\mathcal{K}$ como sendo a função núcleo Normal (veja 2.33), temos

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \boldsymbol{\theta}_g) \approx \frac{1}{k}\sum_{j=1}^{k}\prod_{l=1}^{p} \theta_{lg}^{-1/2}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\theta_{lg}}(x_l^{novo} - x_{ilg}^{(k)})^2\right\}, \tag{3.10}$$

onde θ_{lg} é a l-ésima componente da diagonal principal da matriz de largura de banda $\boldsymbol{\theta}_g$ na classe ω_g. Agora, podemos obter a densidade a posteriori de $\boldsymbol{\theta}_g$, $\pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)}) \propto$

$f_{\omega_g}(\mathbf{x}^{(n_g)} \mid \boldsymbol{\theta}_g)\pi_{\omega_g}(\boldsymbol{\theta}_g)$, admitindo que $\mathcal{K}$ é o produto de funções núcleo Normal. Então, substituindo o produto de funções núcleo Normal na expressão (3.6), temos

$$\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) = \prod_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} \prod_{l=1}^{p} \theta_{lg}^{-1/2} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\theta_{lg}} (y_{jlg}^{(m)} - x_{ilg}^{(k)})^2 \right\}. \quad (3.11)$$

Assim, considerando

$$\delta_{ijlg} = (y_{jlg}^{(m)} - x_{ilg}^{(k)})^2, \quad a_{jg} = \sum_{i=1}^{k} \prod_{l=1}^{p} \delta_{ijlg}^{-1/2}, \quad w_{ijg} = \frac{\prod_{l=1}^{p} \delta_{ijlg}^{-1/2}}{\sum_{i=1}^{k} \prod_{l=1}^{p} \delta_{ijlg}^{-1/2}} \quad e \quad \sum_{i=1}^{k} w_{ijg} = 1,$$

vamos reescrever (3.11) como produto de misturas de densidades,

$$\begin{aligned}
\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) &= \prod_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} \prod_{l=1}^{p} \frac{\theta_{lg}^{-1/2} \theta_{lg}^{-1+1}}{\sqrt{2}\ \Gamma(1/2)} \exp\left\{ -\frac{1}{2\theta_{lg}} \delta_{ijlg} \right\} \\
&= \prod_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} \prod_{l=1}^{p} \theta_{lg} \frac{\theta_{lg}^{-1/2-1}}{\sqrt{2}\ \Gamma(1/2)} \exp\left\{ -\frac{\delta_{ijlg}}{2} \theta_{lg}^{-1} \right\} \\
&= \prod_{j=1}^{m} \frac{1}{k} \prod_{l=1}^{p} \theta_{lg} \left\{ \sum_{i=1}^{k} \prod_{l=1}^{p} \frac{(\delta_{ijlg}/2)^{-1/2}}{(\delta_{ijlg}/2)^{-1/2}} \frac{\theta_{lg}^{-1/2-1}}{\sqrt{2}\ \Gamma(1/2)} \exp\left\{ -\frac{\delta_{ijlg}}{2} \theta_{lg}^{-1} \right\} \right\} \\
&= \prod_{j=1}^{m} \frac{1}{k} \prod_{l=1}^{p} \theta_{lg} \left\{ \sum_{i=1}^{k} \prod_{l=1}^{p} \frac{\delta_{ijlg}^{-1/2}}{\sqrt{2}\ 2^{-1/2}} \prod_{l=1}^{p} \frac{\theta_{lg}^{-1/2-1}}{(\delta_{ijlg}/2)^{-1/2} \Gamma(1/2)} \exp\left\{ -\frac{\delta_{ijlg}}{2} \theta_{lg}^{-1} \right\} \right\} \\
&= \prod_{j=1}^{m} \frac{1}{k} \prod_{l=1}^{p} \theta_{lg} \left\{ \frac{a_{jg}}{a_{jg}} \sum_{i=1}^{k} \prod_{l=1}^{p} \delta_{ijlg}^{-1/2} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2} \right) \right\} \\
&= \prod_{j=1}^{m} \frac{1}{k} \prod_{l=1}^{p} \theta_{lg} \left\{ a_{jg} \sum_{i=1}^{k} \frac{\prod_{l=1}^{p} \delta_{ijlg}^{-1/2}}{a_{jg}} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2} \right) \right\} \\
&\propto \prod_{j=1}^{m} \prod_{l=1}^{p} \theta_{lg} \left\{ \sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2} \right) \right\}, \quad (3.12)
\end{aligned}$$

onde IG é a densidade Gama Inversa multivariada independente, ou seja, é o produto de Gama Inversa univariada, com densidades dada pela Definição (A.2.3) em anexo, dado por:

$$IG(\boldsymbol{\theta}_g \mid \boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) = \prod_{l=1}^{p} \frac{\beta_{lg}^{\alpha_{lg}}}{\Gamma(\alpha_{lg})} \theta_{lg}^{-(\alpha_{lg}+1)} \exp\left\{ -\frac{\beta_{lg}}{\theta_{lg}} \right\}, \quad (3.13)$$

onde $\boldsymbol{\alpha}_g' = (\alpha_{1g}, \alpha_{2g}, \ldots, \alpha_{pg})$ e $\boldsymbol{\beta}_g' = (\beta_{1g}, \beta_{2g}, \ldots, \beta_{pg})$.

3.4.1 Aproximando misturas de densidades

A técnica de aproximação, ou colapso, de misturas de densidades é fundamental para muitas aplicações de multi-processo (West & Harrison, 1997). A mistura de densidades é aproximada por uma densidade $P^*(\boldsymbol{\theta}^*)$, onde $\boldsymbol{\theta}^*$ é um vetor de parâmetros, de forma funcional especificada, onde os parâmetros definem a aproximação a ser escolhida. Por exemplo, uma mistura de Normais pode ser aproximada por uma única densidade Normal. Então, o que é necessário é um critério que indique qual a densidade a ser empregada para representar a mistura satisfazendo um nível de aproximação desejável. Assim, o *critério da divergência logarítmica (Kullback-Leibler)*, apesar de não ser uma medida de distância, apresenta propriedades satisfatórias para ser empregada como uma medida de aproximação de densidade (West & Harrison, 1997). De acordo com o critério da divergência logarítmica, definimos como a divergência entre a densidade aproximada $P^*(\boldsymbol{\theta}^*)$ e a verdadeira densidade $P(\boldsymbol{\theta})$, onde $\boldsymbol{\theta}$ é um vetor de parâmetros, como

$$\delta(\boldsymbol{\theta}^*) \;=\; \int_{\Theta} P(\boldsymbol{\theta}) log\left(\frac{P(\boldsymbol{\theta})}{P^*(\boldsymbol{\theta}^*)}\right) d\boldsymbol{\theta}. \tag{3.14}$$

A propriedade do critério da divergência logarítmica que nos fundamental nas aplicações, é que $\delta(\theta^*) = 0$ se, somente se, $P^*(\theta^*) = P(\theta)$. Então, minimizando $\delta(\theta^*)$ obtemos "uma boa"aproximação de $P(\theta)$ por $P^*(\theta^*)$.

Teorema 3.4.1 *Seja* $f(\boldsymbol{\theta}) = \sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right)$ *uma mistura de distribuições Gamas Inversa, usando o critério da divergência logarítmica (Kullback-Leibler), a melhor aproximação de* $f(\boldsymbol{\theta})$ *por uma densidade* $IG(\boldsymbol{\theta}_g \mid \boldsymbol{\alpha}_g, \boldsymbol{\beta}_g)$ *é dada quando*

$$\alpha_{jlg} \;=\; \frac{1}{2}\left\{1 + log\,\frac{\exp\left\{\sum_{i=1}^{k} w_{ijg} log\,\delta_{ijlg}/2\right\}}{\left[2\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right]^{-1}}\right\}^{-1} \quad e \tag{3.15}$$

$$\beta_{jlg} \;=\; 2\;\alpha_{lg}\left\{\left[2\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right]^{-1}\right\},\; l = 1, \ldots, p \;\; e \;\; j = 1, \ldots, m. \tag{3.16}$$

Demonstração: Da definição do *critério da divergência logarítmica (Kullback-Leibler)* em (3.14), temos

$$\begin{aligned}
\delta(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) &= \int_\Theta f(\boldsymbol{\theta}) log\left(\frac{f(\boldsymbol{\theta})}{IG(\boldsymbol{\theta}_g \mid \boldsymbol{\alpha}_g, \boldsymbol{\beta}_g)}\right) d\theta \\
&= \int_\Theta [log f(\boldsymbol{\theta})] f(\boldsymbol{\theta}) d\theta - \int_\Theta log\, IG(\boldsymbol{\theta}_g \mid \boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) f(\boldsymbol{\theta}) d\theta \\
&= E[log f(\boldsymbol{\theta})] - \int_\Theta log\left\{\prod_{l=1}^{p} \frac{\beta_{lg}^{\alpha_{lg}}}{\Gamma(\alpha_{lg})} \theta_{lg}^{-(\alpha_{lg}+1)} \exp\left\{-\frac{\beta_{lg}}{\theta_{lg}}\right\}\right\} f(\boldsymbol{\theta}) d\theta \\
&= c - \int_\Theta \sum_{l=1}^{p} \left[\alpha_{lg} log\, \beta_{lg} - log\, \Gamma(\alpha_{lg}) - (\alpha_{lg}+1) log\, \theta_{lg} - \frac{\beta_{lg}}{\theta_{lg}}\right] f(\boldsymbol{\theta}) d\theta \\
&= c - \sum_{l=1}^{p} \int_\Theta \left[\alpha_{lg} log\, \beta_{lg} - log\, \Gamma(\alpha_{lg}) - (\alpha_{lg}+1) log\, \theta_{lg} - \frac{\beta_{lg}}{\theta_{lg}}\right] f(\boldsymbol{\theta}) d\theta \\
&= c - \sum_{l=1}^{p} \left[\alpha_{lg} log\, \beta_{lg} - log\, \Gamma(\alpha_{lg}) - (\alpha_{lg}+1) E(log\, \theta_{lg}) - \beta_{lg} E(\theta_{lg}^{-1})\right], \qquad (3.17)
\end{aligned}$$

onde $c = E[log f(\boldsymbol{\theta})]$. Portanto, para minimizar a expressão (3.17), devemos solucionar o sistema de derivadas parciais simultaneamente,

$$\frac{\partial}{\partial \boldsymbol{\alpha}_g} \delta(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) = 0 \; e \; \frac{\partial}{\partial \boldsymbol{\beta}_g} \delta(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) = 0.$$

Assim,

$$\begin{aligned}
\frac{\partial}{\partial \alpha_{lg}} \delta(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) &= 0, \text{ implica em} \\
-log\, \beta_{lg} + \psi(\alpha_{lg}) + E[log\, \theta_{lg}] &= 0, \text{ logo} \\
E[log\, \theta_{lg}] &= log\, \beta_{lg} - \psi(\alpha_{lg}), \quad l = 1, \ldots, p,
\end{aligned}$$

onde $\psi(z) = \frac{\partial log \Gamma(z)}{\partial z}$ é a função *Digamma*. E

$$\begin{aligned}
\frac{\partial}{\partial \beta_{lg}} \delta(\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g) &= 0, \text{ então} \\
-\frac{\alpha_{lg}}{\beta_{lg}} - E(\theta_{lg}^{-1}) &= 0, \text{ implica em} \\
E(\theta_{lg}^{-1}) &= \frac{\alpha_{lg}}{\beta_{lg}}, \quad l = 1, \ldots, p.
\end{aligned}$$

Portanto, a expressão (3.17) é minimizada se, somente se,

$$E[log\,\theta_{lg}] = log\,\beta_{lg} - \psi(\alpha_{lg}) \;\; e \;\; E(\theta_{lg}^{-1}) = \frac{\alpha_{lg}}{\beta_{lg}}. \tag{3.18}$$

Como $\boldsymbol{\theta}_g \sim \sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right)$ e usando as propriedades da distribuição Gama Inversa em (A.2.3) e (A.2.4) em anexo, sendo $\boldsymbol{\theta}_g^* = (\theta_{1,g}, \ldots, \theta_{l-1,g}, \theta_{l+1,g}, \ldots, \theta_{p,g})$ e Θ^* é o espaço gerado por $\boldsymbol{\theta}_g^*$, temos que

$$\begin{aligned}
E[log\,\theta_{lg}] &= \int_0^\infty log\,\theta_{lg} \left[\int_{\Theta^*} \sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right) d\boldsymbol{\theta}_g^*\right] d\theta_{lg} \\
&= \int_0^\infty log\,\theta_{lg} \left[\sum_{i=1}^{k} w_{ijg} \int_{\Theta^*} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right) d\boldsymbol{\theta}_g^*\right] d\theta_{lg} \\
&= \int_0^\infty log\,\theta_{lg} \sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_{lg} \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right) d\theta_{lg} \\
&= \sum_{i=1}^{k} w_{ijg} \int_0^\infty log\,\theta_{lg} IG\left(\boldsymbol{\theta}_{lg} \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right) d\theta_{lg} \\
&= \sum_{i=1}^{k} w_{ijg} E[log\,\theta_{lg}], \text{ pela propriedade (A.2.3)}, \\
&= \sum_{i=1}^{k} w_{ijg} (log\,\delta_{ijlg}/2 - \psi(1/2)),
\end{aligned} \tag{3.19}$$

e ainda,

$$\begin{aligned}
E[\theta_{lg}^{-1}] &= \int_0^\infty log\,\theta_{lg} \left[\int_{\Theta^*} \sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right) d\boldsymbol{\theta}_g^*\right] d\theta_{lg} \\
&= \sum_{i=1}^{k} w_{ijg} \int_0^\infty \theta_{lg}^{-1} IG\left(\boldsymbol{\theta}_{lg} \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right) d\theta_{lg} \\
&= \sum_{i=1}^{k} w_{ijg} E[\theta_{lg}^{-1}], \text{ pela propriedade (A.2.4)}, \\
&= \sum_{i=1}^{k} w_{ijg} 1/2(\delta_{ijlg}/2)^{-1} \\
&= \sum_{i=1}^{k} w_{ijg}\, \delta_{ijlg}^{-1}.
\end{aligned} \tag{3.20}$$

Voltando ao sistema de equações,

$$\begin{aligned} E(\theta_{lg}^{-1}) = \frac{\alpha_{lg}}{\beta_{lg}} &\Rightarrow 2\left[\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right] = 2\frac{\alpha_{lg}}{\beta_{lg}} \\ &\Rightarrow \beta_{lg} = 2\alpha_{lg}\left[2\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right]^{-1} \\ &\Rightarrow \beta_{lg} = 2\alpha_{lg}\beta_{jlg}^{(1)}, \end{aligned} \tag{3.21}$$

onde $\beta_{jlg}^{(1)} = \left[2\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right]^{-1}$, e

$$\begin{aligned} E[log\,\theta_{lg}] &= log\,\beta_{lg} - \psi(\alpha_{lg}) \\ \sum_{i=1}^{k} w_{ijg}(log\,\delta_{ijlg}/2 - \psi(1/2)) &= log\,\beta_{lg} - \psi(\alpha_{lg}) \\ \sum_{i=1}^{k} w_{ijg} log\,\delta_{ijlg}/2 - \psi(1/2)\sum_{i=1}^{k} w_{ijg} &= log\,\beta_{lg} - \psi(\alpha_{lg}) \\ -\sum_{i=1}^{k} w_{ijg} log\,\delta_{ijlg}/2 - \psi(\alpha_{lg}) &= -\psi(1/2) - log\,\beta_{lg} \\ -log\,\exp\left\{\sum_{i=1}^{k} w_{ijg} log\,\delta_{ijlg}/2\right\} - \psi(\alpha_{lg}) &= -\psi(1/2) - log(2\alpha_{lg}\beta_{jlg}^{(1)}) \\ -log\,\beta_{jlg}^{(0)} - \psi(\alpha_{lg}) &= -\psi(1/2) - log\,2 - log\,\alpha_{lg} - log\,\beta_{jlg}^{(1)} \\ log\,\alpha_{lg} - \psi(\alpha_{lg}) &= log\,(1/2) - \psi(1/2) + log\,\frac{\beta_{jlg}^{(0)}}{\beta_{jlg}^{(1)}}, \end{aligned} \tag{3.22}$$

onde $\beta_{jlg}^{(0)} = \exp\left\{\sum_{i=1}^{k} w_{ijg} log\,\delta_{ijlg}/2\right\}$. Assim a melhor aproximação da mistura de Gamas Inversas por um densidade Gama Inversa, segundo o critério da divergência logarítmica, é dada pela solução do sistema de equações não lineares

$$\begin{cases} log\,\alpha_{lg} - \psi(\alpha_{lg}) = log\,(1/2) - \psi(1/2) + log\,\frac{\beta_{jlg}^{(0)}}{\beta_{jlg}^{(1)}} \\ \beta_{lg} = 2\alpha_{lg}\beta_{jlg}^{(1)} \end{cases}$$

Uma aproximação para a solução do sistema de equações acima pode ser obtida empregando a aproximação de Stirling para a função digamma, $log(y) - \psi(y) = (2y)^{-1}$, levando

a

$$\alpha_{lg} = \frac{1}{2}\left\{1 + log\,\frac{\beta_{jlg}^{(0)}}{\beta_{jlg}^{(1)}}\right\}^{-1} \tag{3.23}$$

$$\beta_{lg} = 2\;\alpha_{lg}\;\beta_{jlg}^{(1)}. \tag{3.24}$$

Ficando demonstrado o Teorema. □

Pelo Teorema 3.4.1, temos então que a melhor aproximação, pelo critério da divergência logarítmica, para a mistura de densidades $\sum_{i=1}^{k} w_{ijg} IG\left(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}\right)$ é

$$\sum_{i=1}^{k} w_{ijg} IG(\boldsymbol{\theta}_g \mid \frac{1}{2}, \frac{\delta_{ijlg}}{2}) \approx IG(\boldsymbol{\theta}_g \mid \alpha_{jlg}, \beta_{jlg}), \tag{3.25}$$

onde

$$\alpha_{jlg} = \frac{1}{2}\left\{1 + log\,\frac{\exp\left\{\sum_{i=1}^{k} w_{ijg} log\,\delta_{ijlg}/2\right\}}{\left[2\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right]^{-1}}\right\}^{-1} \quad e \tag{3.26}$$

$$\beta_{jlg} = 2\;\alpha_{lg}\left\{\left[2\sum_{i=1}^{k} w_{ijg}\delta_{ijlg}^{-1}\right]^{-1}\right\}. \tag{3.27}$$

Agora substituindo (3.25) em (3.12), teremos

$$\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) \;\propto\; \prod_{j=1}^{m}\prod_{l=1}^{p} \theta_{lg}\; IG(\boldsymbol{\theta}_g \mid \alpha_{jlg}, \beta_{jlg}). \tag{3.28}$$

Para um θ_{lg} particular temos a priori não informativa $\hat{\pi}_{\omega_g}(\theta_{lg}) \propto \theta_{lg}^{-1}$. Como as componentes do vetor $\boldsymbol{\theta}_g$ são independentes, então

$$\hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g) \propto \prod_{l=1}^{p} \theta_{lg}^{-1} \tag{3.29}$$

é a priori independente de Jefreys para $\boldsymbol{\theta}_g$ (Sun & Berger, 2006). Conjugando a verossi-

milhança de $\boldsymbol{\theta}_g$, $\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)})$, com a priori $\hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g)$, temos

$$\begin{aligned}
\hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)}) &\propto \prod_{j=1}^{m}\prod_{l=1}^{p} \theta_{lg} IG(\boldsymbol{\theta}_g \mid \alpha_{jlg}, \beta_{jlg}) \;\; \theta_{lg}^{-1} \\
&= \prod_{j=1}^{m}\prod_{l=1}^{p} \theta_{lg} \frac{\beta_{jlg}^{\alpha_{jlg}}}{\Gamma(\alpha_{jlg})} \theta_{lg}^{-(\alpha_{jlg}+1)} \exp\left\{-\frac{\beta_{jlg}}{\theta_{lg}}\right\} \;\; \theta_{lg}^{-1} \\
&= \prod_{l=1}^{p}\prod_{j=1}^{m} \frac{\beta_{jlg}^{\alpha_{jlg}}}{\Gamma(\alpha_{jlg})} \theta_{lg}^{-(\alpha_{jlg}+1)} \exp\left\{-\frac{\beta_{jlg}}{\theta_{lg}}\right\} \\
&\propto \prod_{l=1}^{p} \theta_{lg}^{-1} \theta_{lg}^{-m\bar{\alpha}_{lg}} \exp\left\{-\frac{m\bar{\beta}_{lg}}{\theta_{lg}}\right\}, \qquad (3.30)
\end{aligned}$$

que identificamos como sendo o núcleo de uma densidade Gamma Inversa, assim obtemos a distribuição a posteriori de $\boldsymbol{\theta}_g$,

$$\pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)}) \approx IG(\boldsymbol{\theta}_g \mid m\bar{\alpha}_{lg}, m\bar{\beta}_{lg}) \qquad (3.31)$$

com

$$\bar{\alpha}_{lg} = \frac{1}{m}\sum_{j=1}^{m} \alpha_{jlg} \; e \; \bar{\beta}_{lg} = \frac{1}{m}\sum_{j=1}^{m} \beta_g.$$

Agora, usando a distribuição a posteriori de $\boldsymbol{\theta}_g$ em (3.7) a *densidade preditiva* de uma nova observação $\mathbf{x}^{novo}$ dada a *Informação* $\mathbf{x}^{(n_g)}$ pode ser aproximada por

$$\begin{aligned}
\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) &\approx \frac{1}{k}\sum_{i=1}^{k} \int_{\Theta} |\boldsymbol{\theta}_g|^{-1/2} \mathcal{K}(\boldsymbol{\theta}_g^{-1/2}(\mathbf{x}^{novo} - x_{ilg}^{(k)})) \pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)}) d\boldsymbol{\theta}_g \\
&\propto \frac{1}{k}\sum_{i=1}^{k} \int_{\Theta} \prod_{l=1}^{p} \theta_{lg}^{-1/2} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\theta_{lg}}(x_l^{novo} - x_{ilg}^{(k)})^2\right\} \\
&\quad \times \frac{m\bar{\beta}_{lg}^{m\bar{\alpha}_{lg}}}{\Gamma(m\bar{\alpha}_{lg})} \theta_{lg}^{-(m\bar{\alpha}_{lg}+1)} \exp\left\{-\frac{m\bar{\beta}_{lg}}{\theta_{lg}}\right\} d\boldsymbol{\theta}_g. \qquad (3.32)
\end{aligned}$$

Pelo Corolário A.3.2 em anexo, a conjugação das densidades de $\mathbf{X}^{novo} \mid \theta_g \sim Normal(x_{ilg}^{(k)}, \theta_g)$

com $\theta_g \sim IG(m\bar{\alpha}_{lg}, m\bar{\beta}_{lg})$, na expressão (3.32), produz

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \approx \frac{1}{k}\sum_{i=1}^{k}\prod_{l=1}^{p} t\left(x_l^{novo} \mid x_{ilg}^{(k)}, \sqrt{\bar{\beta}_{lg}/\bar{\alpha}_{lg}}, \ 2m\bar{\alpha}_{lg}\right).$$

onde $t(\cdot)$ denota a distribuição t-Student com parâmetros de *locação* $x_{ilg(q)}^{(k)}$, *escala* $\sqrt{\frac{\bar{\beta}_{lg(q)}}{\bar{\alpha}_{lg(q)}}}$ e *graus de liberdade* $2m\bar{\alpha}_{lg(q)}$. Substituindo (3.33) em (3.8), considerando as Q partições da *Informação* $\mathbf{x}^{(n_g)}$, temos:

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \approx \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{k}\sum_{i=1}^{k}\prod_{l=1}^{p} t\left(x_l^{novo} \mid x_{ilg(q)}^{(k)}, \sqrt{\frac{\bar{\beta}_{lg(q)}}{\bar{\alpha}_{lg(q)}}}, \ 2m\bar{\alpha}_{lg(q)}\right). \quad (3.33)$$

Empregando (3.33) na Definição 3.2.1, obtemos então uma regra para Análise Discriminante empregando essas distribuições preditivas aproximadas.

3.5 Estimação da Densidade Preditiva Bayesiana por Produto de Funções Núcleo Normais empregando Análise de Componentes Independentes

Assim como no caso do classificador *Naive Bayes*, nem sempre a suposição de independência nos dados é adequada. Então, uma alternativa é usar a *Análise de Componentes Independentes*, afim de produzir vetores de características transformados, cuja as componentes são aproximadamente independentes. Assim temos uma transformação $\mathbf{Y} = \mathbf{MX}$, onde $\mathbf{X}$ é a amostra observada, $\mathbf{M}$ é uma matriz de transformação e $\mathbf{Y}$ são as *Componentes Independentes*. O objetivo então é encontrar uma matriz $\hat{\mathbf{M}}$ que seja uma matriz de transformação ótima, esse procedimento de encontrar $\hat{\mathbf{M}}$ pode ser realizado por diversas abordagens descritas na literatura(*ver* Hyvarinen & Oja (2000), Hyvarinen *et al.* (2001), Hastie & Tibshirani (2003) e Stone (2004)).

No contexto de Análise de Discriminante, como agora temos as componentes de $\mathbf{Y}$ aproximadamente independentes, a densidade da distribuição de probabilidade de $\mathbf{Y}$ é aproximadamente dada pelo produto das densidades marginais. Então o procedimento de classificação pode ser resumido pelo seguinte algoritmo de Amato *et al.* (2003):

1. Para cada classe ω_g, $g = 1, \ldots, r$, usamos a amostra $\{\mathbf{x}_1^{(\omega_g)}, \mathbf{x}_2^{(\omega_g)}, \ldots, \mathbf{x}_{n_g}^{(\omega_g)}\}$ de tamanho n_g da classe ω_g para estimar a média $\boldsymbol{\mu}_g$ da classe ω_g, e usamos essa estimativa para centralizar toda a informação $\mathbf{x}^{(n_g)}$ em relação a essa classe. Então, usamos o algoritmo ICA (Hyvarinen & Oja, 2000) nos dados centralizados afim de obter a matriz de transformação ótima $\hat{\mathbf{M}}_{\omega_g}$.

2. Usando a matriz $\hat{M}_{\omega_g}$, calculamos a amostra, centrada em $\boldsymbol{\mu}_g$, transformada por $\mathbf{Y} = \hat{\mathbf{M}}_{\omega_g}\mathbf{X}_i,\ i = 1, \ldots, n$.

3. Para essa nova observação $\mathbf{x}^{novo}$ calculamos para cada classe $\omega_g; \forall g\ :\ 1, \ldots, r$ a estimativa da densidade preditiva por:

$$f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \ \propto\ \frac{1}{k}\sum_{i=1}^{k}\prod_{l=1}^{p} t\left(\hat{\mathbf{M}}_g\, x_l^{novo} \mid \hat{\mathbf{M}}_g x_{ilg}^{(k)}, \sqrt{\frac{\bar{\beta}_{lg}}{\bar{\alpha}_{lg}}},\ 2\bar{\alpha}_{lg}\right) \mid det(\hat{\mathbf{M}}_g) \mid . \quad (3.34)$$

Agora, substituindo (3.34) em (3.8), considerando as Q partições da *Informação* $\mathbf{x}^{(n_g)}$, temos:

$$f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \ \approx\ \frac{1}{Qk}\sum_{q=1}^{Q}\sum_{i=1}^{k}\prod_{l=1}^{p} t\left(\hat{\mathbf{M}}_g x_j^{novo} \mid \hat{\mathbf{M}}_g x_{ilg(q)}^{(k)}, \sqrt{\frac{\bar{\beta}_{lg(q)}}{\bar{\alpha}_{lg(q)}}},\ 2\bar{\alpha}_{lg(q)}\right) \mid det(\hat{\mathbf{M}}_g) \mid \quad (3.35)$$

Empregando (3.35) na Definição 3.2.1, obtemos então uma regra para Análise Discriminante empregando essa distribuições preditivas aproximadas.

3.6 Estimação da Densidade Preditiva Bayesiana por Função Núcleo Multivariada Normal

Em geral, tanto a suposição de independência como a transformação por *Componentes Independentes* são inadequadas, num ponto de vista em que as correlações entre as

componentes do vetor de características são importantes. Assim, vamos apresentar um procedimento para o caso geral, que nada mais é do que a extensão das abordagens anteriores para o caso multivariado. Semelhantemente aos casos anteriores, vamos ainda considerar a *Informação* $\mathbf{x}^{(n_g)} = \{\mathbf{x}_1^{(\omega_g)}, \mathbf{x}_2^{(\omega_g)}, \ldots, \mathbf{x}_{n_g}^{(\omega_g)}\}$, e suas partições $\{\mathbf{x}^{(k)}\}$ e $\{\mathbf{y}^{(m)}\}$. Para obter $f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)})$, vamos novamente considerar o estimador por função núcleo, mas em vez de um produto de funções núcleo, vamos usar o estimador de núcleo multivariado (Wand & Jones, 1993):

$$\mathcal{K}(\mathbf{x}^{novo} \mid \mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g) = \frac{1}{k} \sum_{i=1}^{k} |\boldsymbol{\theta}_g|^{-1/2} \, \mathcal{K}\left(\boldsymbol{\theta}_g^{-1/2}(\mathbf{x}^{novo} - \mathbf{x}_{ig}^{(k)})\right), \tag{3.36}$$

onde $\boldsymbol{\theta}_g$ é a *matriz de largura de banda*. Usando (3.7) para obter a densidade preditiva $f_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)})$, então o estimador de núcleo de $\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \boldsymbol{\theta}_g)$, considerando a *Função Núcleo Normal Multivariada* (Scott, 1992), é dado por:

$$\frac{1}{k} \sum_{i=1}^{k} |\boldsymbol{\theta}_g|^{-1/2} \, \frac{1}{(2\pi)^{p/2}} \exp\{-\frac{1}{2}(\mathbf{x}^{novo} - \mathbf{x}_{ig}^{(k)})' \boldsymbol{\theta}_g^{-1} (\mathbf{x}^{novo} - \mathbf{x}_{ig}^{(k)})\}. \tag{3.37}$$

Para obter $\hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)}) \propto \hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) \, \hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(k)})$, vamos considerar novamente o Núcleo Normal Multivariado para aproximar a *função de verossimilhança*

empírica de $\boldsymbol{\theta}_g$. Assim,

$$\begin{aligned}
\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) &= \prod_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} |\boldsymbol{\theta}_g|^{-1/2} \frac{1}{(2\pi)^{p/2}} \exp\{-\frac{1}{2}(\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})' \boldsymbol{\theta}_g^{-1} (\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})\} \\
&= \prod_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} |\boldsymbol{\theta}_g|^{-1/2-p+p} \frac{1}{(2\pi)^{p/2}} \exp\{-\frac{1}{2} tr\left[(\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})(\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})' \boldsymbol{\theta}_g^{-1}\right]\} \\
&= |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \frac{1}{k} \sum_{i=1}^{k} |\boldsymbol{\theta}_g|^{-1/2-p} \frac{|\Delta_{ijg}|^{p/2-p/2}}{(2\pi)^{p/2}} \exp\{-\frac{1}{2} tr\left[\Delta_{ijg} \boldsymbol{\theta}_g^{-1}\right]\} \\
&= |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \frac{2^p \Gamma(p/2)}{k\,(2\pi)^{p/2}} \sum_{i=1}^{k} |\boldsymbol{\theta}_g|^{-1/2-p} \frac{|\Delta_{ijg}|^{p/2-p/2}}{2^p\,\Gamma(p/2)} \exp\{-\frac{1}{2} tr\left[\Delta_{ijg} \boldsymbol{\theta}_g^{-1}\right]\} \\
&= |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \frac{2^p \Gamma(p/2)}{k\,(2\pi)^{p/2}} \sum_{i=1}^{k} \frac{|\Delta_{ijg}|^{p/2-p/2}}{2^p\,\Gamma(p/2)} |\boldsymbol{\theta}_g|^{-1/2-p} \exp\{-\frac{1}{2} tr\left[\Delta_{ijg} \boldsymbol{\theta}_g^{-1}\right]\} \\
&\propto |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \sum_{i=1}^{k} |\Delta_{ijg}|^{-p/2} \frac{|\Delta_{ijg}|^{p/2}}{2^p\,\Gamma(p/2)} |\boldsymbol{\theta}_g|^{-p-1/2} \exp\{-\frac{1}{2} tr\left[\Delta_{ijg} \boldsymbol{\theta}_g^{-1}\right]\} \\
&= |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \left[\sum_{i=1}^{m} |\Delta_{ijg}|^{-p/2}\right] \sum_{i=1}^{k} \frac{|\Delta_{ijg}|^{-p/2}}{\sum_{i=1}^{m} |\Delta_{ijg}|^{-p/2}} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg}) \\
&\propto |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \sum_{i=1}^{k} \omega_{ijg} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg}), \qquad (3.38)
\end{aligned}$$

onde $\mathcal{WI}_p$ é a densidade Wishart Inversa com p graus de liberdade e parâmetro de escala $\Delta_{ijg} = (\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})(\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})'$, com densidade dada na Definição (A.2.2) em anexo, e

$$\omega_{ijc} = \frac{|\Delta_{ijg}|^{-p/2}}{\sum_{i=1}^{m} |\Delta_{ijg}|^{-p/2}} \qquad (3.39)$$

Agora usaremos novamente a ideia de aproximar misturas de densidades por um única densidade, empregando outra vez o *critério da divergência logarítmica (Kullback-Leibler)* como uma medida de divergência entre as densidades.

Teorema 3.6.1 *Seja* $f(\boldsymbol{\theta}_g) = \sum_{i=1}^{k} \omega_{ijg} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg})$ *uma mistura de densidades Wishart Inversa. Usando o critério da divergência logaritmica (Kullback-Leibler) temos que a melhor aproximação para a mistura* $f(\boldsymbol{\theta}_g)$ *é*

$$WI_p(\boldsymbol{\theta}_g \mid d_{jg}^*, \boldsymbol{B}_{jg}), \qquad (3.40)$$

onde $\boldsymbol{B}_{jg} = \frac{d^*_{jg}}{p}\left(\sum_{i=1}^{k}\omega_{ijg}\Delta_{ijg}^{-1}\right)^{-1}$ *e* d^*_{jg} *é o valor de* d_{jg} *que satisfaz*

$$log\,|\boldsymbol{B}_{jg}| - \sum_{v=1}^{p}\Psi\left(\frac{d_{jg}-v+1}{2}\right) \;=\; \sum_{i=1}^{k}\omega_{ijg}log\,|\Delta_{ijg}| - \sum_{v=1}^{p}\Psi\left(\frac{v}{2}\right).$$

Demonstração: Devemos encontrar valores para d^*_{jg} e $\mathbf{B}_{jg}$, tal que a densidade $WI_p(\boldsymbol{\theta}_g \mid d^*_{jg}, \mathbf{B}_{jg})$ seja a melhor aproximação para a mistura de Wishart Inversa $f(\boldsymbol{\theta}_g)$, minimizando o critério da divergência logarítmica, definida em (3.14), pela solução simultânea do sistema de derivadas parciais

$$\frac{\partial}{\partial \mathbf{B}_{jg}}\delta(d^*_{jg}, \mathbf{B}_{jg}) = 0 \;\; e \;\; \frac{\partial}{\partial d^*_{jg}}\delta(d^*_{jg}, \mathbf{B}_{jg}) = 0. \tag{3.41}$$

Assim,

$$\begin{aligned}
\delta(d_{jg}, \mathbf{B}_{jg}) &= \int_{\Theta} f(\boldsymbol{\theta}_g) log\left(\frac{f(\boldsymbol{\theta}_g)}{\mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_{jg}, \mathbf{B}_{jg})}\right) d\boldsymbol{\theta}_g \\
&= \int_{\Theta} \left[log\, f(\boldsymbol{\theta}_g)\right] f(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g - \int_{\Theta} \left[log\, \mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_{jg}, \mathbf{B}_{jg})\right] f(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g \\
&= c - \int_{\Theta} log\left\{\frac{|\mathbf{B}_{jg}|^{d_{jg}/2}\,|\boldsymbol{\theta}_g|^{\frac{-(d_{jg}+p+1)}{2}}}{2^{d_{jg}p/2}\Gamma_p(d_{jg}/2)} \exp\left\{-\frac{1}{2}tr\left(\mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right\}\right\} f(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g \\
&= c - \int_{\Theta} log\left[\frac{|\mathbf{B}_{jg}|^{d_{jg}/2}}{2^{d_{jg}p/2}\Gamma_p(d_{jg}/2)}\right] f(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g - \int_{\Theta} log\left[|\boldsymbol{\theta}_g|^{-(d_{jg}+p+1)/2}\right] f(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g \\
&\quad + \int_{\Theta} log\left\{\exp\left\{-\frac{1}{2}tr\left(\mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right\}\right\} f(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g \\
&= c - \frac{d_{jg}}{2}\left[log\,|\mathbf{B}_{jg}| - p\,log 2\right] + log\,\Gamma_p\left(\frac{d_{jg}}{2}\right) + \frac{d_{jg}+p+1}{2}E\left[log\,|\boldsymbol{\theta}_g|\right] \\
&\quad + \frac{1}{2}E\left[tr\left(\mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right],
\end{aligned}$$

onde $c = E\left[log\, f(\boldsymbol{\theta}_g)\right]$. Agora, para solucionar a primeira derivada parcial, vamos usar dois resultados de Teoria das Matrizes (Harville, 2008),

i. Seja A uma matriz simétrica. Então $\frac{\partial log|A|}{\partial A} = A^{-1}$, e

ii. Seja B outra matriz simétrica. Então $\frac{\partial tr(AB)}{\partial A} = B$.

Assim,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{B}_{jg}}\delta(d^*_{jg},\mathbf{B}_{jg}) &= -\frac{d^*_{jg}}{2}\frac{\partial log\ |\mathbf{B}_{jg}|}{\partial \mathbf{B}_{jg}} + \frac{1}{2}E\left[\frac{\partial}{\partial \mathbf{B}_{jg}} tr\left(\mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right] \\ &= -\frac{d^*_{jg}}{2}\mathbf{B}_{jg}^{-1} + \frac{1}{2}E\left[\boldsymbol{\theta}_g^{-1}\right],\end{aligned}$$

e como solução da equação,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{B}_{jg}}\delta(d^*_{jg},\mathbf{B}_{jg}) = 0 &\implies -\frac{d^*_{jg}}{2}\mathbf{B}_{jg}^{-1} + \frac{1}{2}E\left[\boldsymbol{\theta}_g^{-1}\right] = 0 \\ &\implies -\frac{d^*_{jg}}{2}\mathbf{B}_{jg}^{-1} = -\frac{1}{2}E\left[\boldsymbol{\theta}_g^{-1}\right] \\ &\implies \mathbf{B}_{jg} = \left[\frac{1}{d^*_{jg}}E\left[\boldsymbol{\theta}_g^{-1}\right]\right]^{-1}, \qquad (3.42)\end{aligned}$$

mas, como $\boldsymbol{\theta}_g \sim \sum_{i=1}^{k}\omega_{ijg}\mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg})$, logo

$$\begin{aligned}E\left[\boldsymbol{\theta}_g^{-1}\right] &= \int_\Theta \boldsymbol{\theta}_g^{-1}\sum_{i=1}^{k}\omega_{ijg}\mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg})d\boldsymbol{\theta}_g \\ &= \sum_{i=1}^{k}\omega_{ijg}\int_\Theta \boldsymbol{\theta}_g^{-1}\mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg})d\boldsymbol{\theta}_g,\end{aligned}$$

pela propriedade (A.2.1) da distribuição Wishart Inversa, temos

$$E\left[\boldsymbol{\theta}_g^{-1}\right] = \sum_{i=1}^{k}\omega_{ijg}p\Delta_{ijg}^{-1}. \qquad (3.43)$$

Substituindo (3.43) em (3.42), logo

$$\mathbf{B}_{jg} = \frac{d^*_{jg}}{p}\left[\sum_{i=1}^{k}\omega_{ijg}\Delta_{ijg}^{-1}\right]^{-1}. \qquad (3.44)$$

Agora, para obtermos d^*_{jg}, temos

$$\frac{\partial}{\partial d^*_{jg}}\delta(d^*_{jg},\mathbf{B}_{jg}) = -\frac{1}{2}\left[log\ |\mathbf{B}_{jg}| - p\,log2\right] + \frac{1}{2}\sum_{v=1}^{p}\Psi\left(\frac{d_{jg}-v+1}{2}\right) + \frac{1}{2}E\left[log\ |\boldsymbol{\theta}_g|\right]. \qquad (3.45)$$

Entretanto, se $\boldsymbol{\theta}_g \sim \sum_{i=1}^{k} \omega_{ijg} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg})$, então

$$E\left[log\ |\boldsymbol{\theta}_g|\right] = \sum_{i=1}^{k} \omega_{ijg} \int_{\Theta} log\ |\boldsymbol{\theta}_g|\, \mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg}) d\boldsymbol{\theta}_g,$$

pela propriedade (A.2.2) da distribuição Wishart Inversa, temos

$$E\left[log\ |\boldsymbol{\theta}_g|\right] = \sum_{i=1}^{k} \omega_{ijg} log\,|\Delta_{ijg}| - p\,log\,2 - \sum_{v=1}^{p} \Psi\left(\frac{p-p+v}{2}\right). \tag{3.46}$$

Substituindo (3.46) em (3.45), a solução para a equação

$$\frac{\partial}{\partial d_{jg}^{*}} \delta(d_{jg}^{*}, \mathbf{B}_{jg}) = 0,$$

é o valor de d_{jg} que satisfaz

$$log\ |\mathbf{B}_{jg}| - \sum_{v=1}^{p} \Psi\left(\frac{d_{jg}-v+1}{2}\right) = \sum_{i=1}^{k} \omega_{ijg} log\,|\Delta_{ijg}| - \sum_{v=1}^{p} \Psi\left(\frac{v}{2}\right). \tag{3.47}$$

Portanto o teorema fica demonstrado. □

Pelo Teorema 3.6.1, a melhor aproximação para a mistura de *Wishart Inversa* por uma *densidade Wishart Inversa* é da forma

$$\sum_{i=1}^{k} \omega_{ijg} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid p, \Delta_{ijg}) \approx \mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_{jg}^{*}, \mathbf{B}_{jg}), \tag{3.48}$$

Observe que, $\Delta_{ijg} = (\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})(\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})'$ é uma matriz singular. Uma alternativa para contornar essa singularidade é usarmos o *pseudo-determinante* e a *pseudo-inversa da matriz*, regularizando a matriz Δ_{ijc} por (McLachlan, 2004):

$$\Delta_{ijg}^{*} = \Delta_{ijg} + \delta I_p \tag{3.49}$$

onde δ é um parâmetro relativamente pequeno tal que Δ_{ijc} seja positiva definida.

Portanto, a densidade aproximada para $\mathbf{y}^{(m)}$ é

$$\hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)}) \;\; \propto \;\; |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_{jg}^*, \mathbf{B}_{jg}), \tag{3.50}$$

um produto de Wishart Inversas. Assim, podemos usar a priori objetiva de Jefreys para $\boldsymbol{\theta}_g$ (Sun & Berger, 2006),

$$\hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(k)}) \propto |\boldsymbol{\theta}_g|^{-(p+1)/2}. \tag{3.51}$$

Assim, conjugando a densidade aproximada para $\mathbf{y}^{(m)}$ e a priori de $\boldsymbol{\theta}_g$, obtemos a densidade a posteriori de $\boldsymbol{\theta}_g$, dada por:

$$\begin{aligned}
\hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)}) &\propto \hat{f}_{\omega_g}(\mathbf{y}^{(m)} \mid \boldsymbol{\theta}_g, \mathbf{x}^{(k)})\, \hat{\pi}_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(k)}) \\
&\propto |\boldsymbol{\theta}_g|^{mp} \prod_{j=1}^{m} \mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_{jg}^*, \mathbf{B}_{jg})\, |\boldsymbol{\theta}_g|^{-(p+1)/2} \\
&= |\boldsymbol{\theta}_g|^{mp-(p+1)/2} \prod_{j=1}^{m} \frac{|\mathbf{B}_{jg}|^{d_{jg}^*/2}}{2^{d_{jg}^* p/2}\Gamma_p(d_{jg}^*/2)}\, |\boldsymbol{\theta}_g|^{-(d_{jg}^*+p+1)/2} \exp\left\{-\frac{1}{2} tr\left(\mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right\} \\
&\propto |\boldsymbol{\theta}_g|^{mp}\, |\boldsymbol{\theta}_g|^{-\sum_{j=1}^{m}(d_{jg}^*+p+1)/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m} tr\left(\mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right\} |\boldsymbol{\theta}_g|^{-(p+1)/2} \\
&\propto |\boldsymbol{\theta}_g|^{-(\sum_{j=1}^{m} d_{jg}^* + mp + m - 2mp + p + 1)/2} \exp\left\{-\frac{1}{2} tr\left(\sum_{j=1}^{m} \mathbf{B}_{jg}\boldsymbol{\theta}_g^{-1}\right)\right\} \\
&= |\boldsymbol{\theta}_g|^{-(d_g^*+p+1)/2} \exp\left\{-\frac{1}{2} tr\left(\mathbf{B}_g^*\boldsymbol{\theta}_g^{-1}\right)\right\}.
\end{aligned} \tag{3.52}$$

Identificamos (3.52) como o núcleo de uma distribuição Wishart Inversa, temos

$$\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)} \sim \mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_g^*, \mathbf{B}_g^*) \tag{3.53}$$

onde $\mathbf{B}_g^* = \sum_{j=1}^{m} \mathbf{B}_{ig}$ e $d_g^* = \sum_{j=1}^{m} d_{jg}^* + m - mp$. Portanto, voltando a densidade preditiva

aproximada de uma nova observação $\mathbf{x}^{novo}$ dado $\mathbf{x}^{n_g}$, temos:

$$\begin{aligned}\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) &= \frac{1}{k}\sum_{i=1}^{k}\int_{\Theta}|\boldsymbol{\theta}_g|^{-1/2}\,\mathcal{K}(\boldsymbol{\theta}_g^{-1/2}(\mathbf{x}^{novo}-\mathbf{x}_{ig}^{(k)}))\pi_{\omega_g}(\boldsymbol{\theta}_g \mid \mathbf{x}^{(n_g)})d\boldsymbol{\theta}_g \\ &= \frac{1}{k}\sum_{i=1}^{k}\int_{\Theta}\frac{|\boldsymbol{\theta}_g|^{-1/2}}{(2\pi)^{p/2}}\exp\{-\frac{1}{2}(\mathbf{x}^{novo}-\mathbf{x}_{ig}^{(k)})'\boldsymbol{\theta}_g^{-1}(\mathbf{x}^{novo}-\mathbf{x}_{ig}^{(k)})\} \\ &\quad \times \mathcal{WI}_p(\boldsymbol{\theta}_g \mid d_g^*, \mathbf{B}_g^*)d\boldsymbol{\theta}_g. \qquad (3.54)\end{aligned}$$

Pelo Corolário A.3.1, a conjugação das densidades de $\mathbf{X}^{novo} \mid \boldsymbol{\theta}_g \sim N_p(\mathbf{x}_{ig}^{(k)}, \boldsymbol{\theta}_g)$ com $\boldsymbol{\theta}_g \sim WI(d_g^*, \mathbf{B}_g^*)$, na expressão (3.54), resulta em

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \approx \frac{1}{k}\sum_{i=1}^{k} t_p(\mathbf{x}^{novo} \mid \mathbf{x}_{ig}^{(k)}, \mathbf{B}_g^*, d_g^*). \qquad (3.55)$$

Logo, em (3.55) temos uma mistura de k distribuições t matriz-variada $T_{1,p}(\mathbf{x}^{novo} \mid d_g^*, 2, \mathbf{x}_{ig}^{(k)}, 1, \mathbf{B}_g^*)$. Finalmente, para Q partições de $\mathbf{x}^{(n_g)} = \left\{\mathbf{x}_{(q)}^{(k)}, \mathbf{y}_{(q)}^{(m)}\right\}$, $q \in \{1, \ldots, Q\}$, obtemos:

$$\hat{f}_{\omega_g}(\mathbf{x}^{novo} \mid \mathbf{x}^{(n_g)}) \approx \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{k}\sum_{i=1}^{k} T_{1,p}(\mathbf{x}^{novo} \mid, d_{g(q)}^*, 2, \mathbf{x}_{ig(q)}^{(k)}, 1, \mathbf{B}_{g(q)}^*). \qquad (3.56)$$

Empregando (3.56) na Definição 3.2.1, obtemos então uma regra para Análise Discriminante empregando essas distribuições preditivas aproximadas.

Capítulo 4

Exemplos Computacionais e Aplicação

Na literatura existem diversas formas para avaliação de classificadores em Análise Discriminante, dentre elas, a *taxa de erro de classificação* que se constitui em uma estimativa para a probabilidade de má classificação. A *taxa de erro de classificação* é determinada pela proporção de objetos que o classificador aloca em classes que não são suas classes de origem. E para avaliar a aplicabilidade dos procedimentos em estudos reais, submetemos os classificadores a alguns conjuntos de dados bastante empregados na literatura.

Neste trabalho, consideramos a *taxa de erro de classificação* associada a um dado classificador C, como sendo a média da proporção de erros de classificação em um determinado número de repetições independentes do experimento, ou seja,

$$\bar{err}_C = \frac{1}{M}\sum_{m=1}^{M} \hat{err}(C \mid \mathbf{X}_{\{m\}}), \tag{4.1}$$

onde $\bar{err}_C$ é a taxa de erro de classificação do classificador C, M é o número de repetições independentes do experimento e $\hat{err}(C \mid \mathbf{X}_{\{m\}})$ é a taxa de erro do classificador para um determinado conjunto de dados $\mathbf{X}_{\{m\}}$ cuja as observações são identificada com relação as classes.

Afim de avaliar a habilidade dos procedimentos discriminantes em classificar os objetos, consideramos apenas as abordagens de estimação de densidades nas classes, propostas nesse trabalho, sem qualquer influência do *custo de classificação* e das probabilidades a priori. Este procedimento foi adotado tanto nos estudos de simulação quanto na aplicação em dados reais, ou seja, todos os classificadores consideram a *função de perda 0-1* e as probabilidade a priori iguais, sendo então uma classificação baseada apenas na máxima densidade estimada.

4.1 Experimentos com Dados Simulados

Nesse estudo de simulação, o objetivo principal é comparar os métodos propostos com os métodos mais tradicionais. As abordagens de Análise Discriminante empregadas foram: *Densidade Preditiva Bayesiana por Função Núcleo Multivariada Normal*, *Densidade Preditiva Bayesiana por Produto de Funções Núcleo Normais* e *Densidade Preditiva Bayesiana por Produto de Funções Núcleo Normais empregando Componentes Independentes*, representados por **GBPKDA**, **BPKDA 1** e **BPKDA 2**, respectivamente, e *Análise Discriminante Linear* (**ADL**); *Análise Discriminante Quadrática* (**ADQ**) e *Naive Bayes Normal*(**NNBDA**). Esses procedimentos foram submetidos a 6 estruturas de dados, considerando apenas duas classes em todas as estruturas, baseadas em misturas de densidades de modelos simétricos (Normal e t) e assimétricos(Normal assimétrica e t assimétrica), que constituem situações bem mais desafiadoras para esses procedimentos.

Os métodos foram avaliados com tamanhos de amostra 100 e 300 em 1000 réplicas independentes do experimento, onde em cada réplica é obtida a taxa de erro de classificação e ao final do experimento é tomado a média dessas taxas como estimativa da taxa de erro de cada classificador. Além disso, os métodos GBPKDA, BPKDA 1 e BPKDA 2 têm dentro de seus procedimentos internos uma Validação Cruzada "Q-fold", dessa forma nesse trabalho vamos considerar 5, 7 e 9 partições internas dos dados.

4.1.1 Particularidades na Implementação dos Classificadores

- **GBPKDA:** No procedimento de regularização da matriz Δ_{ijg} o valor de δ escolhido foi 0,001. No processo de otimização para minimizar a derivada parcial afim de obter um valor para d_g^*, empregamos a função *uniroot.all* do pacote *rootSolve* da linguagem de programação R.

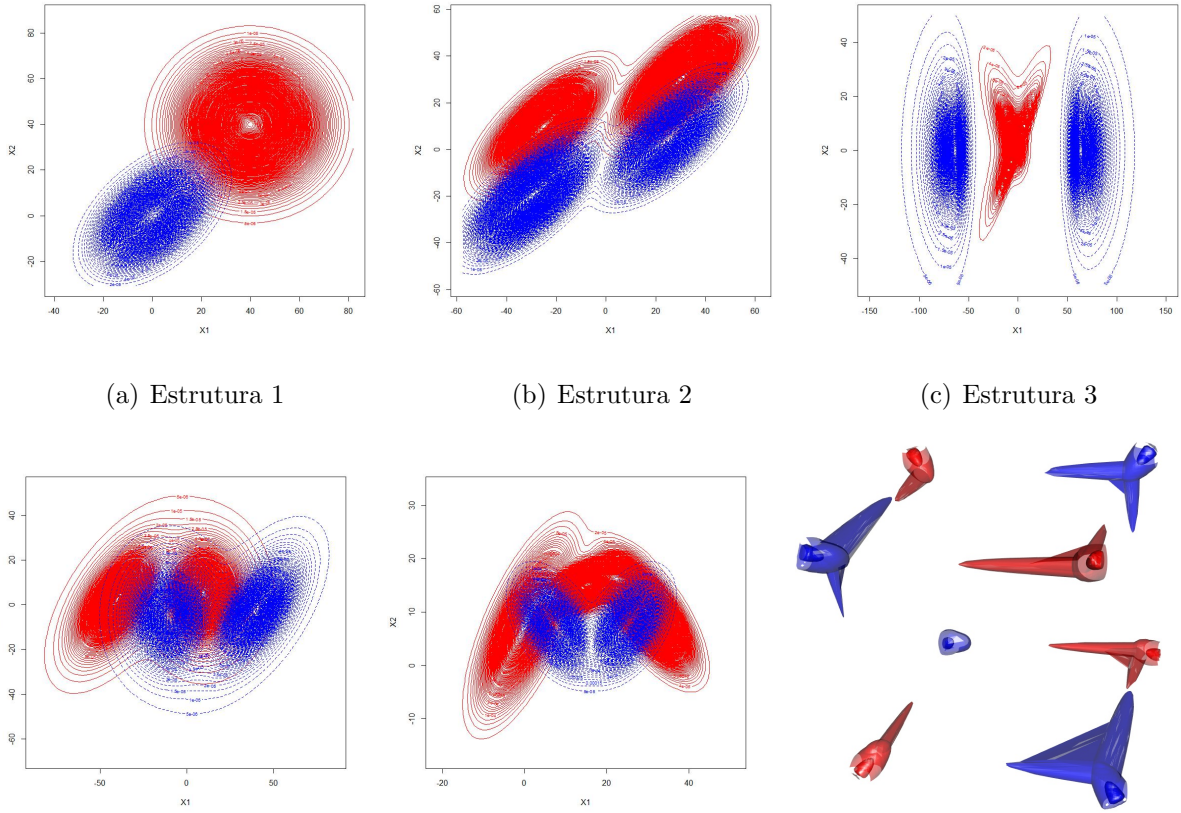
- **BPKDA 2:** As componentes independentes são obtidas pela função *fastICA* do pacote *fastICA* da linguagem de programação R, que emprega o algoritmo de Hyvarinen & Oja (2000).

4.1.2 Conjunto de Treinamento e Teste

No processo de geração das observações, cada observação tem probabilidade 0,5 de ser gerada segunda as características de cada uma das classes, por isso o número de observações em cada classe é aleatório. Em cada repetição do experimento, são geradas $n \in \{100, 300\}$ observações das classes e essas observações compõem o *conjunto de treinamento*, que são usados para treinar as funções discriminantes. Em seguida, é gerado de forma independente 1000 observações das classes que compõem o *conjunto de teste*, e essas observações são classificadas segundo as regras já treinadas e sobre elas são calculadas as *taxas de erro de classificação*.

Na Figura 4.1, apresentamos observações das diferentes estruturas para as distribuições das classes consideradas no experimento computacional. Os valores dos parâmetros para simular as observações dessas estruturas estão no Apêndice B.

Figura 4.1: Exemplo das Estruturas Simuladas



(a) Estrutura 1 (b) Estrutura 2 (c) Estrutura 3

(d) Estrutura 4 (e) Estrutura 5 (f) Ilustração da Estrutura 6

A Estrutura 1 é a situação mais simples empregada, com as densidades das classes sendo distribuições Normais onde em uma classe as observações possuem correlação positiva e a outra com correlação nula entre as variáveis. Na Estrutura 2, as observações são simuladas empregando um modelo de mistura de densidades Normais. A Estrutura 3 segue um conceito um pouco mais complicado em termos de estimação da densidade, onde a classe localizada no centro é um modelo de mistura de densidades empregando uma distribuição t assimétrica, e na outra classe temos uma mistura de densidades t assimétricas, sendo as modas de cada componente da mistura bem afastadas. Na Estrutura 4, segue a mesma ideia da Estrutura 3, onde as modas das componentes da mistura são separadas pela outra classe, geradas por misturas de Normais, e há uma razoável sobreposição entre as classes. A Estrutura 5 também é gerada por modelos de misturas de Normais, porém com uma sobreposição bem maior entre as classes. Na Estrutura 6, as observações são ge-

radas a partir de uma mistura de densidades t assimétricas, sua estrutura é semelhante a um hipercubo em 4 dimensões, onde em cada componente das misturas o vetor de locação é um vértice do hipercubo, também, os vértices adjacentes neste hipercubo são ocupados pelas componente da mistura de outra classe. Na Figura 4.2(f), temos uma ilustração em 3 dimensões da Estrutura 6.

As observações foram geradas pelo função *rmmix* do pacote *mixsmsn* da linguagem de programação R (R Development Core Team, 2012), usando as densidades como definidas no Apêndice (A) (para mais informações ver Cabral *et al.* (2012)).

4.1.3 Resultados das simulações

Tabela 4.1: Média e desvio padrão das estimativa da taxa de erro de classificação.

Estrutura	Método	Q	N = 100		N = 300		Estrutura	Método	Q	N = 100		N = 300	
			Média	Sd	Média	Sd				Média	Sd	Média	Sd
		5	0.0148	0.0040	0.0139	0.0041			5	0.2421	0.0417	0.2605	0.0302
	GBPKDA	7	0.0158	0.0043	0.0137	0.0041		GBPKDA	7	0.2378	0.0408	0.2391	0.0300
		9	0.0158	0.0042	0.0143	0.0041			9	0.2360	0.0406	0.2308	0.0292
		5	0.0194	0.0064	0.0180	0.0052			5	0.1880	0.0211	**0.1747**	0.0164
	BPKDA 1	7	0.0194	0.0061	0.0180	0.0053		BPKDA 1	7	0.1871	0.0203	0.1750	0.0160
1		9	0.0194	0.0060	0.0180	0.0051	**4**		9	0.1867	0.0198	0.1762	0.0162
		5	0.0187	0.0065	0.0180	0.0054			5	0.1878	0.0222	0.1753	0.0154
	BPKDA 2	7	0.0185	0.0062	0.0178	0.0054		BPKDA 2	7	0.1872	0.0219	0.1769	0.0151
		9	0.0185	0.0061	0.0177	0.0053			9	**0.1862**	0.0216	0.1771	0.015
	ADL		0.0149	0.0041	0.0146	0.0039		ADL		0.2830	0.0263	0.2823	0.0216
	ADQ		**0.0141**	0.0040	**0.0128**	0.0037		ADQ		0.2850	0.0272	0.2822	0.0216
	NNBDA		0.0152	0.0045	0.0143	0.0040		NNBDA		0.3047	0.0368	0.3028	0.0280
		5	0.1003	0.0266	0.1178	0.0228			5	0.2867	0.0488	0.3144	0.0359
	GBPKDA	7	0.0964	0.0240	0.1082	0.0204		GBPKDA	7	0.2939	0.0445	0.2749	0.0371
		9	0.0940	0.0226	0.1037	0.0187			9	0.2898	0.0438	0.2631	0.0332
		5	0.0845	0.0153	0.0757	0.0096			5	**0.2055**	0.0191	0.1919	0.0154
	BPKDA 1	7	0.0850	0.0151	0.0763	0.0098		BPKDA 1	7	0.2060	0.0189	0.1931	0.0154
2		9	0.0854	0.0151	0.0767	0.0098	**5**		9	0.2066	0.0188	0.1935	0.0153
		5	0.0834	0.0149	0.0760	0.0102			5	0.2061	0.0212	0.1908	0.0163
	BPKDA 2	7	0.0839	0.0149	0.0766	0.0100		BPKDA 2	7	0.2065	0.0210	**0.1917**	0.0164
		9	0.0843	0.0147	0.0766	0.0099			9	0.2070	0.0211	0.1926	0.0164
	ADL		**0.0658**	0.0090	**0.0641**	0.0077		ADL		0.3703	0.0480	0.3559	0.0304
	ADQ		0.0675	0.0088	0.0656	0.0079		ADQ		0.2647	0.0361	0.2423	0.0269
	NNBDA		0.2084	0.0193	0.2084	0.0145		NNBDA		0.2532	0.0308	0.2357	0.0221
		5	0.0500	0.056	0.0299	0.0334			5	0.1352	0.0703	0.2358	0.1586
	GBPKDA	7	0.0539	0.0433	0.0261	0.0238		GBPKDA	7	0.1247	0.0623	0.1051	0.0988
		9	0.0544	0.0430	0.0284	0.0173			9	0.1202	0.0583	0.0590	0.0505
		5	0.0431	0.0109	0.0328	0.0079			5	0.0450	0.0196	0.0273	0.0106
	BPKDA 1	7	0.0438	0.0104	0.0335	0.0076		BPKDA 1	7	0.0485	0.0193	0.0282	0.0105
3		9	0.0444	0.0101	0.0340	0.0076	**6**		9	0.0510	0.0199	0.0287	0.0098
		5	0.0177	0.0068	0.0147	0.0052			5	**0.0266**	0.0134	0.0225	0.0117
	BPKDA 2	7	**0.0174**	0.0065	0.0146	0.0051		BPKDA 2	7	0.0276	0.0139	**0.0221**	0.0103
		9	0.0175	0.0060	**0.0145**	0.0049			9	0.0283	0.0140	0.0227	0.0105
	ADL		0.3840	0.0590	0.3883	0.0578		LDA		0.4993	0.0298	0.4992	0.0216
	ADQ		0.0758	0.0726	0.0672	0.0643		QDA		0.4574	0.0762	0.4793	0.0675
	NNBDA		0.0781	0.0883	0.0698	0.0824		NNBDA		0.4991	0.0305	0.4992	0.0210

Sd: Desvio padrão.

Na Tabela 4.1, temos as média e os desvios padrão das estimativas das taxas de erro de classificação das 1000 réplicas independentes do experimento em todas as estruturas de dados para todos os procedimento discriminantes. Na Estrutura 1, o classificador com menor taxa de erro foi o ADQ com 1,41% de classificações erradas, os métodos propostos nesse trabalho obtiveram resultados muito próximos do ADQ mostrando-se competitivos nessa situação, em geral as taxas de erros foram pequenas, o que era esperado pois essa estrutura era a situação mais fácil (veja Figura 4.2). Na Estrutura 2, novamente os métodos clássicos ADL e ADQ obtiveram menores taxas de erro, com uma diferença entre 2% e 4% aproximadamente, entretanto com o aumento do tamanho da amostra os métodos BPKDA 1 e 2 tiveram um redução de aproximadamente 1% nas taxas de erro, o método NNBDA obteve resultados nada satisfatórios (veja Figura 4.3).

Nas Estruturas 3, 4 e 5 os métodos propostos nesse trabalho tiveram resultados bem melhores que os procedimentos clássicos, exceto o GBPKDA na Estrutura 5, destacando-se o procedimento BPKDA 2 que obteve as melhores taxas de erro dos classificadores estudados, no máximo apresentou resultados próximos do BPKDA 1. Na comparação entre BPKDA 1 e 2, o BPKDA 2 foi bem mais satisfatório, o que era esperado pois, mesmo que ambos considerem independência entre as variáveis, o BPKDA 2 emprega as componentes independentes que tem o objetivo de produzir variáveis independentes, logo espera-se que esse procedimento obtivesse melhor resultado. De um modo geral, o método ADL foi o que teve os piores resultados (veja Figuras 4.4, 4.5 e 4.6).

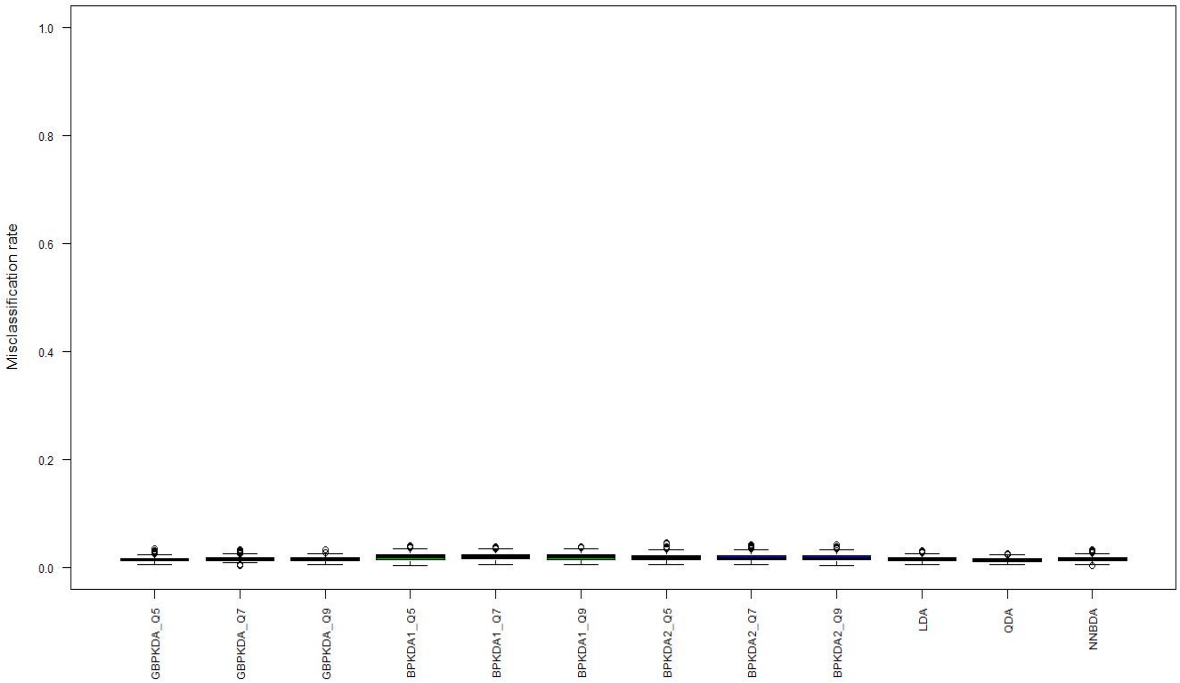
Os resultados na Estrutura 6 foram os mais divergentes entre os métodos propostos empregando função núcleo e os procedimento clássicos da literatura, estes últimos apresentaram taxas de erro de classificação próximas a 50% em ambos os tamanhos de amostra, enquanto os procedimentos por função núcleo apresentaram taxas de erro entre 2,66% e 13,52% para tamanho de amostra 100, e taxas de erro entre 2,21% e 23,58%, mostrando então a flexibilidade desses procedimentos mesmo em estruturas de dados mais complexas. Entre os métodos por função núcleo, o que apresentou piores resultados foi o GBPKDA em ambos os tamanhos de amostra, o BPKDA 2 apresentou resultados melhores que o BPKDA 1 para tamanho de amostra 100, mas com o aumento do tamanho da amostra para 300, o BPKDA 1 mostrou resultados próximos do BPKDA 2 (veja Figura 4.7).

Além das médias das taxas de erro de classificação, a Tabela 4.1 mostra ainda a estimativa dos desvios padrão. Em geral, as estimativas são pequenas o que indica que os classificadores têm uma variabilidade pequena. Entretanto, como forma de avaliar de modo empírico essa variabilidade, temos as seguintes figuras que mostram os boxplots de cada classificador para cada estrutura e tamanho de amostra. Na Figura 4.2, vemos que a variabilidade dos classificadores realmente são pequenas e semelhantes na estrutura de dados 1. Na Figura 4.3, vemos uma diferença nos resultados mais clara que na estrutura anterior, onde ADL e ADQ lideram com os melhores resultados e pequena variabilidade, e os métodos BPKDA 1 e 2, praticamente empatados, entretanto o GBPKDA com o aumento da amostra mostra uma queda significativa da mediana das taxas de erros com o aumento do número de partições internas do Q-fold CrossValidation, o NNBDA mostra resultados muito ruins em relação aos outros, o que mostra que por mais que a suposição de independência seja inadequada, os métodos BPKDA 1 e 2 conseguem obter bons resultados.

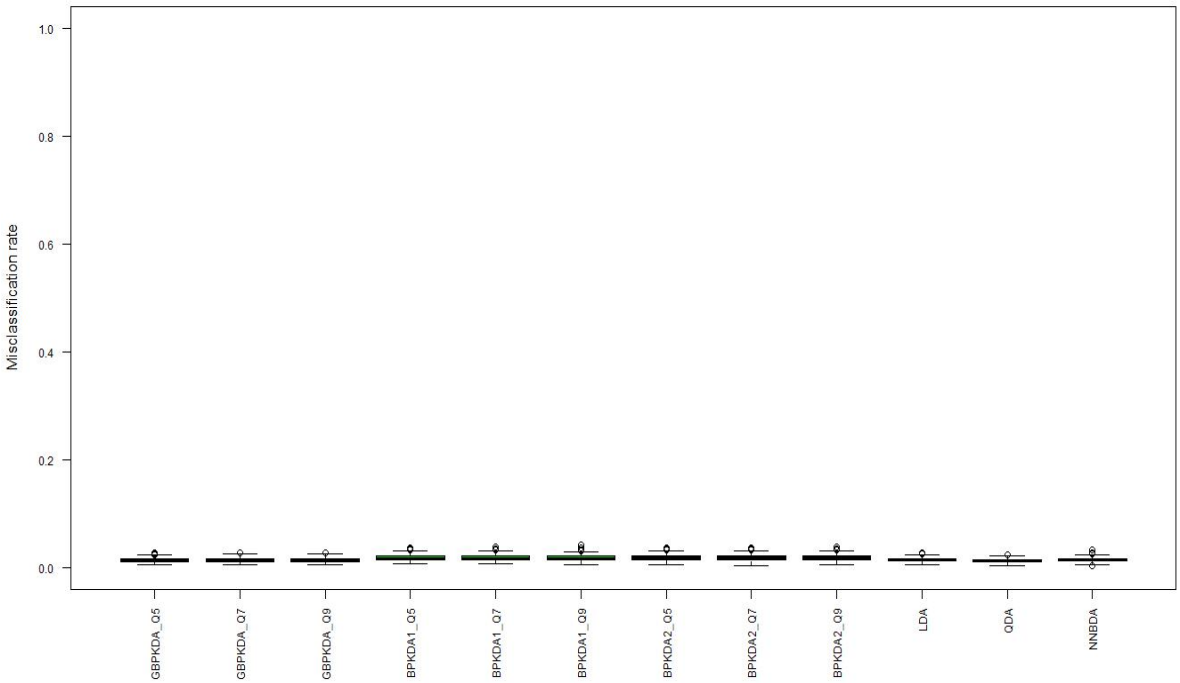
Na Figura 4.4, vemos que os métodos clássicos têm um desempenho inferior aos métodos por função núcleo, e apresentam vários pontos discrepantes, chegando a ter pontos que indicam que a taxa de erro superou os 60%. Entre os procedimentos por função núcleo, o que obteve uma maior variabilidade e maior número de discrepâncias foi o GBPKDA. Pela Tabela 4.1 o GBPKDA teve desempenho inferior aos BPKDA 1 e 2, entretanto pelo boxplot vemos que, em termos da mediana, ele obteve o melhor resultado ou equivalente aos outros métodos, principalmente com o aumento do tamanho da amostra. Na estrutura 4, pela Figura 4.5, vemos claramente a superioridade dos BPKDA 1 e 2 em relação aos outros métodos, e todos os classificadores apresentaram uma variabilidade próxima da normalidade. O método GBPKDA, apresentou novamente a característica de redução das taxas de erro com o aumento da amostra e partições internas. Na Figura 4.6, vemos um padrão semelhante ao anterior, porém o ADL apresentou desempenho inferior aos demais. Na Figura 4.7, podemos observar a clara diferença entre os métodos por função núcleo e os clássicos. Os métodos BPKDA 1 e 2 apresentam os melhores resultados, sendo o BPKDA 2 o melhor, com as menores taxas de erros de classificação e um variabilidade menor em relação aos demais. O GBPKDA, apresentou resultados competitivos aos dos

outros métodos por função núcleo apresentados, mas para tamanho de amostra 300 e com um número maior de partições internas do Validação Cruzada "Q-fold", ele aproximou-se bastante dos métodos BPKDA 1 e 2, o que pode indicar que para um tamanho de amostra e partições internas maiores, o GBPKDA pode melhorar mais ainda seu desempenho.

Figura 4.2: Taxa de erro de classificação da Estrutura 1

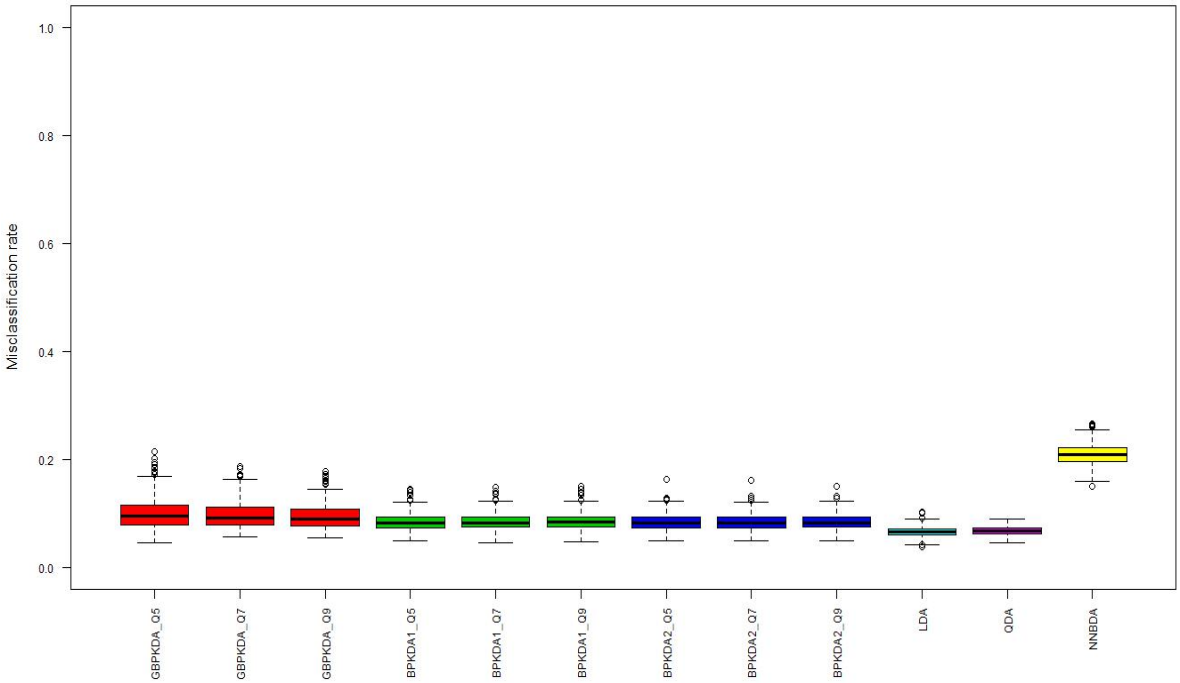


(a) Tamanho Amostral: 100

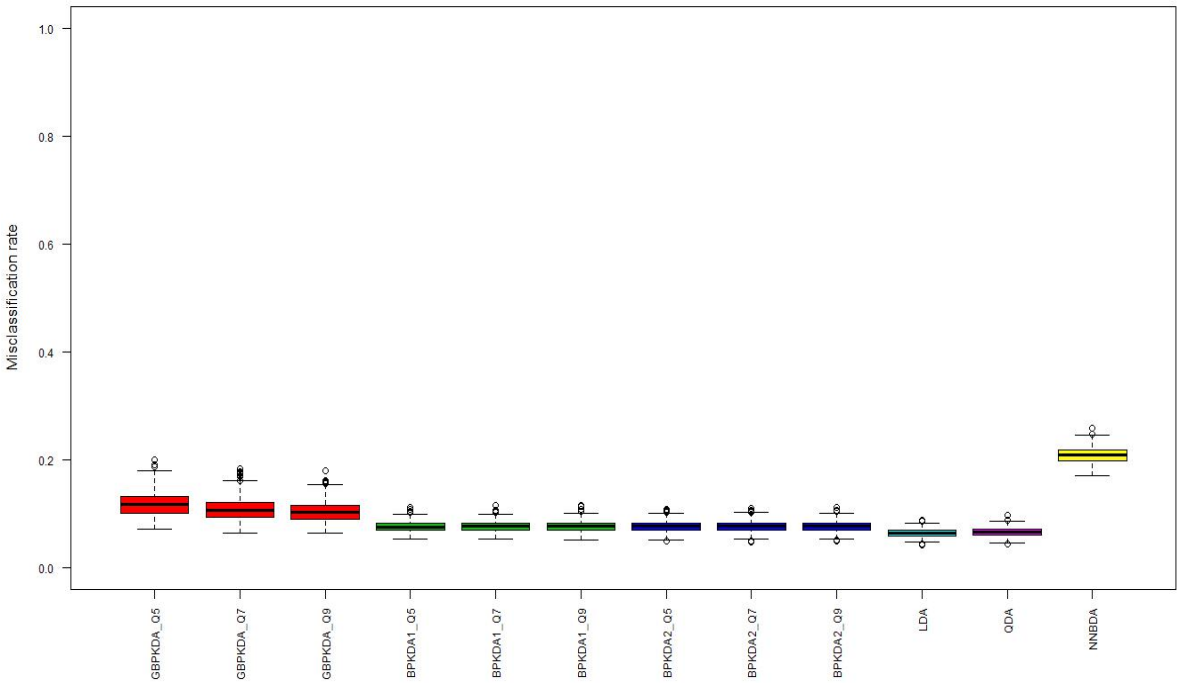


(b) Tamanho Amostral: 300

Figura 4.3: Taxa de erro de classificação da Estrutura 2

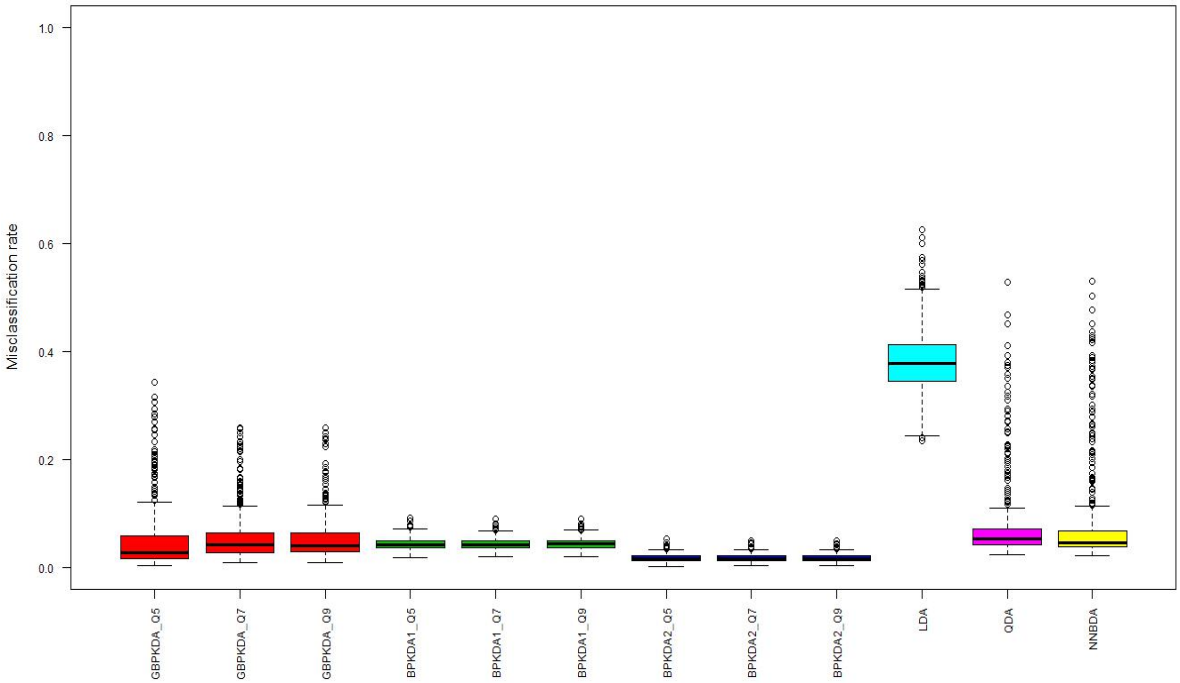


(a) Tamanho Amostral: 100

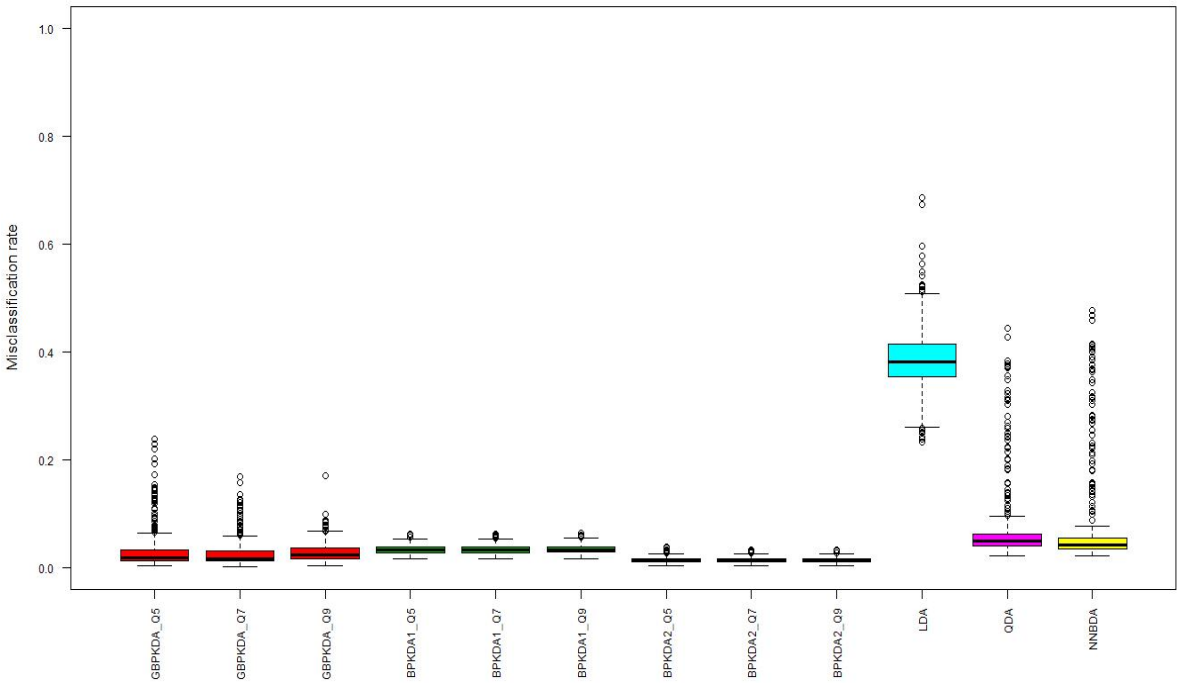


(b) Tamanho Amostral: 300

Figura 4.4: Taxa de erro de classificação da Estrutura 3

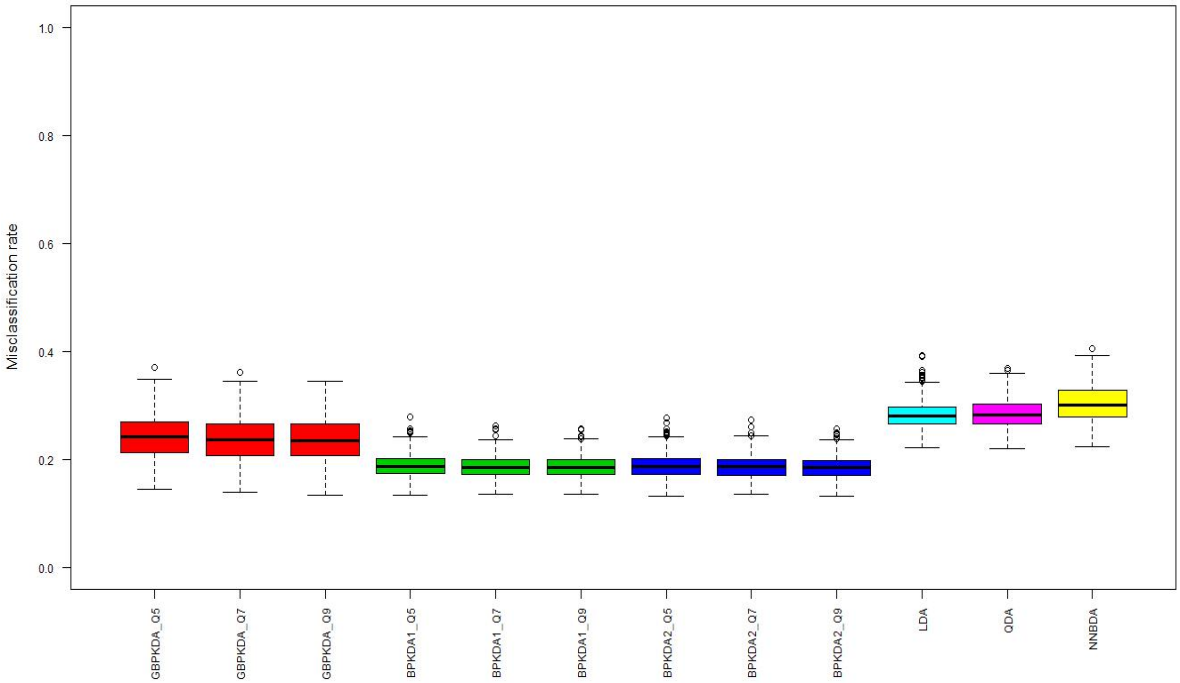


(a) Tamanho Amostral: 100

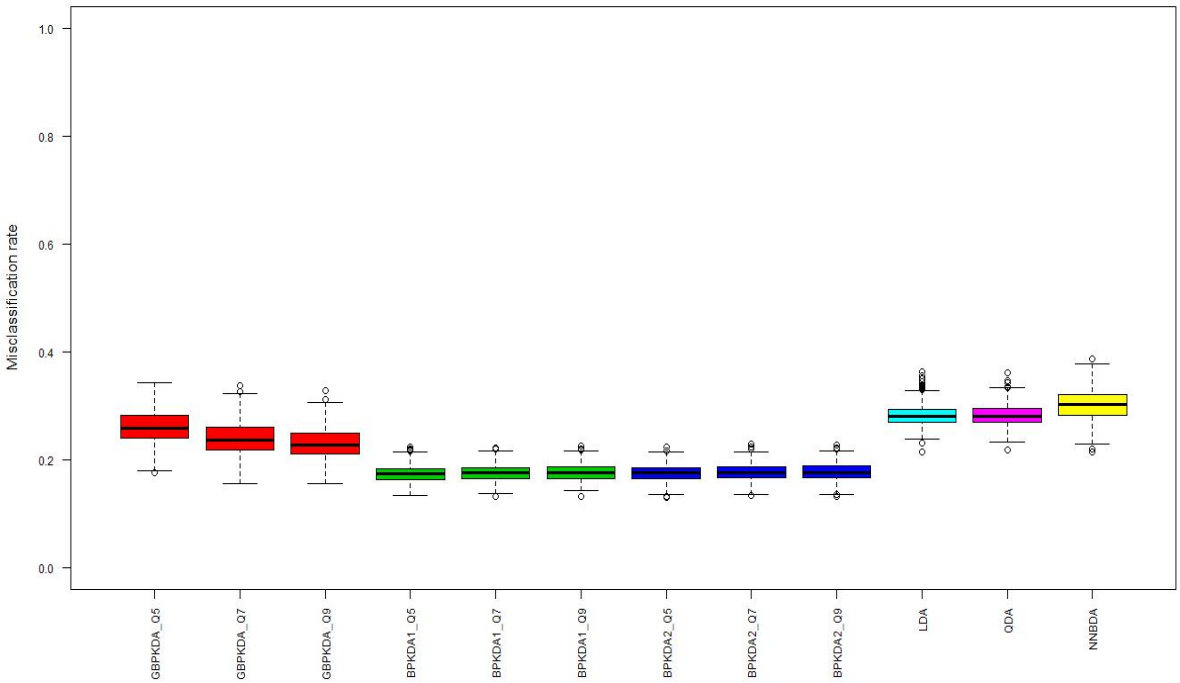


(b) Tamanho Amostral: 300

Figura 4.5: Taxa de erro de classificação da Estrutura 4

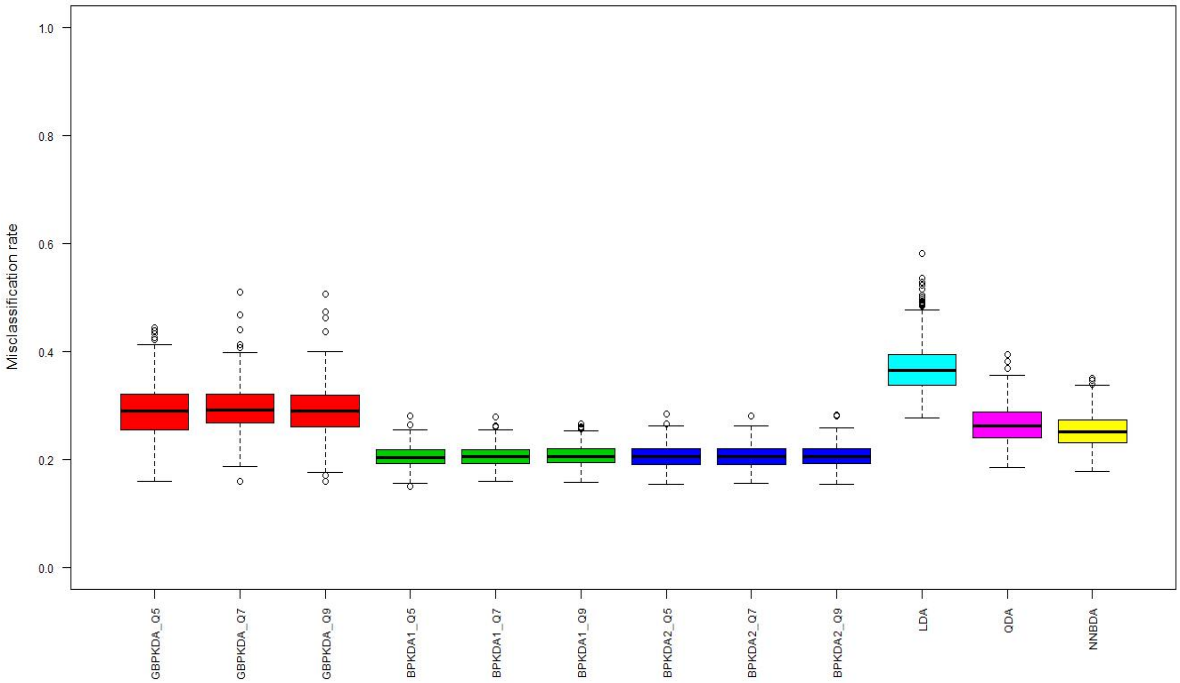


(a) Tamanho Amostral: 100

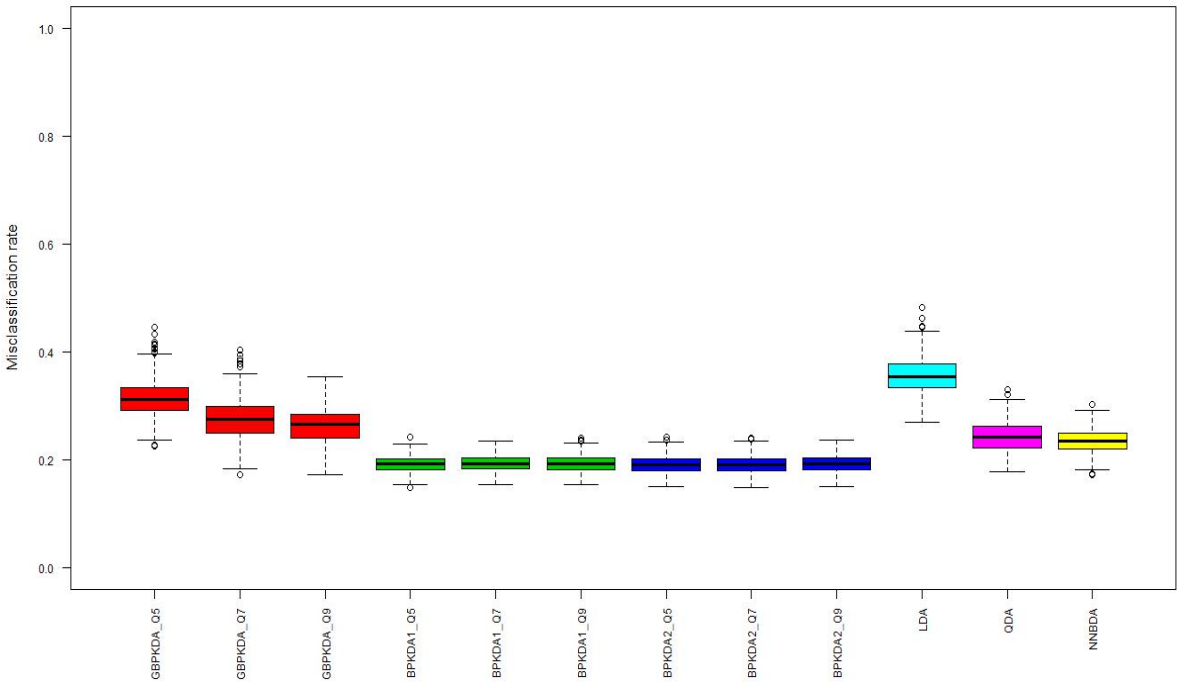


(b) Tamanho Amostral: 300

Figura 4.6: Taxa de erro de classificação da Estrutura 5

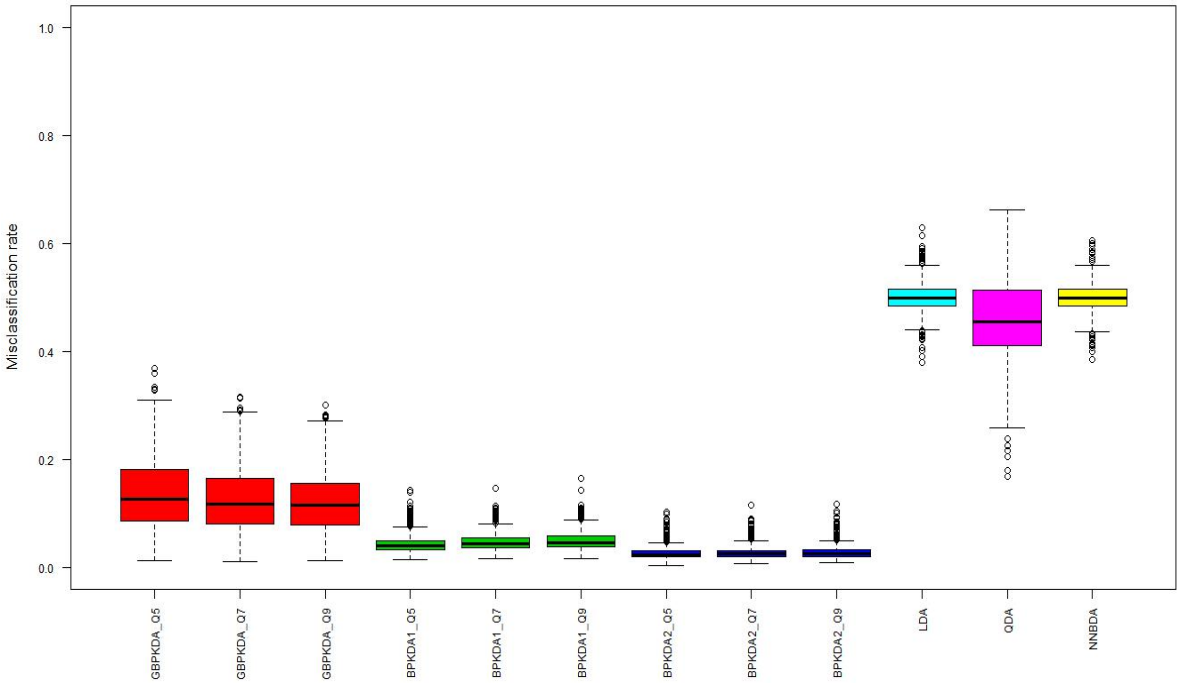


(a) Tamanho Amostral: 100

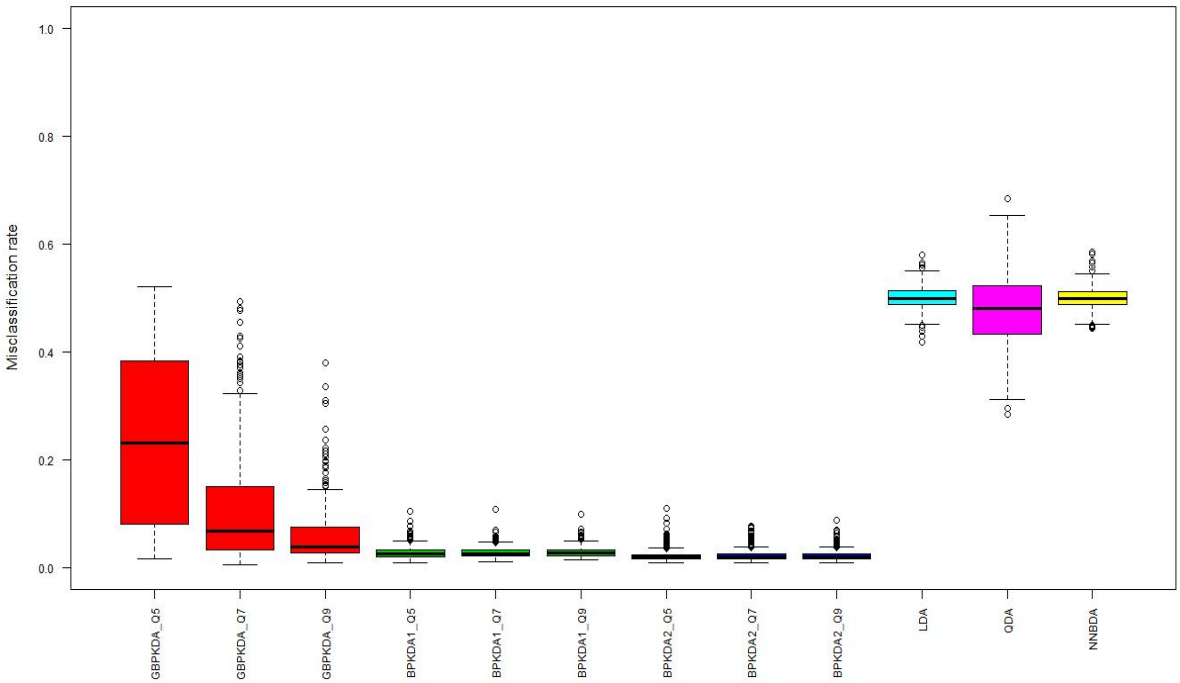


(b) Tamanho Amostral: 300

Figura 4.7: Taxa de erro de classificação da Estrutura 6



(a) Tamanho Amostral: 100



(b) Tamanho Amostral: 300

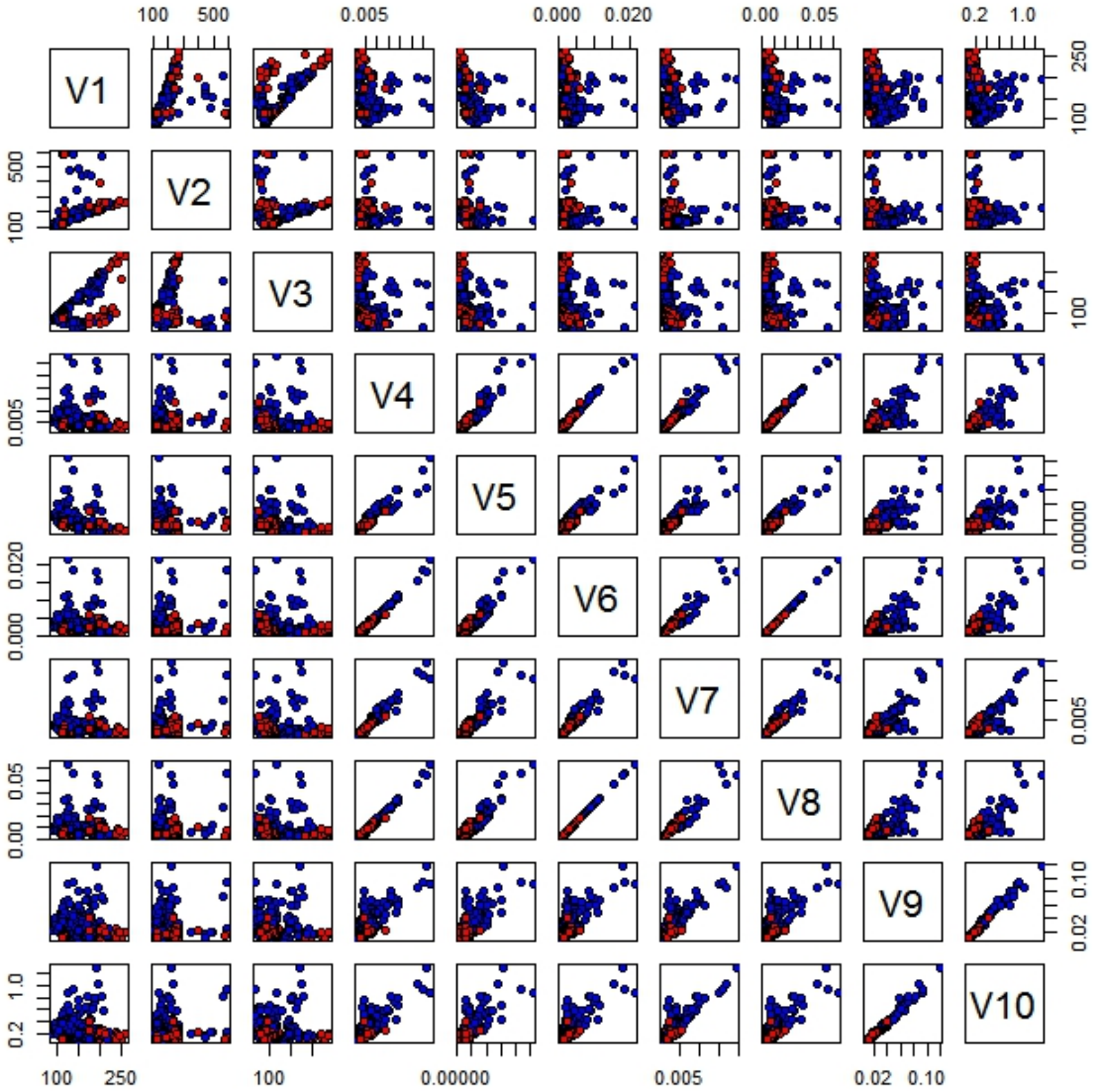
4.2 Aplicação em dados reais

Para a avaliação da aplicabilidade dos classificadores, a aplicação em dados reais tem como resultado uma estimativa da taxa de erro diferente do caso anterior. Como não é possível a repetição do experimento, empregamos o procedimento de Validação Cruzada "K-fold"(Efron, 1983), com número de partições iguais a 5 e 10, recomendados pela literatura. A avaliação consiste em empregar os procedimento em 5 conjuntos de dados bastante difundidos na literatura, que podem ser obtidos no site da *UCI Machine Learning Repository* (Frank & Asuncion, 2010). Os conjuntos de dados são:

1. **Wisconsin Diagnostic Breast Cancer:** Trata-se de um estudo de câncer de mama, onde as observações das características são extraídas de imagens digitalizadas da massa da mama, afim de classificar através das imagens se o tumor é benigno ou maligno. Estudo de 1995, com 569 observações, sendo 357 classificados como Benigno e 212 como maligno, com 30 características observadas.

   Na Figura 4.8 a seguir, apresentamos a distribuição conjunta das observações em duas dimensões para cada combinação das 10 primeiras variáveis. Em todas as combinações com as variáveis V1 e V2, podemos observar que há uma sobreposição das classes, e nas combinações entre as variáveis de V4 a V10, vemos uma alta correlação entre essas variáveis, que pode implicar em problemas para os procedimentos de classificação que supõem independência. De modo geral, a Figura 4.8 mostra que há uma presença de observações discrepantes, e em alguns pares de variáveis, podemos perceber que as observações apresentam um comportamente assimétrico.

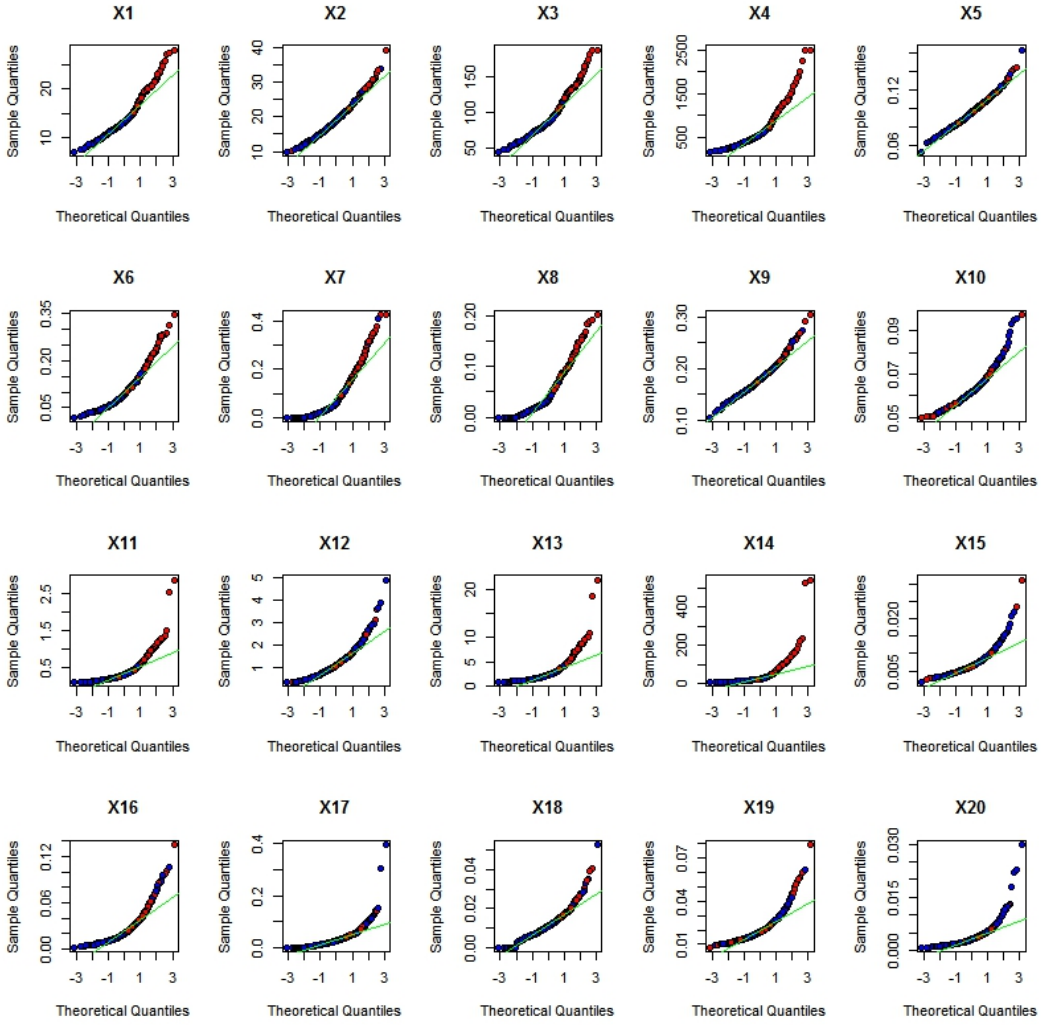
Figura 4.8: Distribuição conjunta das observações das 10 primeiras variáveis de *Wisconsin Diagnostic Breast Cancer*.



Na Figura 4.9, temos gráficos de quantis Normais por quantis observados, afim de avaliar de um modo independente se há uma possível normalidade nos dados. De modo univariado, na maioria das variáveis podemos identificar uma estrutura de classes, mas ainda com sobreposição entre as classes. Nenhuma das variáveis apresentam normalidade, o que pode ser visto se observarmos que os quantis teóricos e

observados não são lineares. Essa não normalidade pode implicar em sérios problemas para os procedimentos clássicos que admitem normalidade.

Figura 4.9: Gráfico Q-Q Normal das variáveis de *Wisconsin Diagnostic Breast Cancer*.

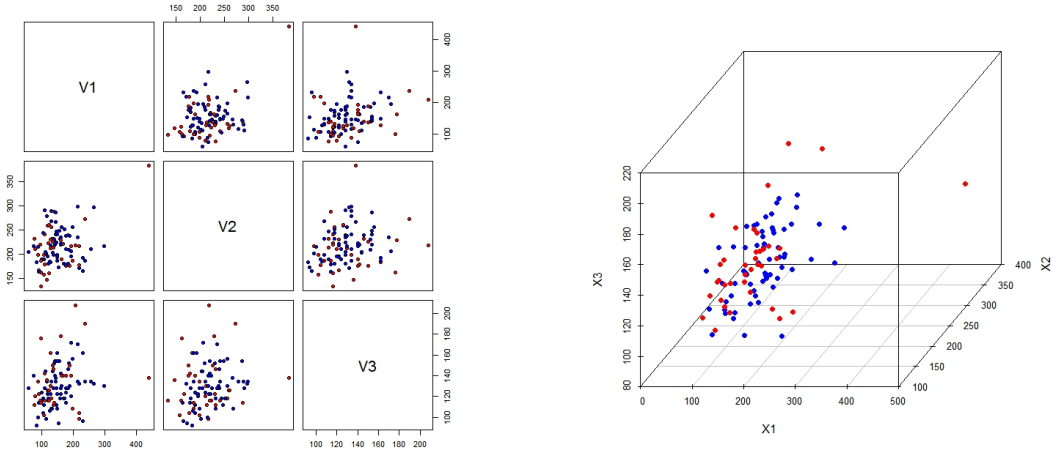


2. **Honolulu:** Estudo realizado na população de Honolulu - Havaí em 1969. Dos 7.683 indivíduos da população foram selecionados 100 e pesquisados a glicemia(mg/dL), colesterol sérico(mg/dL) e pressão sanguínea sistólica(mmHg), e 37 desses indivíduos foram classificados como *fumantes* e 63 como *não fumantes*.

   Pela Figura 4.10, podemos observar que há sobreposição entre as classes, uma correlação positiva considerável entre as variáveis, e também há presença de observações discrepantes. Tal estrutura de sobreposição pode implicar em grandes problemas paras os métodos de classificação, principalmente para aqueles que se baseiam na

distância entre as observações, que é o caso dos procedimento por função núcleo, mas também podem implicar em dificuldade para os métodos clássicos, pois os centroides das classes estão muito próximos.

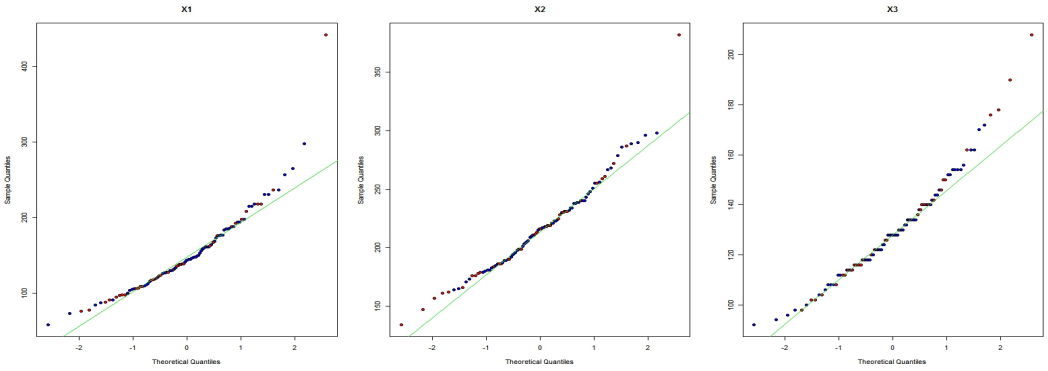
Figura 4.10: Distribuição conjunta das observações das variáveis de *Honolulu*.



(a) Por pares de variáveis (b) Em 3 dimensões

Pela Figura 4.11, observamos de modo independente que as variáveis provavelmente não possuem normalidade.

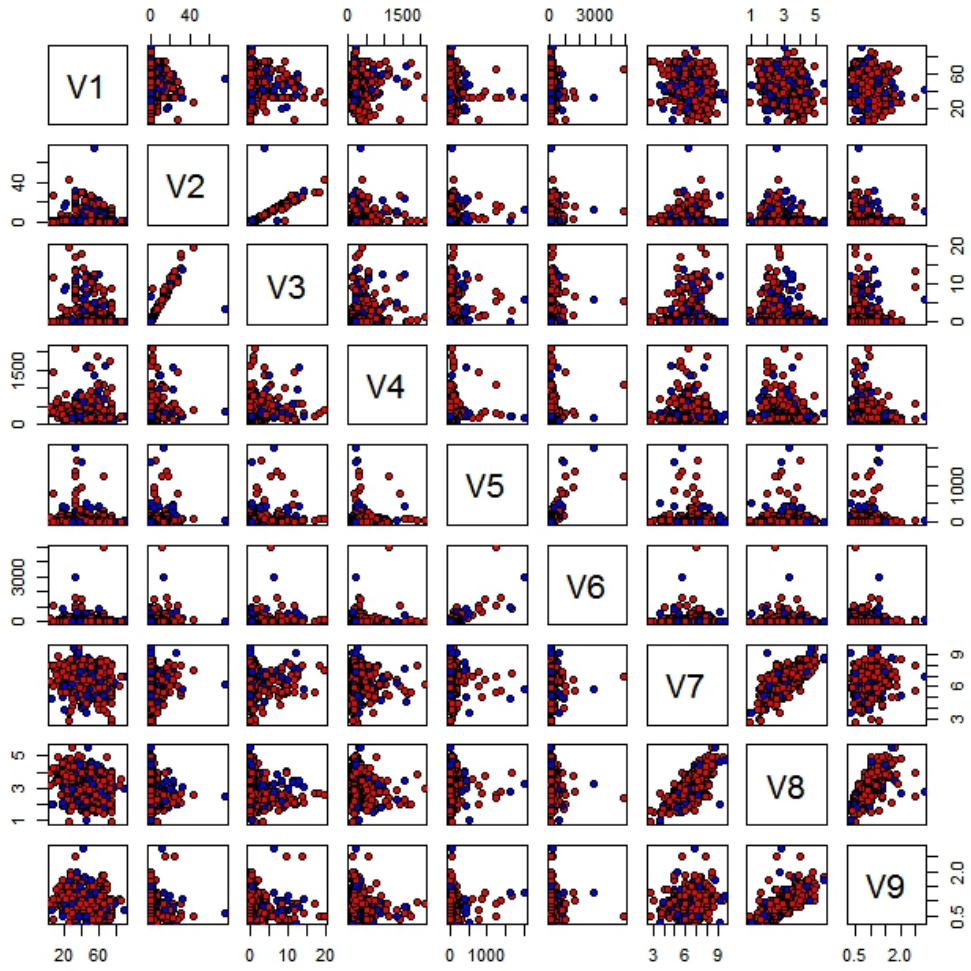
Figura 4.11: Gráfico Q-Q Normal das variáveis de *Honolulu*.



3. **Indian Liver Patient:** Coletado no noroeste de Andhra Pradesh, India. Esse conjunto de dados contém 583 observações de 416 pacientes com doença no fígado e 167 pacientes sem doença no fígado. 9 características são observadas em cada indivíduo.

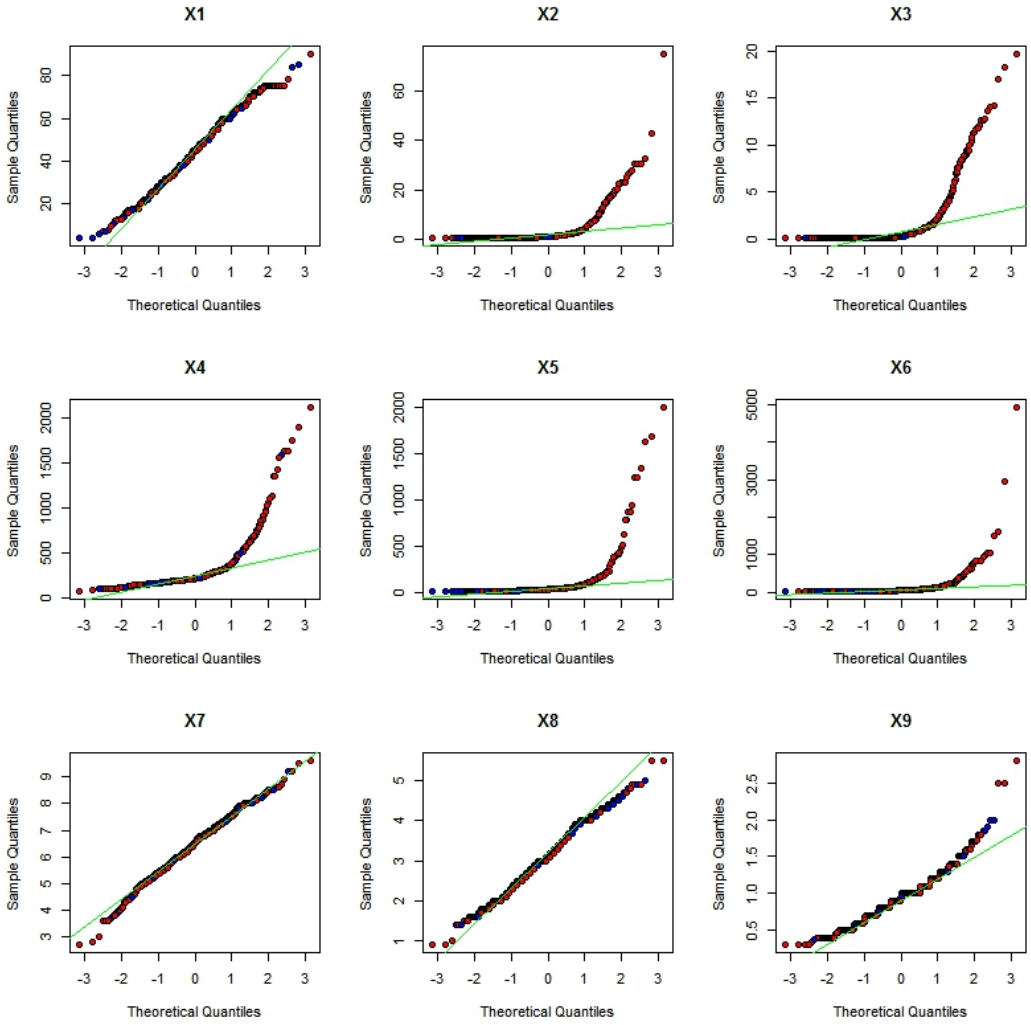
Na Figura 4.12, observamos que há sobreposição das classes em todos os pares de variáveis. Nos pares (V2,V3), (V7,V8) e (V8,V9) observamos uma alta correlação entre essas variáveis. Nas outras combinações entre as variáveis, observamos presença de observações discrepantes, e em algumas delas temos estruturas assimétricas.

Figura 4.12: Distribuição conjunta das observações das variáveis de *Indian Liver Patient*.



Na Figura 4.13, observamos que as variáveis individualmente estão bem afastadas da normalidade, e novamente as observações das classes estão sobrepostas.

Figura 4.13: Gráfico Q-Q Normal das variáveis de *Indian Liver Patient*.

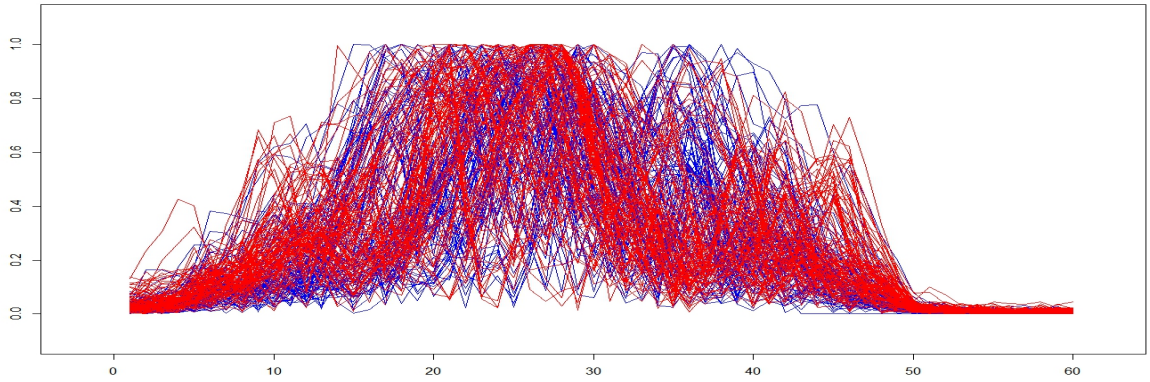


4. **Connectionist Bench (Sonar, Mines vs. Rocks):** Estudo feito para comparar a ressonância em diversas perspectivas diferentes de uma rocha e um mina submarina, afim de encontrar um padrão para que se possam ser distinguidas por um sonar. As observações são de dimensão 60, e são apenas 208 observações sendo 111 de minas e 97 de rochas.

   Na Figura 4.14, podemos observar a grande dificuldade na classificação dessas observações nesse conjunto de dados. As variáveis são altamente correlacionadas pelas características do experimento, assim implicando em maiores dificuldades paras os métodos que supõem independência. Também podemos observar que as classes estão sobrepostas. Uma forma de contornar esses problemas é empregar Componentes

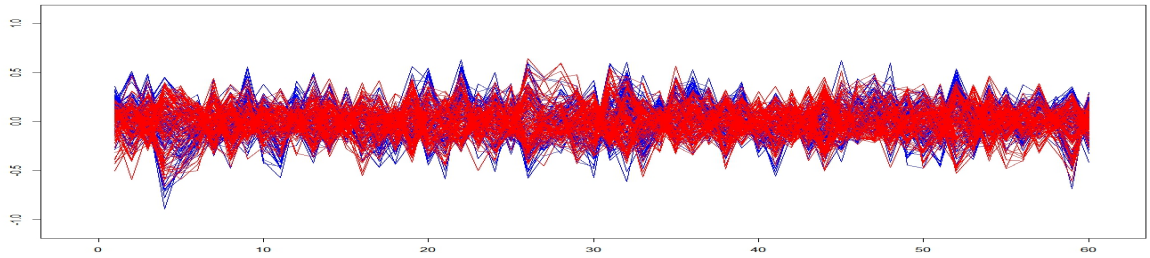
Independentes com o objetivo de separar melhor essas classes.

Figura 4.14: Representação das observações dos dados *Connectionist Bench (Sonar, Mines vs. Rocks)*.

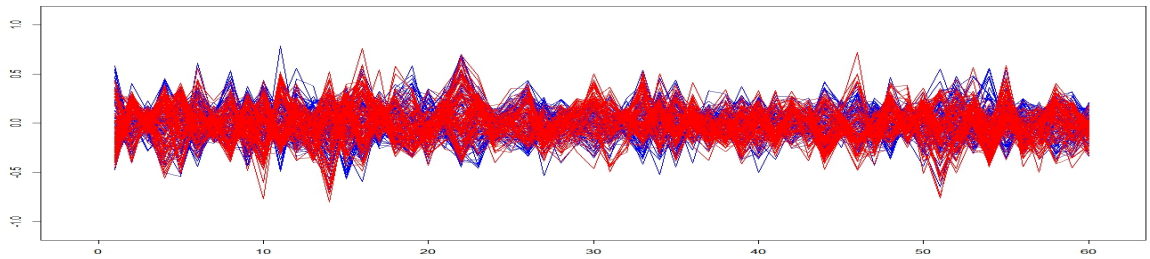


Na Figura 4.15, empregamos a Análise de Componentes Independentes nos dados, na Figura 4.16(a), empregamos um pré-processamento dos dados, centralizando os pelo vetor de médias das observações da Classe 1(representada pela cor vermelha), e na Figura 4.16(b), os dados são centralizados pelo vetor de médias das observações da Classe 2(representada pela cor azul), antes de ser empregado o ICA. Mesmo após esses procedimentos as classes ainda estão sobrepostas.

Figura 4.15: Gráfico em ondas das componentes independentes das variáveis de *Connectionist Bench (Sonar, Mines vs. Rocks)*.



(a) Centradas pelo vetor de médias da Classe 1

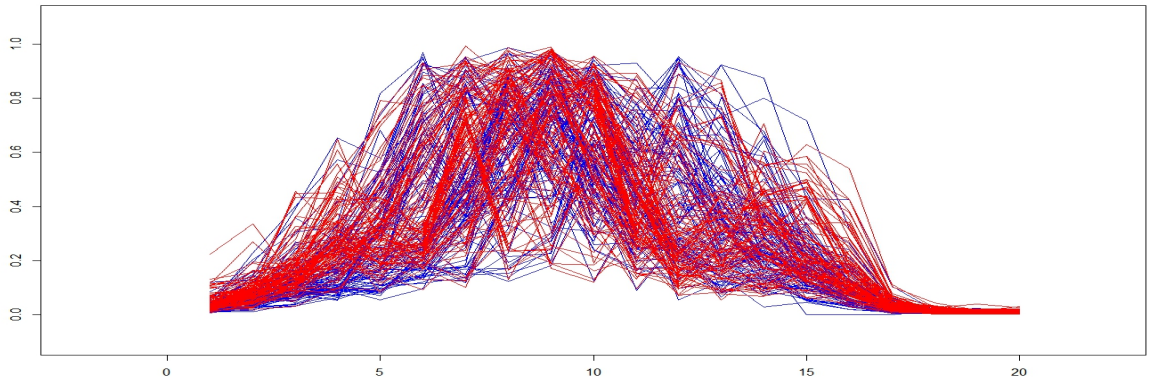


(b) Centradas pelo vetor de médias da Classe 2

5. **Connectionist Bench (Sonar, Mines vs. Rocks) 2:** Similar ao conjunto anterior, com a diferença que a cada 3 observações das variáveis, é calculado um média aritmética dessas 3 observações para formar uma nova variável, assim reduzindo a dimensão de 60 para uma dimensão de 20 variáveis (Cooley & Maceachern, 1998).

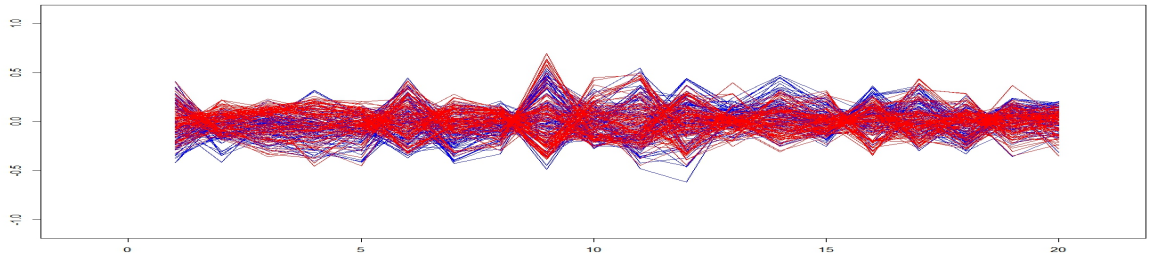
   Na Figura 4.16, mesmo com o procedimento proposto por Cooley & Maceachern (1998), ainda temos uma alta correlação entre as variáveis, e a sobreposição das classes ainda pode ser observada.

Figura 4.16: Representação das observações de *Connectionist Bench (Sonar, Mines vs. Rocks) 2*.

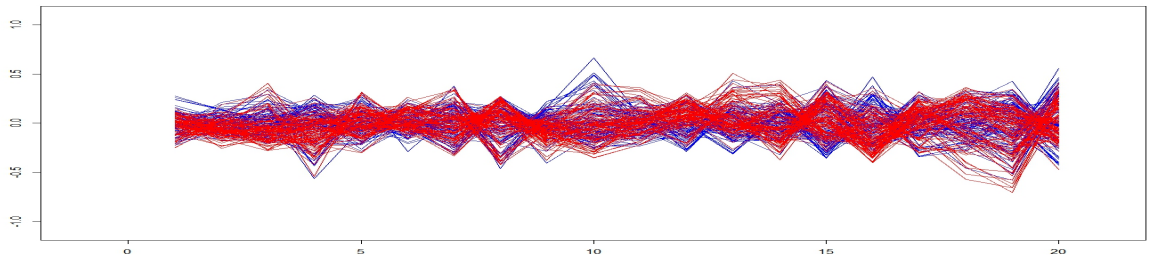


Na Figura 4.17, obtemos os gráficos empregando o mesmo procedimento feito na Figura 4.15. E novamente as classes ainda estão sobrepostas, indicando que possivelmente as componentes independentes não proporcionaram um melhor cenário para classificação.

Figura 4.17: Gráfico em ondas das componentes independentes das variáveis de *Connectionist Bench (Sonar, Mines vs. Rocks)* 2.



(a) Centradas pelo vetor de médias da Classe 1

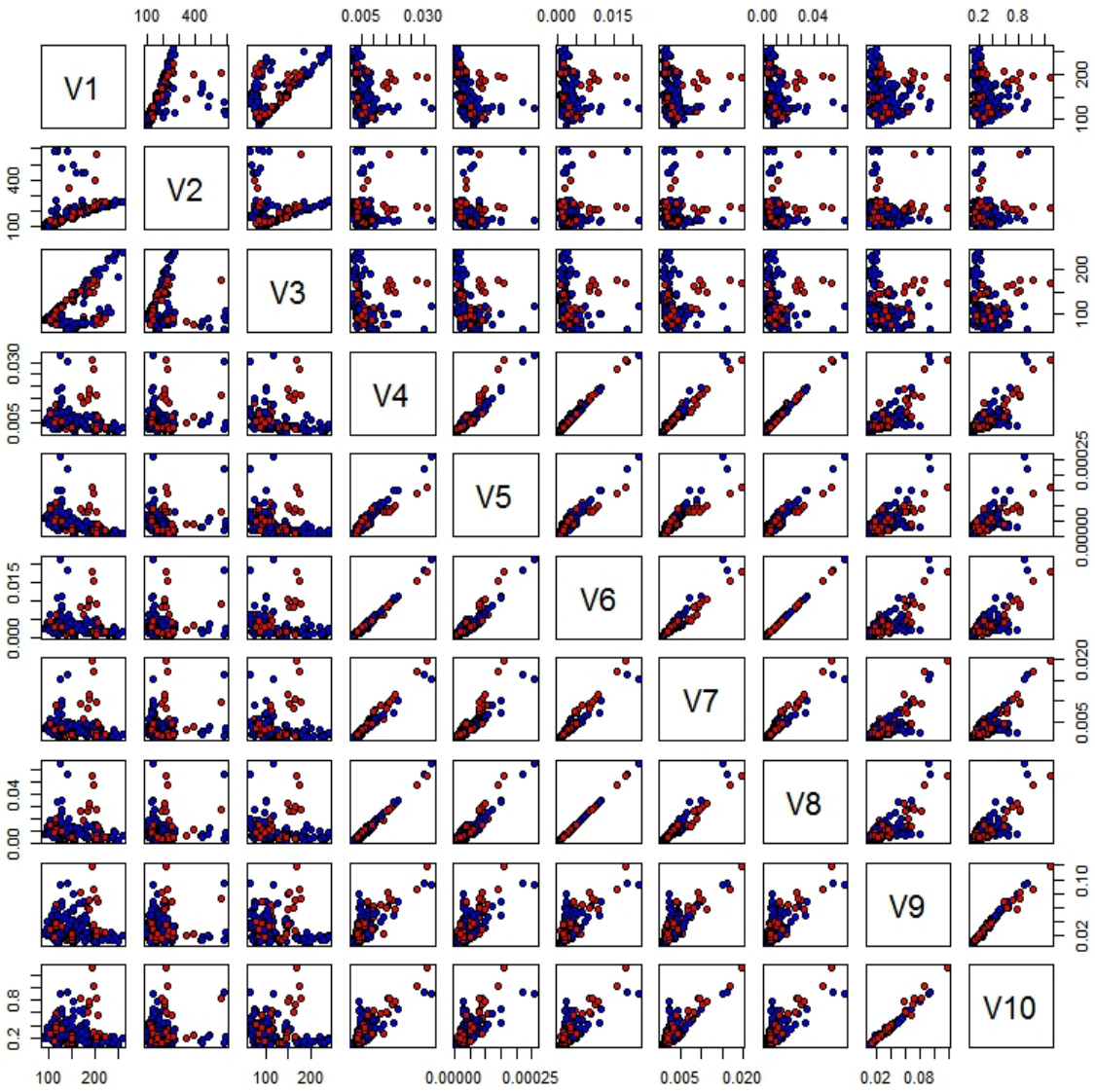


(b) Centradas pelo vetor de médias da Classe 2

6. **Parkinsons Disease:** Estudo feito em 2008 em 195 indivíduos sendo 48 destes identificados como saudáveis e 147 tendo o mal de Parkinson. Em cada indivíduo foram gravadas as vozes e transformadas em sinal digital, extraindo 22 características desse sinal, afim de discriminar um indivíduo com a doença pela voz.

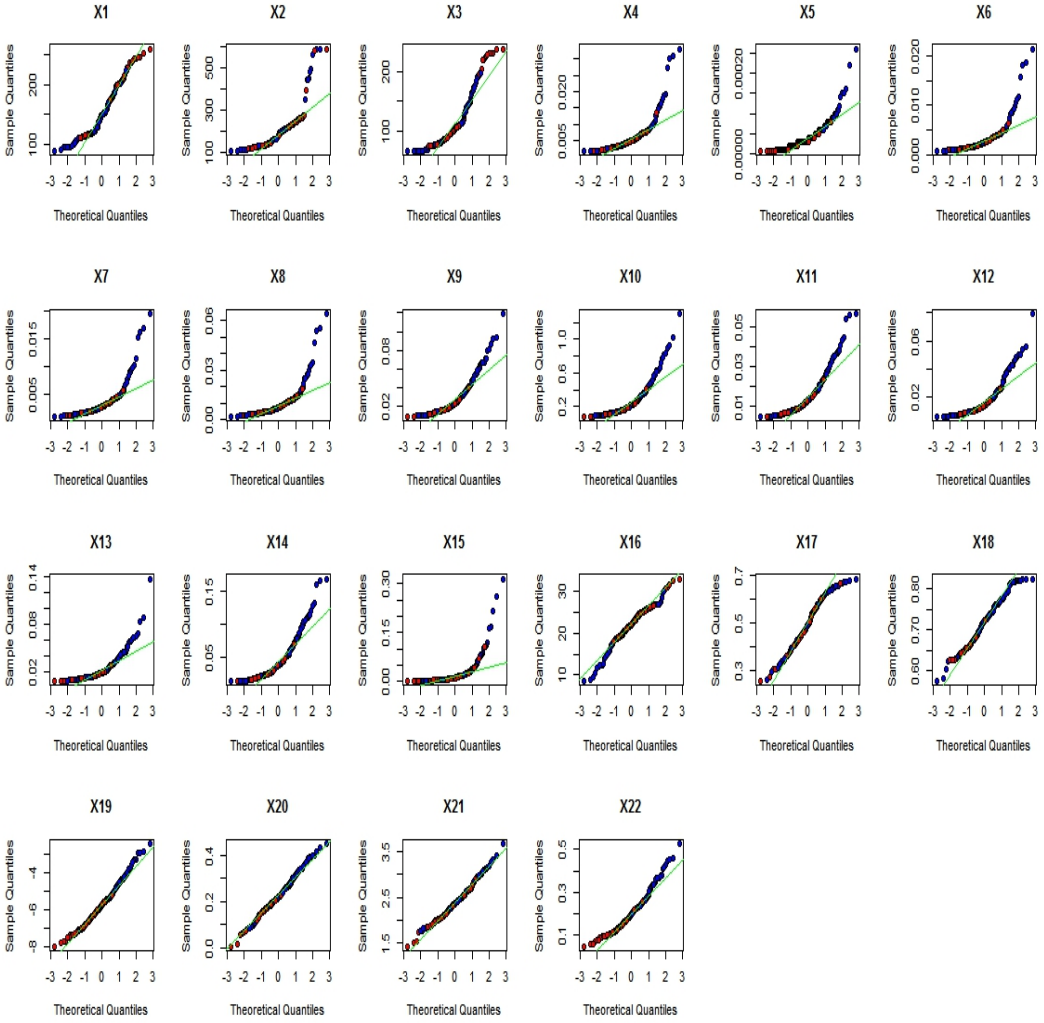
   Na Figura 4.18, apresentamos a distribuição conjunta dos pares de variáveis das 10 primeiras variáveis. Podemos observar que há uma alta correlação entre as variáveis V1, V2 e V3. Em todos os pares de variáveis feitos pela combinação de V4 a V10, também observamos uma correlação positiva muita alta entre essas características. De modo geral, as classes estão bem sobrepostas, e a presença de observações discrepantes e uma estrutura assimétrica, faz desses dados um desafio para os procedimentos de classificação.

Figura 4.18: Distribuição conjunta das observações das 10 primeiras variáveis de *Parkinsons Disease*.



Na Figura 4.19 a seguir, podemos observar que as variáveis não apresentam normalidade individualmente, e isso juntamente com os problemas anteriores implicam em possíveis problemas para os métodos de classificação empregando independência entre as variáveis e normalidade.

Figura 4.19: Gráfico Q-Q Normal das variáveis de *Parkinsons Disease*.



Todos os conjuntos de dados apresentam alta dimensionalidade nos dados e variáveis que não seguem distribuição Normal individualmente. O conjunto que possui menor dimensão é o *Honolulu*, com apenas 3 variáveis, mas tem bem menos observações que os demais. Essa alta dimensionalidade pode implicar em sérios problemas de estimação, principalmente nas abordagens paramétricas por conta da estimação dos parâmetros. O conjunto de dados *Wisconsin Diagnostic Breast Cancer*, apresenta 30 variáveis com apenas 569 observações e separando em classes temos menos observações para estimar as quantidades. É um conjunto de dados bem estudado na literatura e mostra-se um cenário bastante desafiador, em termos de classificação. Os dados de *Honolulu* apesar de terem apenas 3 dimensões, apresentam a característica de maior informação para uma

classe do que a outra, pois uma classe tem bem mais observações que a outra. O *Indian Liver Patient* é um conjunto de dados recente e ainda pouco analisado na literatura (ver informações em Frank & Asuncion (2010)), é uma estrutura de dados que apresenta uma dificuldade por ter 9 dimensões. O conjunto de dados *Connectionist Bench (Sonar, Mines vs. Rocks)*, é a estrutura que apresenta maior desafio para os classificadores, pois tem a maior dimensão dentre os 6 e possuem poucas observações.

Na Tabela 4.2 a seguir, apresentamos as taxas de erro de classificação para os métodos GBPKDA, BPKDA 1, BPKDA 2, ADL, ADQ e NNBDA, aplicados aos 6 conjuntos de dados propostos e empregando o procedimento de Validação Cruzada "K-Fold" com 5 e 10 partições.

Tabela 4.2: Taxa de erro de classificação dos métodos em conjuntos de dados reais.

Validação Cruzada "K-Fold"	**Method**	**Q**	**Data set**					
			1	2	3	4	5	6
5	GBPKDA	5	0.3286	0.3600	0.6021	0.5288	0.4856	0.3590
		7	0.1775	0.3600	0.5386	0.5385	0.4567	0.2256
		9	0.1248	**0.3400**	0.5026	0.5529	0.4567	0.1744
	BPKDA 1	5	0.0808	0.5300	0.3293	0.1731	0.1538	0.1846
		7	0.0598	0.5300	0.3396	0.1683	**0.1442**	0.2051
		9	0.0510	0.5100	0.3756	0.1731	0.1731	0.1436
	BPKDA 2	5	0.0879	0.5000	0.3276	0.1875	0.1827	0.1641
		7	0.0826	0.4900	0.3345	**0.1442**	0.1538	0.1487
		9	0.1213	0.5200	**0.3105**	0.1779	0.1587	0.1744
	ADL		**0.0439**	0.4400	0.3825	0.2067	0.2308	0.1744
	ADQ		0.0492	0.4200	0.4494	0.3077	0.1971	**0.1179**
	NNBDA		0.0703	0.4800	0.4425	0.3173	0.3221	0.3026
10	GBPKDA	5	0.4534	0.3800	0.6329	0.5288	0.5288	0.2051
		7	0.1898	**0.3600**	0.5369	0.5385	0.4567	0.1846
		9	0.1687	0.3800	0.5369	0.5577	0.4567	0.1897
	BPKDA 1	5	0.0756	0.5800	0.3431	0.1538	0.1731	0.1538
		7	0.0773	0.5700	0.3671	**0.1346**	0.1490	0.1692
		9	0.0756	0.4900	0.3688	0.1442	**0.1442**	0.1897
	BPKDA 2	5	0.0844	0.5400	**0.3156**	0.1731	0.1731	0.1744
		7	0.0826	0.5400	0.3225	0.1971	0.1827	0.1949
		9	0.0738	0.5000	0.3259	0.1587	0.1683	0.1436
	ADL		**0.0404**	0.4600	0.3739	0.2260	0.2067	0.1692
	ADQ		0.0439	0.4100	0.4528	0.2644	0.2019	**0.1231**
	NNBDA		0.0738	0.4600	0.4528	0.3173	0.3221	0.3077

Conjunto de Dados: (1) Breast Cancer Wisconsn; (2) Honolulu; (3) Indian Liver Patient; (4) Sonar, Mines vs. Rocks; (5) Sonar, Mines vs. Rocks 2(COOLEY et al, 1998); (6) Parkinsons Disease.

No Conjunto de Dados 1, o método com menor taxa de erro foi o ADL com 4,39%, seguido do ADQ com 4,92%. Os métodos propostos empregando função núcleo, não obtiveram resultados satisfatórios em relação aos métodos clássicos, entretanto o BPKDA 1 com Q igual a 5 e 5 partições na Validação Cruzada, obteve taxa de erro de 5,10% que está bem próximo dos resultados do ADL e ADQ, o que indica que um melhor ajuste em

relação ao valor de Q pode melhorar esse resultado. No Conjunto de Dados 2, apesar das elevadas taxas de erro, o GBPKDA apresentou a menor taxa de erro dentre os métodos, em ambas as partições(5 e 10) da Validação Cruzada. O procedimento BPKDA 2 foi o que obteve as menores taxas de erro no Conjunto de Dados 3 com 31,05% e 31,56% nas Validações Cruzadas considerando 5 e 10 partições, respectivamente. O GBPKDA foi o que apresentou as piores taxas com até 63,29% de erro.

No Conjunto de Dados 4, os métodos com melhores resultados foram o BPKDA 2 com $Q = 7$, que obteve 14,42% erro na Validação Cruzada com 5 partições, e BPKDA 1 com $Q = 7$, apresentando 13,46% de erro na Validação Cruzada com 10 partições. Esses métodos apresentaram os melhores resultados em comparação com os demais, apesar das variáveis serem altamente correlacionadas. No Conjunto de Dados 5, as taxas de erros não tiveram mudanças significativas, mostrando que o método de pré-processamento empregado por Cooley & Maceachern (1998) não influenciou nos resultados. Os métodos com melhores resultados foram o BPKDA 2 com $Q = 7$, que obteve 14,42% erro na Validação Cruzada com 5 partições, e BPKDA 1 com $Q = 7$, apresentando 13,46% de erro na Validação Cruzada com 10 partições. No Conjunto de Dados 6, o método que obteve melhores resultados foi o ADQ com 11,79% e 12,31% nas Validações Cruzadas com 5 e 10 partições, respectivamente. E os métodos propostos por função núcleo não obtiveram resultados próximos do resultado do ADQ.

Capítulo 5

Conclusão e Trabalhos Futuros

5.1 Conclusão

Neste trabalho, apresentamos três novas abordagens não paramétricas de Análise Discriminante baseadas em estimadores por função núcleo adotando uma abordagem Bayesiana, a *Densidade Preditiva Bayesiana por Função Núcleo Multivariada Normal* (GBPKDA), a *Densidade Preditiva Bayesiana por Produto de Funções Núcleo Normais* (BPKDA 1) e a *Densidade Preditiva Bayesiana por Produto de Funções Núcleo Normais empregando Componentes Independentes* (BPKDA 2). As principais vantagens desses métodos são:

1. a não imposição de um modelo probabilístico paramétrico aos dados, ou seja, o procedimento é livre de modelo, com isso, são eliminados os problemas associados a estimação de parâmetros e, como ocorre em algumas aplicações reais, o problema decorrente da alta dimensionalidade e o número de observações insuficientes para estimação;

2. o emprego da modelagem preditiva, permitiu a exclusão de processos de estimação da matriz de largura de banda, bem como sua influência nos resultados da classificação;

3. com a exceção do método GBPKDA, nos demais obtivemos expressões fechadas

para a estimação das densidades sem a necessidade de processos de estimação complicados, e apresentaram resultados satisfatórios em comparação com os métodos tradicionais abordados, considerando a taxa de erro de classificação.

Nos estudo de simulação, os métodos propostos nesse trabalho mostram desempenho similares aos métodos clássicos nos experimentos desenvolvidos, em que os cenários apresentavam menor dificuldade em termos de classificação(Estruturas 1 e 2). E nos demais cenários, os métodos propostos empregando função núcleo apresentaram resultados melhores que os procedimentos tradicionais, principalmente no caso em que havia maior dificuldade para o problema de Análise Discriminante(Estrutura 6), os métodos BKPDA 1 e BPKDA 2 mostraram estimativas das taxas de erro de classificação muito inferiores em relação aos métodos clássicos.

De um modo geral, considerando todos os resultados obtidos nas simulações, o método que mais se destacou foi o BPKDA 2, que mostrou ótimo desempenho, flexibilidade ao apresentar bons resultados em cenários bastante diferentes e empregando misturas finitas de densidades, robustez ao modelar estruturas que apresentam pontos discrepantes e assimetria. Segundo Scott (1992), a abordagem empregando produto de funções núcleo é recomendado na prática, em vez do procedimento empregando função núcleo multivariada, o que pode indicar o porquê do GBPKDA ter tido um desempenho inferior aos BPKDA 1 e BPKDA 2. Apesar do mau desempenho do GBPKDA, em relação com o BPKDA 1 e BPKDA 2 na maior parte dos experimentos, mais estudos sobre sua aplicabilidade em outros tipos de estruturas. Uma melhor análise nos processos de regularização das matrizes $\Delta_{ijc} = (\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})(\mathbf{y}_{jg}^{(m)} - \mathbf{x}_{ig}^{(k)})'$, e a otimização em torno do parâmetro de escalas d_{jg}^{*} pode levar a resultados melhores, uma vez que para um dos conjuntos de dados reais, o método GBPKDA obteve o melhor resultado dentre os quais ele foi comparado.

5.2 Trabalhos Futuros

Nesta seção apresentamos algumas ideias para futuros trabalhos relacionados com o que foi desenvolvido neste trabalho.

- Fazer novos estudos com tamanhos de amostra maiores.
- Avaliar a relação entre o tamanho da partição de treino e a partição de teste no desempenho dos classificadores propostos nesse trabalho, ou seja, verificar se é melhor direcionar mais informação para a estimação dos parâmetros da distribuição t, no caso univariado e multivariado, ou ter mais observações para a estimação do valor da densidade da observação a ser classificada.
- Comparar os procedimentos de classificação com outros critérios, como a curva ROC, do inglês *Receiver Operating Characteristic*.
- Avaliar a influência da quantidade δ que regulariza a matriz Δ_{ijg}, considerando outros valores diferentes do que foi adotado nesse trabalho. E empregar ou desenvolver novas formas de otimizar a escolha dessa quantidade (veja por exemplo Minka (2000)).
- Comparar os métodos propostos com abordagens mais robustas do que ADL, ADQ e NNBDA, como por exemplo a Análise Discriminante empregando modelos de misturas finitas.
- Empregar os métodos propostos juntamente com técnicas de redução de dimensão.

Apêndice A

Algumas Distribuições, Propriedades e Resultados

A.1 Distribuição Normal

Definição A.1.1 (Distribuição Normal matriz variada) *Seja* $\boldsymbol{X}_{(nxp)} \in \mathbb{R}^{nxp}$ *uma matriz aleatória com distribuição Normal matriz variada com média* $\boldsymbol{M}_{(nxp)} \in \mathbb{R}^{nxp}$ *e matriz de covariâncias* $\boldsymbol{\Lambda} \otimes \boldsymbol{\Sigma}$*, onde* $\boldsymbol{\Lambda}_{(nxn)}$ *e* $\boldsymbol{\Sigma}_{(pxp)}$ *são matrizes positivas definidas, com densidade dada por (Iranmanesh* et al.*, 2010):*

$$f(\boldsymbol{X} \mid \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = \frac{1}{(2\pi)^{np/2} \left|\boldsymbol{\Lambda}\right|^{p/2} \left|\boldsymbol{\Sigma}\right|^{n/2}} \exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Lambda}^{-1}(\boldsymbol{X}-\boldsymbol{M})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{M})\right]\right\}. \quad \text{(A.1)}$$

Notação: $vec(\boldsymbol{X}) \sim N_{np}(vec(\boldsymbol{M}), \boldsymbol{\Lambda} \otimes \boldsymbol{\Sigma})$*, onde* $\otimes$ *é o produto de Kronecker e* $vec(\cdot)$ *é a operação de vetorização para notação de matrizes.*

Considerando $n = 1$ e $\boldsymbol{\Lambda}_{(nxn)} = \mathbf{I}_{(nxn)}$, onde $\mathbf{I}$ é a matriz identidade e substituindo em (A.1), temos a seguinte situação.

Definição A.1.2 (Distribuição Normal multivariada) *Seja* $\boldsymbol{X}_{(px1)} \in \mathbb{R}^{p}$ *um vetor aleatório com distribuição Normal multivariada com vetor de médias* $\boldsymbol{\mu}_{(px1)} \in \mathbb{R}^{p}$ *e matriz*

de covariâncias $\mathbf{\Sigma}_{(pxp)}$ *positiva definida, com densidade dada por (Johnson & Wichern, 2007):*

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{p/2} \left|\mathbf{\Sigma}\right|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \tag{A.2}$$

Notação: $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Considerando um caso particular, com $\mathbf{X}$ um vetor aleatório com distribuição Normal multivariada com $p = 1$, $\boldsymbol{\mu} = \mu$ e $\mathbf{\Sigma} = \sigma^2$, dizemos que $\mathbf{X} = X$ tem distribuição Normal univariada, e a transformação $Z = \frac{X-\mu}{\sqrt{\sigma^2}}$ substituindo em (A.2), temos:

Definição A.1.3 (Distribuição Normal Padrão) *Seja Z uma variável aleatória com distribuição Normal univariada com média 0 e variância 1, então sua densidade é dada por (James, 2010)*

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \quad z \in \mathbb{R}, \tag{A.3}$$

Notação: $Z \sim N(0, 1)$.

Assim, podemos fazer a seguinte definição.

Definição A.1.4 (Função de distribuição da distribuição Normal Padrão) *Seja a variável aleatória* $Z \sim N(0, 1)$*, a função*

$$\Phi(z) = \int_{-\infty}^{z} f(z)dz = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz \quad z \in \mathbb{R} \tag{A.4}$$

é denominada de função de distribuição da distribuição Normal Padrão (James, 2010).

A.2 Distribuição Gama Inversa

Definição A.2.1 (Distribuição Gamma Inversa matriz variada) *Seja* $\mathbf{\Sigma}$ *uma matriz aleatória positiva definida de ordem p com distribuição Gama Inversa matriz variada*

com parâmetros $\alpha > (p-1)/2$, $\beta > 0$ *e* $\mathbf{\Omega}(pxp) > 0$, *com densidade dada por (Iranmanesh* et al.*, 2010):*

$$f(\mathbf{\Sigma} \mid \alpha, \beta, \mathbf{\Omega}) = \frac{|\mathbf{\Omega}|^{\alpha}}{\Gamma_p(\alpha)\beta^{\alpha p}} |\mathbf{\Sigma}|^{-\alpha-(p+1)/2} \exp\left\{-\frac{1}{\beta} tr(\mathbf{\Omega}\mathbf{\Sigma}^{-1})\right\} \tag{A.5}$$

onde tr *é a função traço e* Γ_p *é a função gama multivariada definida como*

$$\Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(\alpha - (j-1)/2), \tag{A.6}$$

onde $\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt$. *Notação:* $\mathbf{\Sigma} \sim IMG_p(\alpha, \beta, \mathbf{\Omega})$

Se considerarmos $\alpha = \nu/2$ e $\beta = 2$, e substituindo em (A.5), temos o seguinte caso particular da distribuição Gamma Inversa matriz variada.

Definição A.2.2 (Distribuição Wishart Inversa) *Seja* $\mathbf{\Sigma}$ *uma matriz aleatória positiva definida de ordem* p *com distribuição Wishart Inversa com parâmetro de escala* $\mathbf{\Omega}_{(pxp)} > 0$ *e* $\nu > p-1$ *graus de liberdade, sua densidade é (Gupta & Nagar, 2000):*

$$f(\mathbf{\Sigma} \mid \nu, \mathbf{\Omega}) = \frac{|\mathbf{\Omega}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\nu/2)} |\mathbf{\Sigma}|^{-(\nu+p+1)/2} \exp\left\{-\frac{1}{2} tr(\mathbf{\Omega}\mathbf{\Sigma}^{-1})\right\}. \tag{A.7}$$

Notação: $\mathbf{\Sigma} \sim WI(\nu, \mathbf{\Omega})$.

Sejam $\mathbf{\Sigma} \sim WI(\nu, \mathbf{\Omega})$ e $E(\mathbf{X})$ o valor esperado de $\mathbf{X}$(James, 2010), temos as seguintes propriedades:

Propriedade A.2.1 $E[\mathbf{\Sigma}^{-1}] = \nu\mathbf{\Omega}^{-1}$ *(Gupta & Nagar, 2000).*

Propriedade A.2.2 $E[log\,|\mathbf{\Sigma}|] = log\,|\mathbf{\Omega}| - p\ log2 - \sum_{t=1}^{p} \psi(\nu-p+t)$, *onde* $\psi(z) = \frac{\partial log\Gamma(z)}{\partial z}$ *é a função digamma (Gupta & Srivastava, 2010).*

Agora, considere $p = 1$ e $\beta = 1/\theta$ onde $\theta > 0$ é um parâmetro de escala, e substitua em (A.5), com isso obtemos a seguinte densidade:

Definição A.2.3 (Distribuição Gamma Inversa) *Seja S uma variável aleatória com distribuição Gama Inversa com parâmetros $\alpha > 0$ de forma e $\theta > 0$ de escala, com densidade dada por:*

$$f(s \mid \alpha, \theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} s^{-\alpha-1} \exp\left\{-\frac{\theta}{s}\right\}, \quad s > 0. \tag{A.8}$$

Notação: $S \sim IG(\alpha, \theta)$.

Sejam $S \sim IG(\alpha, \theta)$ e $E(S)$ o valor esperado de S (James, 2010), temos as seguintes propriedades:

Propriedade A.2.3 $E[S^{-1}] = \alpha/\theta$.

Propriedade A.2.4 $E[logS] = logS - \psi(S)$.

A.3 Distribuição t-Student

Definição A.3.1 (Distribuição t matriz variada generalizada) *Seja $\boldsymbol{T}_{(nxp)} \in \mathbb{R}^{(nxp)}$ uma matriz aleatória com distribuição t matriz variada generalizada com parâmetros $\boldsymbol{M} \in \mathbb{R}^{(nxp)}$, $\boldsymbol{\Psi}_{(nxn)} > 0$, $\boldsymbol{\Omega}_{(pxp)} > 0$, $\alpha > (p-1)/2$ e $\beta > 0$, com densidade dada por (Iranmanesh* et al.*, 2010):*

$$\begin{aligned} f(\boldsymbol{T} \mid \alpha, \beta, \boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\Omega}) &= \frac{|\boldsymbol{\Psi}|^{-n/2} |\boldsymbol{\Omega}|^{-p/2} \Gamma_p(\alpha + n/2)}{(2\pi/\beta)^{np/2} \Gamma_p(\alpha)} \\ &\quad \times \left| \boldsymbol{I}_{(nxn)} + \frac{\beta}{2} \boldsymbol{\Omega}^{-1} (\boldsymbol{T} - \boldsymbol{M}) \boldsymbol{\Psi}^{-1} (\boldsymbol{T} - \boldsymbol{M})' \right|^{-(\alpha+n/2)}. \end{aligned} \tag{A.9}$$

Notação: $\boldsymbol{T} \sim T_{n,p}(\alpha, \beta, \boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\Omega})$.

Definição A.3.2 (Distribuição t multivariada) *Seja $\boldsymbol{T}_{(px1)} \in \mathbb{R}^{(px1)}$ um vetor aleatório com distribuição t multivariada com parâmetros $\boldsymbol{\mu} \in \mathbb{R}^{(p)}$ de locação, $\boldsymbol{\Omega}_{(pxp)} > 0$ de*

escala e ν graus de liberdade, com densidade dada por (Johnson & Wichern, 2007):

$$f(\boldsymbol{t} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) = \frac{|\boldsymbol{\Omega}|^{-1/2}\,\Gamma[(\nu+p)/2]}{\pi^{p/2}\Gamma(\nu/2)\nu^{p/2}}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-(\nu+p)/2} \tag{A.10}$$

Notação: $\boldsymbol{T} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu)$.

Definição A.3.3 (Distribuição t) *Seja $T \in \mathbb{R}$ uma variável aleatória com distribuição t univariada com parâmetros $\mu \in \mathbb{R}$ de locação, $\sigma > 0$ de escala e ν graus de liberdade, com densidade dada por (Johnson & Wichern, 2007):*

$$f(t \mid \mu, \sigma, \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}\sigma}\left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2} \tag{A.11}$$

Notação: $T \sim t(\mu, \sigma, \nu)$.

Teorema A.3.1 *Sejam $\boldsymbol{X} \mid \boldsymbol{\Sigma} \sim N_{n,p}(\boldsymbol{M}, \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})$ e $\boldsymbol{\Sigma} \sim IMG_p(\alpha, \beta, \boldsymbol{\Psi})$ então $\boldsymbol{X} \sim T_{n,p}(\alpha, \beta, \boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\Omega})$.*

Demonstração: A densidade de $\mathbf{X} \mid \boldsymbol{\Sigma}$ é dada por:

$$f(\mathbf{X} \mid \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}\,|\boldsymbol{\Omega}|^{p/2}\,|\boldsymbol{\Sigma}|^{n/2}}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Omega}^{-1}(\mathbf{X}-\mathbf{M})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\right]\right\}, \tag{A.12}$$

e a densidade de $\boldsymbol{\Sigma}$ é

$$h(\boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Psi}|^{\alpha}}{\Gamma_p(\alpha)\beta^{\alpha p}}\,|\boldsymbol{\Sigma}|^{-\alpha-(p+1)/2}\exp\left\{-\frac{1}{\beta}tr(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\}. \tag{A.13}$$

Agora, encontramos a densidade de $\mathbf{X}$ empregando a definição de distribuição condicional

(James, 2010), e fazendo $\mathbf{X}-\mathbf{M}=\mathbf{D}$, assim

$$\begin{aligned}
f(\mathbf{X}) &= \int_{\Theta} f(\mathbf{X}\mid\boldsymbol{\Sigma})\, h(\boldsymbol{\Sigma})\, d\boldsymbol{\Sigma}, \text{ onde } \Theta \text{ é o espaço das matrizes positivas definidas,} \\
&= \int_{\Theta} \frac{1}{(2\pi)^{np/2}\left|\boldsymbol{\Omega}\right|^{p/2}\left|\boldsymbol{\Sigma}\right|^{n/2}} \exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Omega}^{-1}\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}\right]\right\} \\
&\quad \times\frac{\left|\boldsymbol{\Psi}\right|^{\alpha}}{\Gamma_p(\alpha)\beta^{\alpha p}}\left|\boldsymbol{\Sigma}\right|^{-\alpha-(p+1)/2}\exp\left\{-\frac{1}{\beta}tr(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\}d\boldsymbol{\Sigma} \\
&= \frac{\left|\boldsymbol{\Psi}\right|^{\alpha}\left|\boldsymbol{\Omega}\right|^{-p/2}}{(2\pi)^{np/2}\beta^{\alpha p}\Gamma_p(\alpha)}\int_{\Theta}\left|\boldsymbol{\Sigma}\right|^{n/2}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Omega}^{-1}\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}\right]\right\} \\
&\quad \times\left|\boldsymbol{\Sigma}\right|^{-\alpha-(p+1)/2}\exp\left\{-\frac{1}{\beta}tr(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\}d\boldsymbol{\Sigma} \\
&= \frac{\left|\boldsymbol{\Psi}\right|^{\alpha}\left|\boldsymbol{\Omega}\right|^{-p/2}}{(2\pi)^{np/2}\beta^{\alpha p}\Gamma_p(\alpha)}\int_{\Theta}\left|\boldsymbol{\Sigma}\right|^{-\alpha-\frac{(n+p+1)}{2}}\exp\left\{-\frac{1}{\beta}tr\left[\frac{\beta}{2}\boldsymbol{\Omega}^{-1}\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}+\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right]\right\}d\boldsymbol{\Sigma} \\
&= \frac{\left|\boldsymbol{\Psi}\right|^{\alpha}\left|\boldsymbol{\Omega}\right|^{-p/2}}{(2\pi)^{np/2}\beta^{\alpha p}\Gamma_p(\alpha)}\int_{\Theta}\left|\boldsymbol{\Sigma}\right|^{-\alpha-\frac{(n+p+1)}{2}}\exp\left\{-\frac{1}{\beta}tr\left[\left(\frac{\beta}{2}\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}+\boldsymbol{\Psi}\right)\boldsymbol{\Sigma}^{-1}\right]\right\}d\boldsymbol{\Sigma}. \qquad (A.14)
\end{aligned}$$

Agora, seja $\mathbf{B}\sim IMG_p(\alpha+n/2,\beta,\boldsymbol{\Psi}_0)$, então do fato de $\int_{\mathbf{B}>0}h(\mathbf{B})d\mathbf{B}=1$, obtemos que

$$\begin{aligned}
\int_{\mathbf{B}>0}\frac{\left|\boldsymbol{\Psi}_0\right|^{\alpha+n/2}}{\Gamma_p(\alpha+n/2)\beta^{\alpha+np/2}}\left|\mathbf{B}\right|^{-\alpha-(n+p+1)/2}\exp\left\{-\frac{1}{\beta}tr(\boldsymbol{\Psi}_0\mathbf{B}^{-1})\right\}d\mathbf{B} &= 1 \\
\frac{\left|\boldsymbol{\Psi}_0\right|^{\alpha+n/2}}{\Gamma_p(\alpha+n/2)\beta^{\alpha p+np/2}}\int_{\mathbf{B}>0}\left|\mathbf{B}\right|^{-\alpha-(n+p+1)/2}\exp\left\{-\frac{1}{\beta}tr(\boldsymbol{\Psi}_0\mathbf{B}^{-1})\right\}d\mathbf{B} &= 1 \\
\int_{\mathbf{B}>0}\left|\mathbf{B}\right|^{-\alpha-(n+p+1)/2}\exp\left\{-\frac{1}{\beta}tr(\boldsymbol{\Psi}_0\mathbf{B}^{-1})\right\}d\mathbf{B} &= \frac{\Gamma_p(\alpha+\frac{n}{2})\beta^{\alpha p}}{\left|\boldsymbol{\Psi}_0\right|^{\alpha+\frac{n}{2}}\beta^{\frac{np}{2}}}. \qquad (A.15)
\end{aligned}$$

Fazendo $\boldsymbol{\Psi}_0=\frac{\beta}{2}\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}+\boldsymbol{\Psi}$ e substituindo (A.15) em (A.14), temos:

$$\begin{aligned}
f(\mathbf{X}) &= \frac{\left|\boldsymbol{\Psi}\right|^{\alpha}\left|\boldsymbol{\Omega}\right|^{-p/2}}{(2\pi)^{np/2}\beta^{\alpha p}\Gamma_p(\alpha)}\frac{\Gamma_p(\alpha+\frac{n}{2})\beta^{\alpha p+\frac{np}{2}}}{\left|\boldsymbol{\Psi}_0\right|^{\alpha+\frac{n}{2}}} \\
&= \frac{\left|\boldsymbol{\Psi}\right|^{\alpha}\left|\boldsymbol{\Omega}\right|^{-p/2}\Gamma_p(\alpha+n/2)}{(2\pi/\beta)^{np/2}\Gamma_p(\alpha)}\left|\boldsymbol{\Psi}_0\right|^{-(\alpha+n/2)} \\
&= \frac{\left|\boldsymbol{\Psi}\right|^{\alpha}\left|\boldsymbol{\Omega}\right|^{-p/2}\Gamma_p(\alpha+n/2)}{(2\pi/\beta)^{np/2}\Gamma_p(\alpha)}\left|\frac{\beta}{2}\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}+\boldsymbol{\Psi}\right|^{-(\alpha+n/2)} \qquad (A.16)
\end{aligned}$$

Note que:

$$\begin{aligned}
\left|\frac{\beta}{2}\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}+\boldsymbol{\Psi}\right| &= \left|\boldsymbol{\Psi}+\frac{\beta}{2}\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}\right| \\
&= \left|\boldsymbol{\Psi}\right|\left|\mathbf{I}_{(pxp)}+\frac{\beta}{2}\boldsymbol{\Psi}^{-1}\mathbf{D}'\boldsymbol{\Omega}^{-1}\mathbf{D}\right| \\
&= \left|\boldsymbol{\Psi}\right|\left|\mathbf{I}_{(nxn)}+\frac{\beta}{2}\boldsymbol{\Omega}^{-1}\mathbf{D}\boldsymbol{\Psi}^{-1}\mathbf{D}'\right|. \qquad (A.17)
\end{aligned}$$

Substituindo novamente, (A.17) em (A.16) e $\mathbf{D} = \mathbf{X} - \mathbf{M}$, temos

$$\begin{aligned} f(\mathbf{X}) &= \frac{|\mathbf{\Psi}|^{\alpha} |\mathbf{\Omega}|^{-p/2} \Gamma_p(\alpha + n/2)}{(2\pi/\beta)^{np/2}\Gamma_p(\alpha)} |\mathbf{\Psi}|^{-(\alpha+n/2)} \left| \mathbf{I}_{(nxn)} + \frac{\beta}{2}\mathbf{\Omega}^{-1}\mathbf{D}\mathbf{\Psi}^{-1}\mathbf{D}' \right|^{-(\alpha+n/2)} \\ &= \frac{|\mathbf{\Psi}|^{-n/2} |\mathbf{\Omega}|^{-p/2} \Gamma_p(\alpha + n/2)}{(2\pi/\beta)^{np/2}\Gamma_p(\alpha)} \left| \mathbf{I}_{(nxn)} + \frac{\beta}{2}\mathbf{\Omega}^{-1}(\mathbf{X} - \mathbf{M})\mathbf{\Psi}^{-1}(\mathbf{X} - \mathbf{M})' \right|^{-(\alpha+n/2)} . \quad \text{(A.18)} \end{aligned}$$

Portanto o teorema fica verificado. □

Corolário A.3.1 *Sejam* $\boldsymbol{X} \mid \mathbf{\Sigma} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ *e* $\mathbf{\Sigma} \sim WI_p(\nu, \mathbf{\Omega})$ *então* $\boldsymbol{X} \sim T_{1,p}(\nu, 2, \boldsymbol{\mu}, 1, \mathbf{\Omega})$.

Demonstração: Basta consideramos $\mathbf{X} \mid \mathbf{\Sigma} \sim N_{n,p}(\mathbf{M}, \mathbf{\Omega} \otimes \mathbf{\Sigma})$ e $\mathbf{\Sigma} \sim IMG_p(\alpha, \beta, \mathbf{\Psi})$ com $n = 1$, $\mathbf{\Omega} = \mathbf{I}_{(nxn)}$, $\beta = 2$ e $\alpha = \nu/2$, assim a demonstração segue analogamente à prova do teorema (A.3.1).

Corolário A.3.2 *Sejam* $X \mid \theta \sim N(\mu, \theta)$ *e* $\theta \sim IG_p(\alpha, \beta)$ *então* $X \sim t(\mu, \sigma^2, 2\alpha)$ *onde* $\sigma^2 = 2\beta(2\alpha)^{-1}$.

Demonstração: A densidade de $X \mid \theta$ é dada por

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{ -\frac{1}{2\theta}(x - \mu)^2 \right\}, \quad \text{(A.19)}$$

e a densidade de $\theta \in \Theta$ é dada por

$$f(\theta \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left\{ -\frac{\beta}{\theta} \right\}. \quad \text{(A.20)}$$

Assim,

$$\begin{aligned} f(x) &= \int_{\Theta} \frac{1}{\sqrt{2\pi\theta}} \exp\left\{ -\frac{1}{2\theta}(x - \mu)^2 \right\} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left\{ -\frac{\beta}{\theta} \right\} d\theta \\ &= \frac{\beta^{\alpha}}{\sqrt{2\pi}\Gamma(\alpha)} \int_{\Theta} \theta^{-1/2}\theta^{-\alpha-1} \exp\left\{ -\frac{1}{2\theta}(x - \mu)^2 - \frac{\beta}{\theta} \right\} d\theta \\ &= \frac{\beta^{\alpha}}{\sqrt{2\pi}\Gamma(\alpha)} \int_{\Theta} \theta^{-(2\alpha+1)/2-1} \exp\left\{ -\frac{\beta}{\theta}\left[1 + \frac{1}{2\beta}(x - \mu)^2 \right] \right\} d\theta. \quad \text{(A.21)} \end{aligned}$$

Seja $B \sim IG((2\alpha+1)/2, \gamma)$, onde $\gamma = \beta\left[1 + \frac{1}{2\beta}(x-\mu)^2\right]$, do fato $\int_0^\infty f(b)db = 1$, temos que

$$\begin{aligned} \int_0^\infty \frac{\gamma^{(2\alpha+1)/2}}{\Gamma((2\alpha+1)/2)} b^{-(2\alpha+1)/2-1} \exp\left\{-\frac{\gamma}{b}\right\} db &= 1 \\ \int_0^\infty b^{-(2\alpha+1)/2-1} \exp\left\{-\frac{\beta\left[1+\frac{1}{2\beta}(x-\mu)^2\right]}{b}\right\} db &= \frac{\Gamma((2\alpha+1)/2)}{\gamma^{(2\alpha+1)/2}}. \end{aligned} \quad (A.22)$$

Substituindo (A.22) em (A.21), temos

$$\begin{aligned} f(x) &= \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \int_\Theta \theta^{-(2\alpha+1)/2-1} \exp\left\{-\frac{\beta}{\theta}\left[1+\frac{1}{2\beta}(x-\mu)^2\right]\right\} d\theta \\ &= \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \frac{\Gamma((2\alpha+1)/2)}{\beta^{(2\alpha+1)/2}\left[1+\frac{1}{2\beta}(x-\mu)^2\right]^{(2\alpha+1)/2}} \\ &= \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \frac{\Gamma((2\alpha+1)/2)}{\beta^{(2\alpha+1)/2}} \left[1+\frac{1}{2\beta}(x-\mu)^2\right]^{-(2\alpha+1)/2}, \\ &\quad \text{fazendo } \beta = \frac{2\alpha\sigma^2}{2}, \\ &= \frac{\left(\frac{2\alpha\sigma^2}{2}\right)^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \frac{\Gamma((2\alpha+1)/2)}{\left(\frac{2\alpha\sigma^2}{2}\right)^{(2\alpha+1)/2}} \left[1+\frac{1}{2\left(\frac{2\alpha\sigma^2}{2}\right)}(x-\mu)^2\right]^{-(2\alpha+1)/2}, \\ &= \frac{\Gamma(\frac{2\alpha+1}{2})}{\sqrt{2\alpha\pi}\ \sigma\ \Gamma((2\alpha)/2)} \left[1+\frac{1}{2\alpha}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(2\alpha+1)/2}. \end{aligned} \quad (A.23)$$

Então $X \sim t(\mu, \sigma^2, 2\alpha)$ onde $\sigma^2 = 2\beta(2\alpha)^{-1}$.

A.4 Distribuições Assimétrica

Definição A.4.1 (Distribuição Normal Assimétrica Multivariada) *Seja $\boldsymbol{X}$ um vetor aleatório de dimensão p com distribuição Normal Assimétrica Multivariada com vetor de locação $\boldsymbol{\mu}_{px1}$, matriz de escalas $\boldsymbol{\Sigma}_{pxp}$ e vetor de forma(assimetria) $\boldsymbol{\lambda}_{px1}$, com densidade*

dada por (Azzalini & Dalla Valle, 1996)

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2N_p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi\left(\boldsymbol{\lambda}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu})'\right) \tag{A.24}$$

Notação: $\boldsymbol{X} \sim SN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$

Definição A.4.2 (Distribuição t Assimétrica Multivariada) *Seja* $\boldsymbol{X}$ *um vetor aleatório de dimensão p com distribuição t Assimétrica Multivariada com vetor de locação* $\boldsymbol{\mu}_{px1}$*, matriz de escalas* $\boldsymbol{\Sigma}_{pxp}$*, vetor de forma(assimetria)* $\boldsymbol{\lambda}_{px1}$ *e* $\nu > 0$ *graus de liberdade, com densidade dada por (Cabral* et al.*, 2012):*

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = 2t_p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)T(\sqrt{\frac{\nu+p}{\nu\Delta(\boldsymbol{x}, \boldsymbol{\mu})}}\boldsymbol{\lambda}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}) \mid \nu+p) \tag{A.25}$$

onde $\Delta(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ *e* $T(z \mid \gamma)$ *é a função de densidade da distribuição t univariada quando* $Z = \frac{X-\mu}{\sigma}$ *onde* $X \sim t(\mu, \sigma, \gamma)$*, denominada de distribuição t padronizada. Notação:* $\boldsymbol{X} \sim St(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$

Apêndice B

Estruturas usadas nas Simulações

Nas estruturas de usadas na simulação, Π_1 denota a classe 1 e Π_2 representa a classe 2.

Estrutura 1

$$\begin{aligned}
\Pi_1 : \mathbf{X} &\sim N\left(\begin{bmatrix}40\\40\end{bmatrix}, \begin{bmatrix}180 & 0\\0 & 180\end{bmatrix}\right)\\
\Pi_2 : \mathbf{X} &\sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}100 & 50\\50 & 100\end{bmatrix}\right)
\end{aligned}$$

Estrutura 2

$$\begin{aligned}
\Pi_1 : \mathbf{X} &\sim 0.6 \cdot N\left(\begin{bmatrix}30\\30\end{bmatrix}, \begin{bmatrix}100 & 70\\70 & 100\end{bmatrix}\right) + 0.4 \cdot Normal\left(\begin{bmatrix}-25\\10\end{bmatrix}, \begin{bmatrix}100 & 70\\70 & 100\end{bmatrix}\right)\\
\Pi_2 : \mathbf{X} &\sim 0.4 \cdot N\left(\begin{bmatrix}25\\5\end{bmatrix}, \begin{bmatrix}100 & 70\\70 & 100\end{bmatrix}\right) + 0.6 \cdot Normal\left(\begin{bmatrix}-30\\-20\end{bmatrix}, \begin{bmatrix}100 & 70\\70 & 100\end{bmatrix}\right)
\end{aligned}$$

Estrutura 3

$$\begin{aligned}
\Pi_1 : \mathbf{X} &\sim 0.5 \cdot St\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}75 & -60\\-60 & 75\end{bmatrix}, \begin{bmatrix}-2\\2\end{bmatrix}, 3\right) \\
&+ 0.5 \cdot Skew-t\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}75 & 60\\60 & 75\end{bmatrix}, \begin{bmatrix}-2\\2\end{bmatrix}, 3\right) \\
\Pi_2 : \mathbf{X} &\sim 0.5 \cdot St\left(\begin{bmatrix}-60\\0\end{bmatrix}, \begin{bmatrix}100 & 0\\0 & 100\end{bmatrix}, \begin{bmatrix}-2\\0\end{bmatrix}, 2\right) \\
&+ 0.5 \cdot Skew-t\left(\begin{bmatrix}60\\0\end{bmatrix}, \begin{bmatrix}100 & 0\\0 & 100\end{bmatrix}, \begin{bmatrix}2\\0\end{bmatrix}, 2\right)
\end{aligned}$$

Estrutura 4

$$\begin{aligned}
\Pi_1 : \mathbf{X} &\sim 0.5 \cdot t\left(\begin{bmatrix}10\\5\end{bmatrix}, \begin{bmatrix}100 & 0\\0 & 100\end{bmatrix}, 5\right) + 0.5 \cdot t\left(\begin{bmatrix}-40\\2.5\end{bmatrix}, \begin{bmatrix}100 & 60\\60 & 100\end{bmatrix}, 5\right) \\
\Pi_2 : \mathbf{X} &\sim 0.5 \cdot t\left(\begin{bmatrix}-10\\-5\end{bmatrix}, \begin{bmatrix}100 & 0\\0 & 100\end{bmatrix}, 5\right) + 0.5 \cdot t\left(\begin{bmatrix}40\\-2.5\end{bmatrix}, \begin{bmatrix}100 & 60\\60 & 100\end{bmatrix}, 5\right)
\end{aligned}$$

Estrutura 5

$$\begin{aligned}
\Pi_1 : \mathbf{X} &\sim 0.3 \cdot N\left(\begin{bmatrix}0\\7.5\end{bmatrix}, \begin{bmatrix}25 & 24.74\\24.74 & 50\end{bmatrix}\right) + 0.2 \cdot Normal\left(\begin{bmatrix}15\\15\end{bmatrix}, \begin{bmatrix}20 & 5\\5 & 5\end{bmatrix}\right) \\
&+ 0.5 \cdot N\left(\begin{bmatrix}30\\10\end{bmatrix}, \begin{bmatrix}20 & -14\\-14 & 20\end{bmatrix}\right) \\
\Pi_2 : \mathbf{X} &\sim 0.5 \cdot N\left(\begin{bmatrix}7.5\\7.5\end{bmatrix}, \begin{bmatrix}15 & -7\\-7 & 15\end{bmatrix}\right) + 0.5 \cdot N\left(\begin{bmatrix}25\\7.5\end{bmatrix}, \begin{bmatrix}15 & 7\\7 & 15\end{bmatrix}\right)
\end{aligned}$$

Estrutura 6

$$
\begin{aligned}
\Pi_1 : \quad & 0.125St\left(\begin{bmatrix}-100\\-100\\-100\\-100\end{bmatrix},\begin{bmatrix}100&70&70&70\\70&100&70&70\\70&70&100&70\\70&70&70&100\end{bmatrix},\begin{bmatrix}10\\10\\10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\100\\-100\\-100\end{bmatrix},\begin{bmatrix}100&70&-70&-70\\70&100&-70&-70\\-70&-70&100&70\\-70&-70&70&100\end{bmatrix},\begin{bmatrix}-10\\-10\\10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\-100\\100\\-100\end{bmatrix},\begin{bmatrix}100&-70&70&-70\\-70&100&-70&70\\70&-70&100&-70\\-70&70&-70&100\end{bmatrix},\begin{bmatrix}-10\\10\\-10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\-100\\-100\\100\end{bmatrix},\begin{bmatrix}100&-70&-70&70\\-70&100&70&-70\\-70&70&100&-70\\70&-70&-70&100\end{bmatrix},\begin{bmatrix}-10\\10\\10\\-10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\100\\100\\-100\end{bmatrix},\begin{bmatrix}100&-70&-70&70\\-70&100&70&-70\\-70&70&100&-70\\70&-70&-70&100\end{bmatrix},\begin{bmatrix}10\\-10\\-10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\100\\-100\\100\end{bmatrix},\begin{bmatrix}100&-70&70&-70\\-70&100&-70&70\\70&-70&100&-70\\-70&70&-70&100\end{bmatrix},\begin{bmatrix}10\\-10\\10\\-10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\-100\\100\\100\end{bmatrix},\begin{bmatrix}100&70&-70&-70\\70&100&-70&-70\\-70&-70&100&70\\-70&-70&70&100\end{bmatrix},\begin{bmatrix}10\\10\\-10\\-10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\100\\100\\100\end{bmatrix},\begin{bmatrix}100&70&70&70\\70&100&70&70\\70&70&100&70\\70&70&70&100\end{bmatrix},\begin{bmatrix}-10\\-10\\-10\\-10\end{bmatrix},2\right)
\end{aligned}
$$

$$
\begin{aligned}
\Pi_2 : \quad & 0.125St\left(\begin{bmatrix}100\\-100\\-100\\-100\end{bmatrix},\begin{bmatrix}100&-70&-70&-70\\-70&100&70&70\\-70&70&100&70\\-70&70&70&100\end{bmatrix},\begin{bmatrix}-10\\10\\10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\100\\-100\\-100\end{bmatrix},\begin{bmatrix}100&-70&70&70\\-70&100&-70&-70\\70&-70&100&70\\70&-70&70&100\end{bmatrix},\begin{bmatrix}10\\-10\\10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\-100\\100\\-100\end{bmatrix},\begin{bmatrix}100&70&-70&70\\70&100&-70&70\\-70&-70&100&-70\\70&70&-70&100\end{bmatrix},\begin{bmatrix}10\\10\\-10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\-100\\-100\\100\end{bmatrix},\begin{bmatrix}100&70&70&-70\\70&100&70&-70\\70&70&100&-70\\-70&-70&-70&100\end{bmatrix},\begin{bmatrix}10\\10\\10\\-10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\100\\100\\-100\end{bmatrix},\begin{bmatrix}100&70&70&-70\\70&100&70&-70\\70&70&100&-70\\-70&-70&-70&100\end{bmatrix},\begin{bmatrix}-10\\-10\\-10\\10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\100\\-100\\100\end{bmatrix},\begin{bmatrix}100&70&-70&70\\70&100&-70&70\\-70&-70&100&-70\\70&70&-70&100\end{bmatrix},\begin{bmatrix}-10\\-10\\10\\-10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}100\\-100\\100\\100\end{bmatrix},\begin{bmatrix}100&-70&70&70\\-70&100&-70&-70\\70&-70&100&70\\70&-70&70&100\end{bmatrix},\begin{bmatrix}-10\\10\\-10\\-10\end{bmatrix},2\right)\\
+ \quad & 0.125St\left(\begin{bmatrix}-100\\100\\100\\100\end{bmatrix},\begin{bmatrix}100&-70&-70&-70\\-70&100&70&70\\-70&70&100&70\\-70&70&70&100\end{bmatrix},\begin{bmatrix}10\\-10\\-10\\-10\end{bmatrix},2\right)
\end{aligned}
$$

Referências Bibliográficas

Amato, U., Antoniadis, A. & Grégoire, G. (2003). Independent component discriminant analysis. *International Journal of Mathematics*.

Andrews, J. L., McNicholas, P. D. & Subedi, S. (2010). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis*, **55**(1), 520–529.

Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, **83**(4), 715–726.

Bernardo, J. M. (1999). Model-free objective bayesian prediction. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*, **93**(3), 295–302.

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons.

Bezdek, J. C. & Pal, S. K. (1992). *Neural Networks for Pattern Recognition: Methods That Search for Structures in Data*. IEEE Press.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Cabral, C. R. B., Lachos, V. H. & Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics and Data Analysis*, **56**, 126–142.

Cooley, C. A. & Maceachern, S. N. (1998). Classification via kernel product estimators. *Biometrika*, **85**(4), 823–833.

de Lima, M. S. & Atuncar, G. S. (2011). A bayesian method to estimate the optimal bandwidth for multivariate kernel estimator. *Journal of Nonparametric Statistics*, **23**(1), 137–148.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under Zero-One loss. *Machine Learning*, **29**, 103–130.

Duda, R. O., Hart, P. E. & Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience, second edition. ISBN 0471056693.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Jounal of the American Statistical Association*, **78**, 316–331.

Ferraty, F. (2010). High-dimensional: a fascinating statistical challenge. *Journal of Multivariate Analysis*, **101**, 305–306.

Frank, A. & Asuncion, A. (2010). UCI machine learning repository [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences.

Garg, G., Prasad, G., Garg, L. & Coyle, D. (2011). Gaussian mixture models for brain activation detection from fMRI data. *International Journal of Bioelectromagnetism*, **13**(4), 255–260.

Ghosh, J. K. & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer.

Gupta, A. K. & Nagar, D. K. (2000). *Matrix Variate Distribution*. Chapman & Hall.

Gupta, M. & Srivastava, S. (2010). Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, **12**, 818–843.

Hand, D. J. & Yu, K. (2001). Idiot's Bayes - Not so stupid after all. *International Statistical Review*, **69**(3), 385–398.

Harville, D. A. (2008). *Matrix Algebra From a Statisticians Perspective*. Springer, NY.

Hastie, T. & Tibshirani, R. (2003). Independent component analysis through product density estimation. *In Advances in Neural Information Processing Systems*, **15**.

Hastie, T., Tibshirani & R. Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, second edition.

Hyvarinen, A. & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, **13**(4-5), 411–430.

Hyvarinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*. John Wiley & sons.

Iranmanesh, A., Arashi, M. & Tabatabaey, S. M. M. (2010). On conditional applications of matrix variate normal distribution. *Iranian Journal of Mathematical Sciences and Informatics*, **5**(2), 33–43.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. Springer, USA.

Jain, A., Duin, R. & Mao, J. (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence*, **22**(1), 4–37.

James, B. R. (2010). *Probabilidade: um curso em nível intermediário*. Instituto Nacional de Matemática Pura e Aplicada, Rio de Janeiro, third edition.

Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, USA, 6th edition. ISBN 0-13-187715-1.

Lamport, L. (1994). *A Document Preparation System*. Addison-Wesley, Massachusetts, second edition.

Marques de Sá, J. P. (2001). *Pattern Recoanition - Concepts, Methods and Applications*. Springer, first edition.

Maugis, C., Celeux, G. & Martin-Magniette, M.-L. (2011). Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, **102**(10), 1374–1387.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, NY.

McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

Minka, T. P. (2000). Inferring a Gaussian distribution. Technical report, MIT.

Prati, R., Batista, G. & Monard, M. (2008). Curvas ROC para avaliação de classificadores. *IEEE Latin America Transactions*, **6**(2).

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, UK. ISBN 052146086.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Stingo, F. & Vannucci, M. (2010). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, **27**(4), 495–501.

Stone, J. V. (2004). *Independent Component Analysis: A Tutorial Introduction*. A Bradford Book.

Sun, D. & Berger, J. (2006). Objective priors for the multivariate normal model. In *ISBA 8th World Meeting on Bayesian Statistics, Alicante, Spain.*.

Theodoridis, S. & Koutroumbas, K. (2008). *Pattern Recognition*. Academic Press, fourth edition.

Wand, M. & Jones, M. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of American Statistical Association*, **88**, 520Ű528.

Webb, G. I., Boughton, J. R. & Wang, Z. (2005). Not so Naive Bayes: Aggregating one-dependence estimators. *Machine Learning*, **58**, 5–24.

West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika*, **78**(2), 421–425.

West, M. & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.