UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

MISTURAS FINITAS DE DENSIDADES COM APLICAÇÕES EM RECONHECIMENTO ESTATÍSTICO DE PADRÕES

Autor: **José Raimundo Gomes Pereira**

Orientador: **Prof. Dr. Lee Luan Ling**

Banca Examinadora:

Prof. Dr. Aluísio de Souza Pinheiro
Prof. Dr. Armando Mario Infante
Prof. Dr. Dalton Soares Arantes
Prof. Dr. João Batista Destro Filho
Prof. Dr. João Marcos T. Romano

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação (FEEC), da Universidade Estadual de Campinas (UNICAMP), como parte dos requisitos exigidos para obtenção do título de Doutor em Engenharia Elétrica.

Maio - 2001

Campinas - SP

Agradecimentos

Ao meu Deus, por tudo que tornou-se realidade;

Ao Prof. Dr. Lee Luan Ling, pela orientação e incentivos constantes;

Ao Prof. Dr. Aluísio de Souza Pinheiro, pela discussões que foram fundamentais para a realização deste trabalho;

Aos membros da Banca Examinadora, pelas sugestões que muito contribuiram para esta forma final do trabalho;

Aos amigos Aluísio, Antônio, Cardoso, Clauber, Leandro e Niwton, com suas respectivas famílias, pela amizade, solidariedade e apoio em todos os momentos;

Aos colegas do LRPRC, pela amizade e disponibilidade que sempre demonstraram;

Aos colegas do Departamento de Estatística do ICE-UA, em particular ao Prof. Dr. José Cardoso Neto, pela compreensão e apoio;

À minha família, pelo amor, solidariedade, compreensão e tudo que necessitamos para vencer;

À minha esposa e meus filhos, de modo muito especial.

Resumo

Neste trabalho, consideramos o problema de Reconhecimento de Padrões Supervisionado (RPS). A proposta é empregar como regra de classificação uma versão empírica do classificador de Bayes, onde estimamos as densidades condicionais através de misturas finitas de densidades normais. Em teoria, com um número suficientemente grande de componentes, uma mistura de normais pode aproximar qualquer densidade. Na prática, uma das dificuldades nessa abordagem, é a seleção de um número de componentes para a mistura, o menor possível, adequado às observações. Para essa questão, propomos selecionar o número de componentes direcionada pelas próprias observações, isso sendo efetivado por procedimentos de seleção de modelos empregando o Critério de Informação de Akaike e o Critério de Informação Bayesiano. É mostrado que pode ser obtida uma regra de classificação consistente para o risco de Bayes com essa abordagem. No desenvolvimento do trabalho, é feita uma revisão das regras estatísticas, paramétricas e não-paramétricas, mais utilizadas em RPS e da teoria dos modelos de misturas finitas. A abordagem proposta é aplicada em problemas com dados simulados e em problemas reais, com seus resultados comparados aos de outros métodos. Dessas comparações, concluimos que essa abordagem se configura numa boa alternativa aos outros métodos de classificação.

Abstract

This thesis deals with a supervised pattern recognition problem: the implementation of an empirical Bayes classifier. The proposed approach is based on the estimation of the conditional probability density functions through a finite mixture of normal densities. The justification of using the finite mixture technique is based on the fact that in theory any probability density function can be approximated by a finite mixture of normal densities for a sufficiently large number of components. However, from a practical point of view, it is desirable to determine the smallest number of normal components suitable for a given set of data. Our approach to this matter is to seek a suitable number of components addressed by the analysis of the data based on the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). We prove that a Bayes risk consistent classification rule can be obtained with the proposed method. For the validation purpose, the proposed approach was tested by simulated and real data and it was compared to other methods suggested in the literature. From these comparisons, we conclude that the proposed approach shows many interesting advantages being a good alternative for statistical pattern classification problems.

Conteúdo

1 Introdução **1**

1.1 O Problema em Reconhecimento de Padrões 1

1.2 RP Supervisionado e RP Não-Supervisionado 4

1.3 Vetor de Características e Classificadores 5

1.4 Abordagem Estatística para RPS 8

1.4.1 A Modelagem do Problema 8

1.4.2 Estimação Paramétrica e Não-Paramétrica 10

1.4.3 Sobre o Conjunto de Treinamento 12

1.5 Modelos de Mistura Finita de Distribuições 13

1.6 A proposta do trabalho 15

1.7 A organização do trabalho 18

2 Classificação Estatística de Padrões **20**

2.1 O Classificador de Bayes 20

2.2 Classificação com Modelos Normais 27

2.3 Classificação Não-Paramétrica 32

2.3.1 Método dos Estimadores por Função-Núcleo 33

2.3.2 Método dos Estimadores pelos k-Vizinhos Mais Próximos 39

2.4 Estimação do Erro de Classificação 45
2.4.1 Método da Contagem dos Erros 46
2.4.2 Estimação por Validação Cruzada 48
2.5 Redução da Dimensionalidade 50
2.5.1 Análise de Componentes Principais 51

3 Misturas Finitas de Densidades **57**
3.1 O Modelo de Mistura Finita de Densidades 57
3.2 Distribuições Marginais 61
3.3 Misturas Finitas Identificáveis 63
3.4 Estimação dos Parâmetros 67
3.4.1 Estimação de Máxima Verossimilhança 68
3.4.2 Estimação Bayesiana 76
3.5 A Estimação do Número de Componentes 78
3.5.1 Teste de Hipóteses para determinar k 79
3.5.2 Seleção de Modelos 80
3.6 Estimação de Máxima Verossimilhança via o algoritmo EM 84
3.6.1 O Algoritmo EM 85
3.6.2 O Algoritmo EM para Mistura Finita de Densidades 93
3.6.3 O EM para componentes na Família Exponencial 100
3.7 Mistura Finita de Densidades Normais 101
3.7.1 Mistura Finita de Normais 102
3.7.2 O Algoritmo EM para as Componentes Normais 103
3.7.3 O EM para as Marginais em Mistura de Normais 110

4 Classificação com Misturas Finitas de Densidades 113

4.1 A Modelagem das Densidades Condicionais 113

4.2 A Estimação da Regra de Bayes 117

4.3 A Seleção dos Números de Componentes 118

4.4 Consistência da Regra de Classificação 122

4.5 A Implementação do Classificador 129

5 Estudos de Simulação e Aplicação 132

5.1 Problemas com Dados Simulados 132

5.1.1 Problema 1 135

5.1.2 Problema 2 139

5.1.3 Problema 3 144

5.1.4 Problema 4 148

5.1.5 Problema 5 151

5.1.6 Problema 6 157

5.1.7 Considerações Sobre os Problemas Simulados 163

5.2 Aplicação com Dados Reais 166

6 Conclusões e Sugestões 174

6.1 Discussões e Conclusões 174

6.2 Sugestões para a Continuidade da Pesquisa 178

A Definições e Conceitos 180

A.1 Algumas Formas de Convergência 180

A.2 Propriedades de Estimadores 181

A.3 Estatística Suficiente e Família Exponencial 183

Bibliografia **186**

Lista de Figuras

5.1 Distribuição das Classes no Problema 1 136
5.2 Distribuição das Classes no Problema 2 140
5.3 Distribuição das Classes no Problema 3 145
5.4 Distribuição das Classes no Problema 4 150
5.5 Distribuição das Classes no Problema 5 152
5.6 Distribuição das Classes no Problema 6 com $d = 2$ 158
5.7 Distribuição das Classes no Problema 6 com $d = 3$ 159
5.8 Critérios AIC e BIC 171

Lista de Tabelas

5.1 Problema 1 – Estimativas do Erro de Classificação (%) 138
5.2 Problema 2 – Estimativas do Erro de Classificação (%) 143
5.3 Problema 2 – Estimativas do Erro de Classificação (%) 144
5.4 Problema 3 – Estimativas do Erro de Classificação (%) 147
5.5 Problema 4 – Estimativas do Erro de Classificação (%) 151
5.6 Problema 5 – Estimativas do Erro de Classificação (%) 154
5.7 Problema 5 – Método MFP com Seleção do AIC e BIC 156
5.8 Problema 6 – Estimativas do Erro de Classificação 161
5.9 Problema 6 – Método MFP com Seleção do AIC e BIC 163
5.10 Taxas de Erro (%) no Subconjunto de Teste 169
5.11 Número de Componentes para os Modelos 170
5.12 Taxas de Erro (%) nas 100 Repetições 172

Capítulo 1

Introdução

Neste capítulo, estabelecemos o problema de Reconhecimento de Padrões, caracterizando-o como Supervisionado ou Não-Supervisionado. Apresentamos as idéias básicas da metodologia estatística para os problemas de Reconhecimento de Padrões e da abordagem empregando-se Misturas Finitas de Densidades. A proposta do trabalho e sua organização são também apresentadas.

1.1 O Problema em Reconhecimento de Padrões

É de conhecimento geral que os seres humanos têm uma grande capacidade em identificar e classificar objetos no cotidiano, até mesmo em situações bastante adversas. Nós recebemos as informações captadas pelo nossos sentidos e imediatamente identificamos as fontes das informações, muitas das vezes sem um esforço consciente para isso. Por exemplo, nós somos capazes de reconhecer faces de pessoas que há muito tempo não víamos,

identificar vozes no telefone mesmo quando a ligação não é de boa qualidade, identificar dígitos , manuscritos ou impressos mecanicamente, apresentados em diferentes tamanhos e inclinações, entre outras coisas.

Em muitas áreas, como em ciências, tecnologia ou até mesmo atividades comerciais, ocorrem situações similares às citadas no parágrafo anterior. A engenharia, biologia, psicologia, medicina, marketing, visão computacional, inteligência artificial e sensoriamento remoto são exemplos de áreas onde ocorrem situações nas quais se faz necessário identificar e/ou classificar "objetos". As situações incluem:

. diagnosticar doenças;

. detectar células anormais em amostras de sangue;

. identificar cidades a partir da leitura do CEP nas correspondências;

. selecionar clientes de bancos com bons riscos de crédito;

. identificar peças defeituosas em uma linha de produção;

. identificar suspeitos de crimes através da impressão digital;

. classificar assinaturas em cheques como falsas ou verdadeiras;

. identificar alvos através de sinais de radar; e

. indentificar áreas com determinados tipos de plantações em imagens de satélites.

Muitas das tarefas descritas acima podem ser efetuadas por seres humanos de uma forma bastante satisfatória. O avanço tecnológico e as necessidades modernas, no entanto,

pressionam no sentido de automatizar a realização dessas tarefas, visando a uma maior precisão, maior rapidez e menor custo, como também, em alguns casos, evitar a exposição dos seres humanos a situações desagradáveis ou até mesmo de risco. A questão, então, é "ensinar" máquinas a efetuar tais tarefas e esse é um problema abordado em Reconhecimento de Padrões.

De forma mais completa, *Reconhecimento de Padrões* (RP) é o estudo de como as máquinas podem observar o meio ambiente, aprender a distinguir os *objetos* de interese, os *padrões*, e tomar decisões seguras sobre a categoria a qual pertence um dado padrão (Jain, Duin e Mao (2000)). O objetivo em Reconhecimento de Padrões, portanto, é esclarecer os complicados mecanismos de identificação/classificação dos seres humanos e automatizar essas funções usando computadores (Fukunaga (1990)).

Como posto acima, o termo *objeto*, ou *padrão*, diz respeito àquilo que nós temos interesse em identificar. Alguns exemplos de padrões seriam uma imagem de uma impressão digital, um paciente sob tratamento médico, um sinal de radar, um sinal de voz, os *pixels* em uma imagem de satélite, uma imagem de uma assinatura em um cheque, entre outros. Às vezes os padrões estão associados a categorias que serão denominadas de *classes*. Alguns exemplos de classes são: em diagnóstico médico, o paciente pode pertencer a classe de *portador de uma doença D*; o sinal de voz pode pertencer a classe de pessoas com *acesso autorizado* a um laboratório; a imagem da assinatura corresponde uma *assinatura verdadeira* de um correntista.

Os problemas de RP, em geral, são considerados como divididos em duas categorias, descritas na seção a seguir.

1.2 RP Supervisionado e RP Não-Supervisionado

O problema de interesse aqui é o de *classificação de padrões*: observado um padrão, estamos interessados em associá-lo a uma classe que seja a mais apropriada às características exibidas por esse padrão. Em muitos casos, as classes a serem considerados no problema são previamente definidas, como nos exemplos dados no parágrafo anterior. Os problemas desse tipo são denominados *Reconhecimento de Padrões Supervisionado* (RPS). Na literatura, as aplicações de RPS também recebem a denominação de *Análise Discriminante* (*Discriminant Analysis*) ou, ainda, *Discriminação* (*Discrimination*).

Em algumas situações reais, além das classes naturalmente definidas para o problema em mãos, é criada uma classe artificial que corresponde aos objetos para os quais existe alguma forma de dúvida no momento da alocação. Em diagnóstico médico, por exemplo, um paciente deve ser alocado à classe de doente ou de não-doente; entretanto, às vezes, torna-se necessário criar uma classe adicional à qual alocar aqueles pacientes cujas observações geram dúvidas no diagnóstico e necessitam de avaliações complementares.

Existem também situações em que as classes são desconhecidas antes da apresentação dos padrões. O número de classes e a composição das mesmas são definidos a partir das informações provenientes das características dos próprios padrões a serem classificados. Os padrões vão sendo agrupados, ou seja, formando as classes, de modo que cada classe seja constituída por padrões que apresentem características com alguma forma de equivalência. Nesses casos, o problema de classificação é denominado *Reconhecimento de Padrões Não-Supervisionado* (RPNS). Na literatura estatística, os problemas de RPNS recebem a denominação de *Análise de Agrupamentos* (*Cluster Analysis* ou *Clustering*).

Neste trabalho, o enfoque é sobre RPS. Salientamos, no entanto, que os problemas de RPNS ocorrem em muitas situações reais. Em Jain *et al.* (2000), os autores apresentam algumas situações cujo problema de classificação é necessariamente de RPNS e discutem várias abordagens para essas aplicações. Algumas referências para RPNS são Hand (1981, Capítulo 7), Jain e Dubes (1988), Fukunaga (1990, Capítulo 11), Ripley (1996, Capítulo 9) e as diversas referências citadas em Jain *et al.* (2000).

1.3 Vetor de Características e Classificadores

Na seção anterior, mencionamos que a classificação é feita com base na informação contida no padrão. Essa informação é determinada através de mensurações feitas sobre o padrão, quantificando, em princípio, os aspectos considerados mais relevantes para diferenciar os padrões que pertençam a classes distintas. Os aspectos mensurados são considerados como observações de variáveis que caracterizam os padrões. Em geral, vários aspectos do padrão são mensurados e, dessa forma, cada padrão gera um conjunto de mensurações que, tomadas conjuntamente, recebem a denominação de *vetor de características* (*feature vector*), que denotaremos por $\mathbf{X}$, enquanto as variáveis que o compõem são denominadas *variáveis preditoras*. Para realizar a classificação, portanto, o que importa é o valor observado $\mathbf{x}$ de $\mathbf{X}$ no padrão a ser classificado. Em virtude disso, na literatura o padrão é simplesmente denominado de *observação* e o problema de RPS é a *predição* da natureza desconhecida de uma dada observação (Devroye, Györfi e Lugosi (1996)).

Uma questão adicional que, como pode ser deduzido dos comentários acima, é de fundamental importância, é a escolha das variáveis preditoras para compor o vetor de carac-

terísticas. Essa discussão não será abordada aqui, onde assumimos que foram selecionadas as melhores variáveis, ou seja, as características que descrevem os objetos de forma mais completa e com suficiente capacidade de discriminação das classes. Para uma discussão sobre seleção de variáveis indicamos Hand (1981, Capítulo 6), McLachlan (1992, Capítulo 12), Ripley (1996, Capítulo 10) e Devroye *et al.* (1996, Capítulo 32).

O vetor de características pode ser visto como um ponto em um espaço de d-dimensional, onde d é o número de variáveis em $\mathbf{X}$, a dimensão de $\mathbf{X}$ ($d = dim(\mathbf{X})$). Esse espaço é às vezes denominado *Espaço das Características*. O que se busca é escolher as variáveis preditoras que permitam que os vetores de características de classes diferentes ocupem regiões compactas disjuntas no espaço das características. Nesse contexto, o problema de classificação seria estabelecer fronteiras que separem os pontos correspondentes a classes distintas, fronteiras essas que recebem a denominação de *Fronteiras de Decisão*.

Em algumas aplicações, o número de características mensuradas se torna excessivamente grande, como nas aplicações onde o padrão é uma imagem e as classes são texturas. Nesses casos, é importante tentar reduzir a dimensão de $\mathbf{X}$ de uma maneira que sua capacidade de discriminação seja preservada. Essas questões estão associadas aos problemas denominados na literatura como *seleção de características* (*feature selection*) e *extração de características* (*feature extraction*). Na seleção de características, busca-se selecionar um número de características $m < d$ que melhore o desempenho do classificador. Na extração de características, o objetivo é obter transformações do vetor de características visando a uma redução de dimensão para $d^{'} < d$. Alguns autores não fazem distinção entre esses dois problemas e os consideram simplesmente como um problema de redução de dimensão (Duda e Hart (1973), Fukunaga (1990)). Em Jain *et al.* (2000), são discutidos métodos

para lidar com seleção e extração de características e são indicadas várias referências sobre esses assuntos.

Uma questão importante é que, em geral, a relação entre os vetores de características e as classes não é determinística. Na prática, o vetor de características varia entre padrões de uma mesma classe; por exemplo, os sintomas de um paciente com uma determinada doença não são exatamente os mesmos de um outro paciente com a mesma doença ou, ainda, a assinatura de uma indivíduo apresenta variações a cada vez que ele assina um cheque.

Observado o vetor de características para um padrão cuja classe é desconhecida, é necessário um procedimento para tomar a decisão de como classificar esse padrão. O procedimento estabelecido é denominado de *classificador* ou *regra de alocação* ou, ainda, *função de decisão*. O classificador, portanto, é uma função de $\mathbf{X}$ e, sendo $r(\cdot)$ o classificador definido para o problema, o valor de $r(\mathbf{x})$ indicará a qual classe o padrão com observação $\mathbf{x}$ deve ser alocado. Surgem, então, duas questões: como construir o classificador e como avaliar o seu desempenho. Para a construção do classificador, em RPS, dispomos de um conjunto com exemplos distintos de padrões das classes, isto é, observações de $\mathbf{X}$ para padrões em cada uma das classes. A avaliação do desempenho faz-se necessária para termos informação sobre o que podemos esperar com respeito aos erros na classificação dos padrões. Essas questões serão abordadas neste trabalho.

O conjunto de exemplos mencionado no parágrafo acima é denominado de *Conjunto de Treinamento*. Considerando que temos M classes e que cada classe tem n_j exemplos, $j = 1, 2, 3, \cdots, M$, como cada padrão é descrito pela correspondente observação de $\mathbf{X}$, o

conjunto de treinamento será denotado por

$$\mathcal{T}_{(n)} = \{\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \mathbf{x}_{j,3}, ..., \mathbf{x}_{j,n_j};\ j = 1, 2, 3, ..., M;\ \sum_{j=1}^{M} n_j = n\},$$

onde $\mathbf{x}_{j,i}$ representa a observação do vetor de características para i-ésimo objeto da j-ésima classe. O termo *exemplos rotulados* é às vezes empregado na literatura para designar as observações em $\mathcal{T}_{(n)}$. O emprego desse termo é para diferenciar dos exemplos de padrões nos problemas de RPNS, cujos exemplos são ditos *não rotulados* pois, como mencionado, as classes não são conhecidas previamente.

Pelo exposto, o planejamento de um sistema de Reconhecimento de Padrões envolve basicamente três componentes: (1) a aquisição das medições das características; (2) a seleção e extração das características; e (3) a definição e avaliação do classificador. Às vezes, é estabelecido um processo de iteração entre as etapas (2) e (3) visando melhorar o desempenho do classificador. Dessas três etapas, a definição do classificador e a avaliação do seu desempenho, são os interesses centrais deste trabalho.

1.4 Abordagem Estatística para RPS

1.4.1 A Modelagem do Problema

O problema de RPS pode ser colocado da seguinte forma: suponha que tenhamos um problema com M classes e considere uma variável Y que assume valores em $\{1, 2, 3, \ldots, M\}$. O valor de Y indica a classe a qual pertence um dado padrão. Nós conjecturamos que existe uma função desconhecida $\mathcal{F}(\cdot)$ que associa a $\mathbf{X}$ o valor de Y. A relação estabelecida

por $\mathcal{F}(\cdot)$ é modelada como tendo uma parte determinística e uma componente aleatória. O problema, então, é *estimar* essa função desconhecida ou, ainda, *construir* um classificador $r(\cdot)$ que aproxime a função $\mathcal{F}(\cdot)$. Abordagens distintas podem ser consideradas para determinar o classificador $r(\cdot)$.

Na abordagem estatística, que recebe a denominação de *Classificação Estatística de Padrões* ou, num sentido mais amplo, *Reconhecimento Estatístico de Padrões*, o vetor de características é considerado como um vetor aleatório. Para cada classe, o comportamento de $\mathbf{X}$ é modelado por uma distribuição de probabilidade que é denominada *distribuição condicional da classe*. Nessa modelagem, para cada classe, é também incluída a probabilidade do padrão provir da classe, sendo essa probabilidade denominada *probabilidade a priori*. A relação de $\mathbf{X}$ com as classes é, então, estabelecida através das distribuições condicionais e das probabilidades a priori. O objetivo nessa modelagem é descrever a incerteza inerente a $\mathbf{X}$, procurando caracterizá-lo em termos de seu comportamento mais representativo.

Como será visto neste trabalho, as distribuições condicionais e as probabilidades a priori são combinadas de uma forma apropriada, a saber, através do *Teorema de Bayes*, para que sejam obtidas as *probabilidades a posteriori das classes*. A probabilidade a posteriori de uma dada classe é a probabilidade de o padrão provir dessa classe, condicionada ao valor observado de $\mathbf{X}$. A idéia básica, portanto, é que o classificador deve alocar o padrão à classe com a maior probabilidade a posteriori.

Na construção do classificador, podem ser consideradas duas abordagens distintas: estimar diretamente as probabilidades a posteriori ou estimar as distribuições condicionais com as respectivas probabilidades a priori para as classes. Rubinstein e Hastie (1997) denominam

essas abordagens, respectivamente, como *aprendizagem discriminativa* e *aprendizagem informativa*. A aprendizagem informativa, às vezes denominada abordagem por *estimação de densidade*, é a abordagem de interesse neste trabalho e será discutida em maiores detalhes.

Para a aprendizagem discriminativa, várias referências discutem os métodos destinados a essa abordagem, como por exemplo, Ripley (1996). Em Holmström, Koinstinen, Laaksonen e Oja (1997), onde denomina-se a aprendizagem discriminativa como abordagem por *técnicas de regressão*, é feita uma descrição simplificada dos métodos para essa abordagem e são indicadas várias referências para uma discussão mais aprofundada (e atualizada) desses métodos. Nesse artigo, também é discutido o emprego de redes neurais artificiais para construção de classificadores, sendo as redes empregadas para estimar as probabilidades a posteriori das classes. Para o emprego de redes neurais artificiais em problemas de RP em geral, veja Bishop (1995).

1.4.2 Estimação Paramétrica e Não-Paramétrica

Para a modelagem considerada na aprendizagem informativa, é necessário estimar as distribuições condicionais e as probabilidades a priori das classes. Em Holmström *et al.* (1997), é observado que, na prática, a estimação das probabilidades a priori não apresenta grandes dificuldades, sendo a parte mais difícil a estimação das distribuições condicionais. Na estimação dessas distribuições, duas abordagens são consideradas usualmente: a *estimação paramétrica* e a *estimação não-paramétrica*. Na estimação paramétrica, as distribuições condicionais são modeladas através de famílias de distribuições paramétricas,

cuja forma funcional é conhecida sendo desconhecidos apenas os valores dos seus parâmetros. As observações em $\mathcal{T}_{(n)}$ são empregadas para estimar esses parâmetros desconhecidos. A distribuição normal multivariada é o modelo mais empregado na prática, sendo as distribuições das classes diferenciadas pelos parâmetros dessa família de distribuições: o vetor de médias e a matriz de covariâncias.

De forma diferente, na estimação não-paramétrica nenhuma suposição é feita sobre a forma funcional das distribuições condicionais, sendo essas distribuições estimadas diretamente a partir das observações em $\mathcal{T}_{(n)}$. Os *estimadores por função núcleo* (*kernel estimators*), também denominados de *janelas de Parzen*(*Parzen windows*), e os métodos dos *k-vizinhos mais próximos*(*k-nearest neighbour*) são os métodos mais utilizados em estimação não paramétrica, o primeiro por suas propriedades teóricas, exaustivamente estudadas, e o segundo por facilidade de implementação (veja, por exemplo, Hand (1982) e Hand (1981)).

Dessas idéias iniciais, vemos que os métodos paramétricos são mais simples de serem utilizados, uma vez que, sendo suposto um certo modelo para as distribuições (desconhecidas) nas classes, é necessário apenas estimar os valores dos parâmetros. Em muitas aplicações, no entanto, qualquer suposição a respeito das distribuições nas classes é completamente inadequada. Para esses casos, os métodos não-paramétricos são mais seguros por sua flexibilidade em não incorporarem restrições com respeito às distribuições subjacentes. Em teoria, os métodos não-paramétricos podem lidar com distribuições de qualquer nível de complexidade. Na prática, por outro lado, os modelos empregados apresentam uma complexidade crescente com o número de observações e são inadequados para lidar com dados de dimensão alta (Scott (1992)).

No Capítulo 2 deste trabalho, os métodos paramétricos e não-paramétricos citados nesta

seção serão descritos e discutidas suas propriedades mais relevantes com respeito aos problemas de classificação estatística de padrões.

1.4.3 Sobre o Conjunto de Treinamento

Com a modelagem empregada na Classificação Estatística, as observações em $\mathcal{T}_{(n)}$ são consideradas como valores observados de uma variável aleatória. Temos, portanto, que as observações de cada uma das classes é suposta ser uma *amostra aleatória* da respectiva distribuição de probabilidade, ou seja, *observações independentes e identicamente distribuídas*. Nesse ponto, faz-se necessário comentar dois aspectos relativos a $\mathcal{T}_{(n)}$, um com relação ao número de observaçãoes por classe, n_j, e outro com respeito à independência das observações.

O número de observações por classe pode ser obtido de duas formas: fixando-se n_j para cada classe ou selecionando-se aleatoriamente um conjunto de n padrões e determinando-se os $n_{j's}$ pelo número de padrões observados para cada uma das classes. No segundo caso, as proporções amostrais n_j/n dariam informações a respeito das proporções das classes enquanto que, para n_j fixado, $\mathcal{T}_{(n)}$ não conteria essa informação. A situação de interesse aqui é a do segundo caso, pois, na maioria das aplicações, as probabilidades a priori das classes não são conhecidas e, como será visto, elas são estimadas com base nas proporções n_j/n. Em alguns problemas reais pode haver um especialista que disponha de alguma informação sobre essas probabilidades.

Com respeito a independência das observações em $\mathcal{T}_{(n)}$, para algumas aplicações, isso não é válido. Nas aplicações em sensoriamento remoto, por exemplo, as observações do conjunto

de treinamento correspondem aos *pixels* das imagens (às vezes considerando uma vizinhança do *pixel*) e, dessa forma, a independência não se verifica. Na prática, no entanto, métodos que supõem independência no conjunto de treinamento, mesmo sob a condição de dependência, apresentam desempenhos considerados satisfatórios (veja, por exemplo, Popat e Picard (1997b)). Neste trabalho, é suposto a independência das observações e indicamos McLachlan (1992, Capítulo 13) para uma discussão e referências adicionais sobre métodos para análise estatística de imagens que não supõem a independência das observações.

1.5 Modelos de Mistura Finita de Distribuições

Os modelos de *mistura finita de distribuções* considerados aqui, são combinações lineares de distribuições. O modelo com k componentes, ou de *dimensão* k, é da forma

$$p(\mathbf{x}) = \sum_{l=1}^{k} \alpha_l f_l(\mathbf{x}),$$

onde as componentes $f_l(\cdot)$ são densidades (funções densidade de probabilidade ou funções de probabilidade) e os coeficientes são restritos a $\alpha_l \geq 0$ e $\sum_{l=1}^{k} \alpha_l = 1$. Na prática, as densidades $f_l(\cdot)$ são especificadas como distribuições paramétricas e, nesse caso, cada componente é definida por um vetor de parâmetros $\boldsymbol{\theta}_j$. Nesse contexto, o modelo de mistura é completamente determinado pelo vetor de parâmetros $\boldsymbol{\Phi} = (\alpha_1, \alpha_2, \ldots, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$.

O principal emprego dos modelos de misturas finitas tem sido para definir uma abordagem formal para os problemas RPNS. Com esssa abordagem, denominada *agrupamento baseado em modelos* (*model-based clustering*), é suposto que as observações são geradas por uma

mistura de distribuições, sendo cada componente da mistura representante de um grupo (*cluster*) e os respectivos coeficientes as probabilidades de a observação provir do respectivo grupo (Banfield e Raftery (1993), Fraley e Raftery (1998)). Os aspectos teóricos do emprego de modelos de misturas finitas em RPNS são discutidos em maior profundidade em McLachlan e Basford (1988).

O aspecto de interesse neste trabalho, é que os modelos de misturas finitas podem ser empregados como uma *classe geral de modelos para representar densidades de qualquer complexidade* (Ripley (1996, Capítulo 6), Jain *et al.* (2000)). Dessa forma, esses modelos são convenientes para modelar as distribuições condicionais das classes nos problemas de RPS. Do ponto de vista teórico, um modelo de mistura com dimensão e componentes escolhidas "apropriadamente", pode aproximar qualquer distribuição com uma precisão arbitrária. Em particular, qualquer distribuição contínua multivariada pode ser aproximada arbitrariamente por uma mistura finita de densidades normais com a mesma matriz de covariâncias, como observado em McLachlan e Peel (2000, Seção 6.1).

A modelagem por misturas finitas, às vezes denominada *estimação semiparamétrica*, combina as vantagens dos métodos paramétricos e não-paramétricos. Esses modelos não são restritos a uma forma funcional específica, como no caso paramétrico, e sua complexidade (em termos de dimensão) cresce com a complexidade do problema e não simplesmente com o tamanho do conjunto de dados, como ocorre no caso não-paramétrico (Bishop (1995)).

Em algumas aplicações, há indicações de que as distribuições das classes são descontínuas e, nesses casos, fica mais evidente a conveniência do emprego de modelos de mistura. Esses casos são exemplificados em Zhu e Cai (1998), na classificação de letras impressas em fontes distintas e no reconhecimento de palavras pronunciadas por pessoas diferentes.

O modelo de mistura finita de densidades normais tem sido o mais empregado nas aplicações que aparecem na literatura. Oliver, Pulsen, Toussaint e Louis (1979) empregam uma mistura de normais para modelar as densidades das classes para um problema em citologia, a classificação de células como normais ou anormais; Sclove (1983) propõe a segmentação de imagens modelando a distribuição dos *pixels* por uma mistura de densidades normais; Roeder (1990) estima a densidade de observações relativas à velocidade com que determinadas galáxias se afastam da nossa empregando uma mistura de normais; Tråvén (1991) usa uma mistura de normais com uma abordagem por redes neurais, aplicando-a na classificação de *pixels* em imagem e em reconhecimento de voz; Kurita, Otsu e Abdelmalek (1992) modelam histogramas de imagens por uma mistura de normais, utilizando "Thresholding" de Máxima Verossimilhança em segmentação de imagens; Hastie e Tibshirani (1996)utilizam mistura de normais para modelar a distribuição das classes na classificação de dígitos manuscritos.

Os modelos de mistura finita se constituem no interesse central deste trabalho e serão discutidos num contexto mais formal. No Capítulo 3, as suas propriedades e as dificuldades inerentes a seu emprego na prática como, por exemplo, a estimação dos parâmetros e a determinação da dimensão do modelo, serão abordadas em maior profundidade.

1.6 A proposta do trabalho

Foi mencionado que modelar as distribuições condicionais empregando modelos paramétricos específicos impõe uma restrição quanto à forma funcional dessas distribuições. A estimação não paramétrica tem inconvenientes com relação à complexidade dos modelos,

principalmente, com vetores de características de dimensão alta. Em muitas aplicações, o vetor de características tem dimensão alta, por exemplo $d > 10$, e se torna inadequado restringir as distribuições das classes a qualquer forma funcional específica.

A proposta deste trabalho é *modelar as distribuições condicionais empregando misturas finitas de densidades*. A motivação é o fato de que as misturas podem modelar distribuições de probabilidades arbitrárias, mesmo em situações onde não hajam razões para supor que a distribuição verdadeira seja uma mistura (Popat e Picard (1997b)). A precisão da aproximação, no entanto, depende das densidades componentes utilizadas e a determinação da dimensão adequada aos dados, como já mencionado.

Com respeito as componentes, serão empregadas *misturas finitas de densidades normais*. Essa escolha decorre, em parte, da abrangência que a distribuição normal apresenta em termos de modelagem, bem como sua relativa facilidade de manuseio com relação as questões computacionais. Além disso, como mencionado na Seção 1.5, com misturas de densidades normais é possível aproximar qualquer distribuição contínua com uma precisão arbitrária. Definidas as densidades componentes, permanecem as questões de como estimar os parâmetros que definem o modelo e como determinar um número de componentes para o modelo.

O método de *estimação de máxima verossimilhança* será empregado para a estimação dos parâmetros. Será visto que a estimação em misturas finitas apresenta algumas dificuldades. Resultados teóricos, no entanto, permitem contornar essas dificuldades, em particular, para as misturas com componentes normais. Para a determinação das estimativas, faz-se necessário o emprego de métodos iterativos, sendo para esse fim utilizado o algoritmo *EM*(*Expectation and Maximization*) que, no caso de mistura de normais, tem

sua implementação bastante simplificada.

Com relação a determinação do número de compoentes, em algumas aplicações, um número de componentes fixo é imposto ao modelo (por exemplo, Hastie e Tibshirani (1996)) ou, ainda, é feito um agrupamento das observações e a cada grupo formado é associada uma componente (Popat e Picard (1997b)). Para estimar a dimensão do modelo, serão adotados procedimentos baseados na *teoria geral de seleção de modelos*, cuja idéia é estimar a dimensão de forma que uma dada função-critério, envolvendo a qualidade do ajuste e a complexidade do modelo, seja otimizada. Dois critérios serão utilizados: o *Critério de Informação de Akaike* e o *Critério de Informação Bayesiano*. Nós utilizaremos as siglas AIC e BIC da teminologia original para denotar, respectivamente, os dois critérios: AIC de *Akaike's Information Criterion* e BIC de *Bayesian Information Criterion* .

Os critérios AIC e BIC foram desenvolvidos sobre bases teóricas distintas, mas têm o objetivo comum de comparar a adequação de modelos às observações. Nas aplicações tem sido observado que o AIC apresenta uma tendência de selecionar modelos com dimensão maior do que aqueles selecionados pelo BIC (Kass e Raftery (1995)). No caso particular em que o modelo verdadeiro é uma mistura de distribuições, estimando o número de componentes através da otimização desses critérios, é demonstrado que, assintoticamente, esse número de coponentes estimado será tão grande quanto o verdadeira número de componentes (Leroux (1992)). Devido às divergências dessas tendências na prática, a proposta é adotar para as distribuições condicionais um modelo que seja uma ponderação entre os modelos selecionados pelo AIC e pelo BIC quando esses critérios indicarem dimensões diferentes.

No Capítulo 4, a abordagem proposta nesta seção será discutida em maiores detalhes. Na Seção 1.7 a seguir será descrita a organização deste trabalho.

1.7 A organização do trabalho

No Capítulo 2, a abordagem estatística para os problemas de RPS é descrita detalhadamente, onde estabelecemos a terminologia empregada neste trabalho. Definimos o *classificador de Bayes* e discutimos suas propriedades. São descritos os principais métodos estatísticos, paramétricos e não-paramétricos, empregados nos problemas de RPS, abordando suas propriedades e as questões práticas de suas utilizações. Discutimos, também, a estimação do erro de classificação e uma abordagem para redução de dimensão dos dados.

A teoria das Misturas Finitas de Densidades é discutida no Capítulo 3. Nesse capítulo, definimos o modelo e abordamos questões, teóricas e práticas, com relação a estimação dos parâmetros, enfatizando o método da Máxima Verossimilhança, e a determinação da dimensão do modelo. Nas discussões sobre a a determinação do número de componentes, são descritos o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC), abordando algumas de suas propriedades. Para a determinação das estimativas para os parâmetros, definimos o algoritmo EM (*Expectation and Maximization*) e discutimos suas principais propriedades. No contexto de misturas finitas de densidades normais, são desenvolvidos os passos desse algoritmo para determinar as estimativas dos parâmetros.

No Capítulo 4, a proposta do trabalho é apresentada de forma detalhada. Discutimos as razões empíricas que motivaram a proposta, bem como detalhamos sua implementação. Para a regra de classificação obtida com a proposta para a estimação das densidades condicionais, discutimos a propriedade de consistência para o risco de Bayes.

No Capítulo 5, estão os resultados de experimentos computacionais e uma aplicação em um problema real. Os experimentos computacionais, que são aplicações do método proposto com dados simulados, são descritos e têm seus resultados discutidos de forma detalhada. O problema real é uma aplicação em classificação de assinaturas de um indivíduo.

As conclusões e sugestões para o prosseguimento das investigações feitas neste trabalho, são apresentadas no Capítulo 6.

No Apêndice A são apresentadas as várias definições e os conceitos empregados neste trabalho.

Capítulo 2

Classificação Estatística de Padrões

Neste capítulo, descrevemos a abordagem estatística, em termos da Teoria da Decisão, para os problemas de Reconhecimento de Padrões Supervisionado (RPS). Definimos o classificador de Bayes e, considerando a aprendizagem informativa, discutimos os principais métodos paramétricos e não-paramétricos para implementá-lo. Abordamos, também, a metodologia estatística mais usual para avaliar o desempenho das regras de classificação.

2.1 O Classificador de Bayes

Como mencionamos, o objetivo principal em RPS é alocar objetos à classes previamente definidas. Cada objeto é descrito por um vetor de características e o classificador utiliza a informação contida nesse vetor para efetuar a alocação dos objetos.

Para formalizar as idéias acima no contexto da Teoria da Decisão Estatística, considere que $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \ldots, \mathcal{C}_M$ denotam as M classes definidas no problema e que o vetor de características é um vetor aleatório $\mathbf{X}$, que assume valores em $\Re^d$. Em cada $\mathcal{C}_j$, $\mathbf{X}$ é distribuído de acordo com uma função de distribuição $F_j(\cdot)$ e densidade $f_j(\cdot)$ com respeito a uma medida ν sobre $\Re^d$, adequada às variáveis preditoras, de modo que $f_j(\cdot)$ pode ser uma função de densidade de probabilidade ou uma função de probabilidade. Considere ainda que $P(\mathcal{C}_j)$ denota a probabilidade de um objeto provir da classe $\mathcal{C}_j$, para $j = 1, 2, 3, ..., M$. As densidades $f_j(\cdot)$ são denominadas *densidades condicionais das classes* e as probabilidades $P(\mathcal{C}_j)$ denominadas *probabilidades a priori das classes.*

Do ponto de vista probabílístico, os objetos são modelados como um par aleatório $(\mathbf{X}, \mathcal{C})$, onde $\mathcal{C}$ é uma variável aleatória que assume valores em $\{1, 2, 3, ..., M\}$, o conjunto dos índices que identificam as classes e, dessa forma, temos que $P(\mathcal{C}_j) = Pr(\mathcal{C} = j)$.

Sendo conhecida a densidade condicional $f_j(\cdot)$, a probabilidade $P(\mathcal{C}_j)$ e dado que observamos $\mathbf{X} = \mathbf{x}$, empregamos o teorema de Bayes para determinar a *probabilidade a posteriori* da classe $\mathcal{C}_j$, que é dada por

$$P(\mathcal{C}_j|\mathbf{x}) = Pr(\mathcal{C} = j|\mathbf{X} = \mathbf{x}) = \frac{f_j(\mathbf{x})P(\mathcal{C}_j)}{f(\mathbf{x})}, \quad j = 1, 2, 3, ..., M, \tag{2.1}$$

onde $f(\mathbf{x}) = \sum_{j=1}^{M} f_j(\mathbf{x})P(\mathcal{C}_j)$ é a *densidade marginal* de $\mathbf{X}$. A idéia principal é empregar (2.1) na construção do classificador.

Definição 2.1.1 *Um classificador (ou regra de alocação ou função de decisão) é qualquer função $r(\cdot)$ para a qual temos $r : \Re^d \longrightarrow \{1, 2, 3, ..., M\}$.*

Da definição 2.1.1, dado um classificador r e um objeto para o qual observamos $\mathbf{X} = \mathbf{x}$, $r(\mathbf{x}) = j$ significa que o objeto é alocado para a classe $\mathcal{C}_j$. Vemos, ainda, que um classificador induz uma partição $\{R_1, R_2, R_3, ..., R_M\}$ em $\Re^d$, com $R_j = \{\mathbf{x} : r(\mathbf{x}) = j\}$. Em decorrência disso, muitas vezes um classificador é definido simplesmente como uma partição de $\Re^d$ (veja, por exemplo, Breiman, Friedman, Olshen e Stone (1984)).

Pela Definição 2.1.1, temos que, para um dado problema existe uma infinidade de classificadores e surge, então, a necessidade de estabelecer critérios para avaliá-los. Poderíamos considerar na avaliação, por exemplo, as probabilidades

$$e(i,j) = Pr(r(\mathbf{X}) = j|\mathcal{C}_i) = \int_{R_j} f_i(\mathbf{x})\, d\nu, \qquad i,j = 1,2,3,...,M, \tag{2.2}$$

denominadas *taxas de alocação* do classificador r. A taxa de alocação $e(i,j)$ é a probabilidade de alocar um objeto a uma classe $\mathcal{C}_j$, dado que o objeto é da classe $\mathcal{C}_i$ e que, para cada $i \neq j$, é uma probabilidade de classificação errada pelo classificador r.

Usando as idéias da Teoria da Decisão, a maneira usual de formalizar um critério para avaliar classificadores é estabelecer uma *função de perda*. A função de perda, que denotaremos por $\lambda(\cdot,\cdot)$, dá a perda decorrente do processo de alocação. Dessa forma, $\lambda(i,j)$ é a perda decorrente de alocar um objeto da classe $\mathcal{C}_i$ para a classe $\mathcal{C}_j$. Temos que $\lambda(i,i) = 0$, para $i = 1,2,3,...,M$ e, em muitos casos, onde os custos de má classificação podem ser considerados iguais, é comum adotar $\lambda(i,j) = 1$ para $i \neq j$. Nesse último caso $\lambda(\cdot,\cdot)$ é denominada *função de perda 0-1*.

É importante observar que temos $\lambda(i, r(\mathbf{X}))$ e, então, vemos que a função de perda é uma

variável aleatória. Se vamos empregá-la para comparar classificadores, então isso deve ser feito em termos de seu valor esperado. Considerando essas observações, definimos a seguir a *função de risco* e o *risco total* para um classificador com uma dada função de perda.

Definição 2.1.2 *Para um classificador r, com uma função de perda $\lambda(\cdot,\cdot)$, temos que:*
(a) A Função de Risco é a perda esperada como função de uma classe $\mathcal{C}_i$, ou seja

$$\begin{aligned} R(r,i) &= E\{\lambda(i,r(\mathbf{X}))|\mathcal{C}_i\} \\ &= \sum_{i\neq j=1}^{M} \lambda(i,j) Pr(r(\mathbf{X})=j|\mathcal{C}_i). \end{aligned} \tag{2.3}$$

(b) O Risco Total, ou Risco de r, é a perda total esperada como função das variáveis aleatórias $\mathbf{X}$ e $\mathcal{C}$, ou seja,

$$\begin{aligned} R(r) &= E\{R(r,\mathcal{C})\} \\ &= \sum_{i=1}^{M} R(r,i) P(\mathcal{C}_i) \\ &= \sum_{i=1}^{M} \sum_{i\neq j=1}^{M} \lambda(i,j) Pr(r(\mathbf{X})=j|\mathcal{C}_i) P(\mathcal{C}_i). \end{aligned} \tag{2.4}$$

Pela Definição 2.1.2, $R(r,i)$ é a perda esperada na alocação dos objetos da classe $\mathcal{C}_i$ enquanto $R(r)$ é a perda total esperada em todo o processo de alocação empregando o classificador r.

Por exemplo, para a função de perda *0-1*, temos

$$R(r,i) = \sum_{i\neq j=1}^{M} Pr(r(\mathbf{X})=j|\mathcal{C}_i), \tag{2.5}$$

e

$$R(r) = \sum_{i=1}^{M} \sum_{i\neq j=1}^{M} Pr(r(\mathbf{X})=j|\mathcal{C}_i) P(\mathcal{C}_i). \tag{2.6}$$

De (2.5) e (2.6), vemos que $R(r,i)$ e $R(r)$ são funções das taxas de alocação dadas em (2.2). Para a função de perda *0-1*, portanto, $R(r,i)$ é a probabilidade de classificação errada dos objetos da classe $\mathcal{C}_i$ e $R(r)$ é a probabilidade total de classificação errada do classificador r. A probabilidade total de classificação errada para r também recebe a denominação de *erro de classificação* de r.

Estabelecida a função de perda, o objetivo é construir um classificador que minimize o risco total. Para esse fim, considere o seguinte classificador:

$$r^*(\mathbf{x}) = k \quad se \quad \sum_{i=1}^{M} \lambda(i,k) f_i(\mathbf{x}) P(\mathcal{C}_i) = \min_j \sum_{i=1}^{M} \lambda(i,j) f_i(\mathbf{x}) P(\mathcal{C}_i) \qquad (2.7)$$

No caso de o mínimo ocorrer para mais de uma classe, o objeto é associado a qualquer uma das classes que o atingirem.

Como as expressões para as probabilidades a posteriori das classes têm o mesmo denominador $f(\mathbf{x})$ (veja (2.1)), a regra em (2.7) pode ser estabelecida em termos dessas probabilidades, ou seja,

$$r^*(\mathbf{x}) = k \quad \text{se} \quad \sum_{i=1}^{M} \lambda(i,k) P(\mathcal{C}_i|\mathbf{x}) = \min_j \sum_{i=1}^{M} \lambda(i,j) P(\mathcal{C}_i|\mathbf{x}). \qquad (2.8)$$

Teorema 2.1.1 *Para uma dada função de perda $\lambda(\cdot,\cdot)$, o classificador r^* minimiza o risco total, ou seja, $R(r^*) \leq R(r)$ para qualquer classificador r.*

Prova: Usando a propriedade de que para duas variáveis aleatórias Y e Z e uma função integrável $\phi(Y,Z)$ temos $E[\phi(Y,Z)] = E\{E[\phi(Y,Z)|Y]\}$ (veja Ash (1972, Seção 6.5)),

podemos escrever

$$\begin{aligned} R(r) &= E\{E[\lambda(\mathcal{C}, r(\mathbf{X}))|\mathbf{X}]\} \\ &= \int_{\Re^d} E[\lambda(\mathcal{C}, r(\mathbf{x}))|\mathbf{X} = \mathbf{x}] f(\mathbf{x})\, d\nu. \end{aligned} \tag{2.9}$$

Vemos de (2.9) que, para minimizar $R(r)$, é suficiente minimizar a esperança condicional no integrando com respeito a classe $\mathcal{C}$, para cada $\mathbf{x}$. Agora, para uma classe $\mathcal{C}_j$, temos que

$$E[\lambda(\mathcal{C}, r(\mathbf{x}) = j)|\mathbf{X} = \mathbf{x}] = \sum_{i=1}^{M} \lambda(i,j) Pr(\mathcal{C} = i|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^{M} \lambda(i,j) \frac{f_i(\mathbf{x}) P(\mathcal{C}_i)}{f(\mathbf{x})}. \tag{2.10}$$

De (2.10), vemos que a esperança condicional será minimizada se tomarmos uma classe $\mathcal{C}_k$ para a qual $\sum_{i=1}^{M} \lambda(i,k) f_i(\mathbf{x}) P(\mathcal{C}_i)$ é um mínimo. Isso equivale empregar a regra r^*. □

No Teorema 2.1.1, vemos que, para o caso da função de perda *0-1*, o valor mínimo é da forma

$$\sum_{i=1}^{M} \lambda(i,k) Pr(\mathcal{C} = i|\mathbf{X} = \mathbf{x}) = \sum_{k \neq i=1}^{M} Pr(\mathcal{C} = i|\mathbf{X} = \mathbf{x}) = 1 - Pr(\mathcal{C} = k|\mathbf{X} = \mathbf{x}), \tag{2.11}$$

ou seja, temos que a esperança condicional será minimizada se tomarmos a classe $\mathcal{C}_k$ para a qual $P(\mathcal{C}_k|\mathbf{x}) = f_k(\mathbf{x}) P(\mathcal{C}_k)$ é um máximo, a classe com maior probabilidade a posteriori. Para a função de perda *0-1*, portanto, o classificador r^* pode ser expresso como

$$r^*(\mathbf{x}) = k \quad \text{se} \quad f_k(\mathbf{x}) P(\mathcal{C}_k) = \max_j f_j(\mathbf{x}) P(\mathcal{C}_j) \tag{2.12}$$

ou, equivalentemente, usando as probabilidades a posteriori $P(\mathcal{C}_j|\mathbf{x})$ como

$$r^*(\mathbf{x}) = k \quad \text{se} \quad P(\mathcal{C}_k|\mathbf{x}) = \max_j P(\mathcal{C}_j|\mathbf{x}), \tag{2.13}$$

valendo para (2.12) e (2.13) o procedimento já descrito no caso de o máximo das probabilidades a posteriori ser atingido por mais de uma classe (veja página 24).

O classificador ótimo r^* é denominado *Regra de Bayes de Mínimo Risco* ou, no caso da função de perda *0-1*, simplesmente *Regra de Bayes*. Se conhecermos $P(\mathcal{C}_j)$ e $f_j(\cdot)$, $j = 1, 2, 3, ..., M$, podemos determinar o valor do risco $R(r^*)$, denominado *Risco de Bayes*. Uma vez que a regra r^* minimiza o risco total, o valor do risco de Bayes é o menor valor que pode ser atingido por qualquer classificador e, por isso, serve como referência para comparação de classificadores. Denominaremos r^* como *Classificador de Bayes* e o risco $e^* = R(r^*)$ como *Erro de Bayes*. No caso da função de perda *0-1*, e^* é equivalente ao erro de classificação de r^*.

Uma rigorosa descrição teórica da Regra de Bayes para o caso de duas classes ($M = 2$) é apresentada em Devroye *et al.* (1996, Capítulo 2). Em Ripley (1996, Capítulo 2) é feito o desenvolvimento dessa regra considerando a classe correspondente à dúvida na alocação dos objetos.

Em problemas reais, as probabilidades a priori $P(\mathcal{C}_j)$ e as densidades condicionais das classes $f_j(\cdot)$ são desconhecidas, impossibilitando a construção do classificador de Bayes. O procedimento é, então, estimar essas quantidades com o objetivo de obter uma aproximação empírica para r^*. Como mencionado na Seção 1.3, uma abordagem consiste em estimar a regra na forma dada em (2.7), a aprendizagem informativa, ou na forma dada em (2.8), a aprendizagem discriminativa. Neste trabalho, como mencionado também, será considerada a aprendizagem informativa, ou seja, a estimação de r^* através da modelagem das densidades condicionais $f_j(\cdot)$.

Nas seções a seguir, discutiremos os procedimentos mais utilizados na estimação da regra de Bayes considerando a aprendizagem informativa. É assumida a existência de um conjunto de treinamento $\mathcal{T}_{(n)} = \{\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \mathbf{x}_{j,3}, ..., \mathbf{x}_{j,n_j}; \; j = 1, 2, 3, ..., M; \; \sum_{j=1}^{M} n_j = n\}$ (veja Seção 1.3) e os procedimentos serão discutidos considerando a função de perda *0-1*.

Na Seção 1.4, foram introduzidas as idéias básicas da aprendizagem informativa, mencionando as duas abordagens consideradas usualmente para a estimação das distribuições condicionais: a *estimação paramétrica* e a *estimação não-paramétrica*. Na estimação paramétrica, foi mencionado que o procedimento usual é empregar a distribuição normal multivariada para modelar essas distribuições. Na Seção 2.2 a seguir, a classificação empregando a distribuição normal é discutida em maiores detalhes.

2.2 Classificação com Modelos Normais

A idéia básica nesta abordagem é modelar as distribuições das classes como distribuições normais multivariadas, diferenciando-as com base nos seus parâmetros. Na classe $\mathcal{C}_j$, $j = 1, 2, 3, ..., M$, a densidade condicional $f_j(\cdot)$ é suposta ser uma normal multivariada com vetor de parâmetros $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, que denotamos por $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Temos, portanto,

$$f_j(\mathbf{x}) = f_j(\mathbf{x}; \boldsymbol{\theta}_j) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\}, \qquad (2.14)$$

onde $\boldsymbol{\mu}_j$ é o vetor de médias e $\boldsymbol{\Sigma}_j$ a matriz de covariâncias. É suposto que cada $\boldsymbol{\Sigma}_j$ é não-singular. Uma matriz de covariâncias singular implica em que as observações da classe estão num subespaço de $\Re^d$ ou, equivalentemente, satisfazem a uma ou mais restrições

lineares e, dessa forma, o problema pode sempre ser contornado por uma apropriada redução de dimensão.

Sendo conhecidas as probabilidades a priori $P(\mathcal{C}_j)$, o modelo dado em (2.14) é utilizado para determinar r^* na forma dada em (2.12). Empregando a função logarítmica, a regra torna-se: $r^*(\mathbf{x}) = k$ *se*

$$\begin{aligned} \ln\{f_k(\mathbf{x})P(\mathcal{C}_k)\} &= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \ln P(\mathcal{C}_k) \\ &= \max_j \ln\{f_j(\mathbf{x})P(\mathcal{C}_j)\}. \end{aligned} \tag{2.15}$$

Para a regra dada em (2.15), algumas simplificações podem ser obtidas considerando-se casos particulares para as matrizes de covariâncias: matrizes de covarâncias iguais e matrizes de covariâncias distintas. O modelo normal com matrizes de covariâncias iguais é denominado *modelo normal homocedástico* e, com matrizes diferentes, *modelo normal heterocedástico*.

No caso de $\mathbf{\Sigma}_j = \mathbf{\Sigma}$, $j = 1, 2, 3, ..., M$, expandindo-se a forma quadrática em (2.15) e desprezando-se os termos que são constantes para todas as classes, a regra é dada por: $r^*(\mathbf{x}) = k$ *se*

$$\begin{aligned} d_k^L(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \ln P(\mathcal{C}_k) \\ &= \max_j d_j^L(\mathbf{x}) \end{aligned} \tag{2.16}$$

O emprego da regra definida em (2.16) é denominadado *Análise Discriminante Linear* (ADL), isso em virtude de a função $d_k^L(\mathbf{x})$ ser linear em $\mathbf{x}$.

Multiplicando $d_k^L(\mathbf{x})$ por -2 em (2.16), obtemos uma forma equivalente para a regra:

selecionar a classe com o menor valor de

$$d_k^L(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) - 2 \ln P(\mathcal{C}_k). \tag{2.17}$$

O primeiro termo no segundo membro em (2.17) é, por definição, a distância de Mahalanobis ao quadrado, entre $\mathbf{x}$ e $\boldsymbol{\mu}_k$. Se as probabilidades a priori forem iguais para todas as classes, então, para um objeto com observação $\mathbf{x}$, a regra seleciona a classe cujo vetor de médias é o mais próximo de $\mathbf{x}$ em termos da distância de Mahalanobis. Vemos também que, se a matriz $\boldsymbol{\Sigma}$ for proporcional a matriz identidade, essa proximidade pode ser mensurada em termos da distância Euclidiana.

Para o caso em que as matrizes são distintas, somente o termo $-\frac{d}{2}\ln(2\pi)$ pode ser desprezado em (2.15). Para esse caso, a regra torna-se: $r^*(\mathbf{x}) = k$ *se*

$$\begin{aligned} d_k^Q(\mathbf{x}) &= -\frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + \ln P(\mathcal{C}_k) \\ &= \max_j d_j^Q(\mathbf{x}). \end{aligned} \tag{2.18}$$

Para regra formulada em (2.18), temos que $d_k^Q(\mathbf{x})$ é uma função quadrática em $\mathbf{x}$ e, por isso, o emprego desta regra é denominado *Análise Discriminante Quadrática* (ADQ).

Ao modelar as classes por distribuições normais multivariadas, estamos admitindo que as observações do vetor de características pertencem a elipsóides no espaço d-dimensional. Essas elipsóides são centradas nos vetores de médias $\boldsymbol{\mu}_j$ e suas formas são determinadas pelas matrizes $\boldsymbol{\Sigma}_j$. Além disso, as regras de alocação obtidas definem as fronteiras de decisão através de hiperplanos no caso de as matrizes $\boldsymbol{\Sigma}_j$ serem idênticas e, para o caso de serem distintas, essas fronteiras são hiperquádricas (veja Duda e Hart (1973, Seção 2.7) para várias ilustrações dessas fronteiras de decisão).

Para o emprego das regras estabelecidas nesta seção, é necessário que os parâmetros $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ e as probabilidades $P(\mathcal{C}_j)$ sejam estimados. As estimativas são obtidas empregando-se as observações no conjunto de treinamento $\mathcal{T}_{(n)}$.

Em algumas aplicações, um especialista pode dispor de informações sobre as probabilidades a priori, porém, não sendo esse o caso, elas são estimadas pelo seu estimador de Máxima Verossimilhança:

$$\widehat{P}(\mathcal{C}_j) = \frac{n_j}{n}, \tag{2.19}$$

ou seja, a proporção de observações no conjunto de treinamento que pertencem a classe $\mathcal{C}_j$.

Para estimar $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, o Método da Máxima Verossimilhança é o mais empregado na prática. O desenvolvimento dos passos necessários à determinação dos estimadores de máxima verossimilhança para os parâmetros em distribuições normais já são bastante conhecidos na literatura (veja, por exemplo, Mardia, Kent e Bibby (1979, Capítulo 4)) e o estimador $\widehat{\boldsymbol{\theta}}_j = (\widehat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}_j)$ tem as componentes dadas por

$$\widehat{\boldsymbol{\mu}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_{j,i} \tag{2.20}$$

e

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\mathbf{x}_{j,i} - \widehat{\boldsymbol{\mu}}_j)(\mathbf{x}_{j,i} - \widehat{\boldsymbol{\mu}}_j)^T, \tag{2.21}$$

onde o denominador $n_j - 1$ corrige o vício (*bias*) do estimador. De (2.20) e (2.21), vemos que $\widehat{\boldsymbol{\mu}}_j$ e $\widehat{\boldsymbol{\Sigma}}_j$ são, respectivamente, o *vetor de médias amostrais* e a *matriz de covariâncias amostrais* para a classe j, $j = 1, 2, 3, ..., M$.

Para o caso em que $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$, o estimador é dado por

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-M}\sum_{j=1}^{M}\sum_{i=1}^{n_j}(\mathbf{x}_{j,i} - \widehat{\boldsymbol{\mu}}_j)(\mathbf{x}_{j,i} - \widehat{\boldsymbol{\mu}}_j)^T, \tag{2.22}$$

onde $n - M = \sum_{j=1}^{M}(n_j - 1)$. A quantidade em (2.22) é denominada *matriz combinada de covariâncias amostrais* e temos que $\widehat{\boldsymbol{\Sigma}}$ é uma combinação dos estimadores $\widehat{\boldsymbol{\Sigma}}_j$:

$$\widehat{\boldsymbol{\Sigma}} = \frac{\sum_{j=1}^{M}(n_j - 1)\widehat{\boldsymbol{\Sigma}}_j}{\sum_{j=1}^{M}(n_j - 1)}\,. \tag{2.23}$$

Os estimadores $\widehat{P}(\mathcal{C}_j)$, $\widehat{\boldsymbol{\mu}}_j$ e $\widehat{\boldsymbol{\Sigma}}_j$ (ou $\widehat{\boldsymbol{\Sigma}}$) substituem os parâmetros desconhecidos na regra em (2.18) (ou em (2.16)) e obtemos uma versão empírica para essa regra. O emprego desses estimadores, infelizmente, não garante que a regra estimada ainda minimizará o risco de Bayes em uma aplicação particular. A questão é que, embora o estimador $\widehat{\boldsymbol{\theta}}_j$ tenha excelentes propriedades para estimar $\boldsymbol{\theta}_j$, nada nos garante que $f_j(\mathbf{x}; \widehat{\boldsymbol{\theta}}_j)$ seja uma boa estimativa para $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ e, por conseqüência, que a regra estimada, $\widehat{r^*}$, seja uma boa aproximação para r^*. Apesar dessas dificuldades, espera-se que, quando o conjunto de treinamento for grande, isto é, cada um dos n_j for suficientemente grande, a regra estimada possa apresentar um bom desempenho com relação aos erros de classificação.

Outra questão a ser observada com relação ao emprego da ADL ou ADQ na prática, diz respeito ao número de parâmetros a serem estimados. Na ADL, esse número de é $Md + d(d+1)/2$ enquanto que, na ADQ, o número aumenta para $Md + Md(d+1)/2$ e, além disso, nesse segundo caso, é necessário termos $n_j \geq d + 1$ em cada uma das classes para estimar $\boldsymbol{\Sigma}_j$. Isto pode ser bastante problemático em aplicações onde a dimensão dos dados é relativamente grande como, por exemplo, quando lidamos com imagens. Assim, como observado em Ripley (1996, Capítulo 2), embora a ADQ tenha um desempenho

melhor que a ADL para conjuntos de treinamento grandes, é possível que em conjuntos de treinamento de tamanho moderado a ADL tenha um desempenho superior.

A título de ilustração, considere o caso particular de duas classes, com as matrizes de covariâncias e as probabilidades a priori iguais. A regra estimada para esse caso é dada por: $\widehat{r^*}(\mathbf{x}) = 1$ *se*

$$d_1^L(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1 \geq \widehat{\boldsymbol{\mu}}_2^T \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_2^T \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_2 = d_2^L(\mathbf{x}). \tag{2.24}$$

A regra estabelecida em (2.24) foi obtida por Fisher (1936), empregando argumentos inteiramente diferentes dos considerados aqui. Fisher derivou a regra acima, tomando a combinação linear $\mathbf{a}^T\mathbf{x}$ que maximiza o quociente entre a soma de quadrados entre as classes e a soma de quadrados dentro das classes (Mardia *et al.* (1979, Seção 11.5)). A abordagen de Fisher não supõe distribuição normal nas classes, porém, implicitamente assume que as matrizes de dispersão são idênticas. O classificador em (2.24) é denominado *Função Discriminante Linear de Fisher*, sendo uma versão empírica do classificador (ótimo) de Bayes para duas classes, com modelo normal homocedástico e com probabilidades a priori iguais.

2.3 Classificação Não-Paramétrica

Na seção anterior, foi considerada a situação onde é assumido que as distribuições condicionais nas classes são normais multivariadas e, portanto, a questão é estimar os parâmetros dessas distribuições. Nesta seção, consideramos a estimação não-paramétrica das densidades condicionais $f_j(\cdot)$; nesse caso, não é postulado nenhum modelo específico para

essas densidades e as estimativas $\widehat{f}_j(\cdot)$ são obtidas diretamente com base no conjunto de treinamento. Serão considerados os dois principais métodos de estimação para determinar o classificador: os *Estimadores por Função-Núcleo* e os *Estimadores pelos k-Vizinhos Mais Próximos*.

2.3.1 Método dos Estimadores por Função-Núcleo

A idéia aqui é usar os estimadores de densidade por função-núcleo (*Kernel density estimators*) para estimar as densidades condicionais das classes. O estimador por função-núcleo para a densidade $f_j(\mathbf{x})$ é dado por

$$\widehat{f}_j^{(N)}(\mathbf{x}) = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{1}{h_j^d} K_0\left(\frac{\mathbf{x} - \mathbf{x}_{j,i}}{h_j}\right), \tag{2.25}$$

onde $h_j = h_j(n_j)$ é denominado *parâmetro de suavização* (*smoothing parameter*) e K_0 é a *função-núcleo* (*kernel function*). Geralmente é imposto que $K_0(\mathbf{x}) \geq 0$ e $\int K_0 \, d\mu = 1$, de forma que K_0 seja uma densidade. Também é assumido que K_0 é uma função simétrica com o "pico" em torno de $\mathbf{0}$, sendo $K_0(\frac{\mathbf{x}-\mathbf{y}}{h_j})$ usada como uma medida da proximidade de $\mathbf{x}$ e $\mathbf{y}$.

Na literatura de Reconhecimento de Padrões, o método de estimação de densidades por função-núcleo é às vezes denominado *Janelas de Parzen* (*Parzen Windows*). Esse método de estimação é discutido e tem suas propriedades estabelecidas em vária referências na literatura sobre estimação de densidades, entre os quais citamos Devroye e Györfi (1985) e Scott (1992). Entre as propriedades, com algumas condições impostas sobre K_0 e h_j, esses estimadores são assintoticamente não-tendenciosos (veja Definição A.2.1) e pontual-

mente consistentes em *Erro Quadrático Médio* (EQM). Para isso, considere as seguintes condições

$$\sup_{\mathbf{x}\in\Re^d} |K_0(\mathbf{x})| < \infty, \tag{2.26}$$

$$\|\mathbf{x}\|^d K_0(\mathbf{x}) \to 0 \quad \text{quando} \quad \|\mathbf{x}\| \to \infty, \tag{2.27}$$

$$\int |K_0(\mathbf{x})|\, d\nu < \infty \tag{2.28}$$

e, como já mencionado,

$$\int K_0(\mathbf{x})\, d\nu = 1. \tag{2.29}$$

Se as condições (2.26) a (2.29) se verificam e se $h_j \to 0$ quando $n_j \to \infty$, então quando $n_j \to \infty$,

$$E\{\widehat{f}_j^{(N)}(\mathbf{x})\} \to f_j(\mathbf{x}), \qquad \forall \mathbf{x} \in \Re^d, \tag{2.30}$$

ou seja, $\widehat{f}_j^{(N)}(\mathbf{x})$ é assintoticamente não-tendenciososo.

Considere que, além das condições (2.26) a (2.29), temos $K_0(\mathbf{x}) \geq 0$ e $K_0(\mathbf{x}) = K_0(-\mathbf{x})$ e também que $h_j \to 0$ e $n_j h_j^d \to \infty$ quando $n_j \to \infty$. Nesse caso, quando $n_j \to \infty$, temos que

$$EQM[\widehat{f}_j^{(N)}(\mathbf{x})] \stackrel{\text{def}}{=} E\{\widehat{f}_j^{(N)}(\mathbf{x}) - f_j(\mathbf{x})\}^2 \to 0, \tag{2.31}$$

nos pontos de continuidade $\mathbf{x}$ de $f_j(\cdot)$. Então, sob as condições dadas, $\widehat{f}_j^{(N)}$ converge pontualmente em erro médio quadrático (veja Duda e Hart (1973, Seção 4.3) e McLachlan (1992, Seção 9.3.3)).

Considerando a expressão para as probabilidades a posteriori dada em (2.1), com as probabilidades a priori estimadas como em (2.19) e empregando os estimadores $\widehat{f}_j^{(N)}(\cdot)$, temos o estimador para as probabilidades a posteriori

$$\widehat{P}^{(N)}(\mathcal{C}_j|\mathbf{x}) = \frac{(\frac{n_j}{n})\widehat{f}_j^{(N)}(\mathbf{x})}{\sum_{t=1}^{M}(\frac{n_t}{n})\widehat{f}_t^{(N)}(\mathbf{x})} = \frac{(\frac{n_j}{n})\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{1}{h_j^d}K_0(\frac{\mathbf{x}-\mathbf{x}_{j,i}}{h_j})}{\sum_{t=1}^{M}(\frac{n_t}{n})\frac{1}{n_t}\sum_{i=1}^{n_t}\frac{1}{h_t^d}K_0(\frac{\mathbf{x}-\mathbf{x}_{t,i}}{h_t})}. \tag{2.32}$$

Simplificando a expressão dada em (2.32), obtemos o estimador para as probabilidades a posteriori dado por

$$\widehat{P}^{(N)}(\mathcal{C}_j|\mathbf{x}) = \frac{\sum_{i=1}^{n_j}\frac{1}{h_j^d}K_0(\frac{\mathbf{x}-\mathbf{x}_{j,i}}{h_j})}{\sum_{t=1}^{M}\sum_{i=1}^{n_t}\frac{1}{h_t^d}K_0(\frac{\mathbf{x}-\mathbf{x}_{t,i}}{h_t})}, \qquad j = 1, 2, 3, ..., M. \tag{2.33}$$

Usando (2.33), desprezando o denominador (que é constante para todas as classes) obtemos a regra de classificação pelo método das funções-núcleo:

$$r_{(n)}^{(N)}(\mathbf{x}) = k \quad \text{se} \quad \sum_{i=1}^{n_k}\frac{1}{h_k^d}K_0(\frac{\mathbf{x}-\mathbf{x}_{k,i}}{h_k}) = \max_j \sum_{i=1}^{n_j}\frac{1}{h_j^d}K_0(\frac{\mathbf{x}-\mathbf{x}_{j,i}}{h_j}), \tag{2.34}$$

com as consideracões já estabelecidas para o caso de o máximo ser atingido por mais de uma classe (veja página 24). O índice (n) em $r_{(n)}^{(N)}$ visa enfatizar a dependência da regra no tamanho do conjunto de treinamento.

A regra estabelecida em (2.34) é uma aproximação para o classificador de Bayes, logo, uma questão é verificar quão boa é essa aproximação. Pela definição dos estimadores $\widehat{f}_j^{(N)}$, vemos que eles dependem de h_j, de K_0 e também de n_j, por conseqüência, $r_{(n)}^{(N)}$ também depende dessas quantidades. Uma abordagem sobre a convergência dessa regra é dada a seguir.

Considere que $\int_{\Re^d} |K_0|\, d\mu < \infty$ $(K_0 \in L_1(\Re^d))$ com $\int K_0\, d\mu = 1$ e que, para todo j, temos $h_j = h = h_{(n)}$. Seja $e_{(n)}^{(N)}$ o risco associado à regra $r_{(n)}^{(N)}$. Sob essas condições, temos o resultado a seguir.

Teorema 2.3.1 *Se* $\mathbf{X}$ *tem uma densidade,* $h_{(n)} \to 0$ *e* $nh_{(n)}^d \to \infty$*, então,* $\forall\, \epsilon \in (0,1)$ *existe* $n_0 > 0$ *tal que, para* $n \geq n_0$ *temos*

$$Pr(e_{(n)}^{(N)} - e^* > \epsilon) \leq \exp(-an\epsilon^2), \tag{2.35}$$

onde $a > 0$ *é uma constante que depende somente de* K_0*.*

Prova: Devroye e Györfi (1985, Secão 10.3). □

O Teorema 2.3.1 estabelece que o erro de classificação com a regra $r_{(n)}^{(N)}$ se aproxima assintoticamente com taxa exponencial do erro de Bayes. O teorema, entretanto, não nos dá indicações de como escolher K_0 ou $h_{(n)}$ de maneira a encontrar uma "melhor" taxa de convergência. Embora as propriedades assintóticas sejam importantes, na prática estaremos lidando com n finito e possivelmente pequeno. Assim, em vez de conhecermos como h_j converge quando n cresce, nós estaremos interessados em escolher K_0 e h_j para um tamanho particular de conjunto de treinamento.

Uma das dificuldades para o emprego do método de estimação por função-núcleo é a escolha de K_0. Considerando as restrições impostas em K_0, vemos que a função-núcleo pode ser qualquer densidade. Tem sido uma prática comum escolher K_0 como sendo a densidade normal multivariada. Sendo empregada essa densidade, o vetor de médias deve ser $\mathbf{0}$, mas ainda é fundamental a escolha da estrutura da matriz de covariâncias mais apropriada,

tendo em vista que a proximidade dos pontos será mensurada com uma métrica baseada nessa matriz. Uma alternativa para essa escolha tem sido empregar transformações no vetor de características visando eliminar as correlações entre as variáveis preditoras, adotando a função-núcleo como sendo o produto de d densidades univariadas, ou seja,

$$K_0(\mathbf{x}) = \prod_{v=1}^{d} K_1(x_v), \tag{2.36}$$

onde K_1 poderia ser, por exemplo, a densidade normal univariada. A função-núcleo da forma dada em (2.36) é denominada *núcleo produto*. Em McLachlan (1992, Seção 9.3), é apresentada uma discussão sobre a escolha da função-núcleo, onde são consideras algumas situações com respeito ao tipo de variáveis que compõem o vetor de características.

Um aspecto importante a ser considerado na escolha da função-núcleo, deveria ser a capacidade de o estimador identificar corretamente a região do espaço onde $P(\mathcal{C}_j|\mathbf{x})$ é um máximo e as regiões onde ocorrem igualdades entre essas probabilidades. Geralmente, as igualdades das probabilidades ocorrem nos extremos das distribuições sendo, portanto, fundamental que $\widehat{f}_j^{(N)}$ seja capaz de estimar corretamente esses extremos e isso depende da função núcleo empregada.

Embora a escolha da função-núcleo seja de muita importância, o seu efetivo desempenho depende da escolha do parâmetro de suavização h_j. Geralmente, a função-núcleo é fixada e o parâmetro h_j é selecionado com base nas observações no conjunto de treinamento. Se h_j é muito pequeno, $\widehat{f}_j^{(N)}$ terá um "pico" em cada ponto do conjunto de treinamento, mas, por outro lado, se ele for muito grande, ocorre uma excessiva "suavização" o que resultará em uma tendência na estimação (veja Bishop (1995, Seção 2.5)). A escolha de h_j deve, portanto, equilibrar a variância (o grau de suavização) e o viés (a tendência) de

$\widehat{f}_j^{(N)}$ (veja Seção 4.3). Na prática, a escolha de h_j é baseda em um critério a ser otimizado com as observações no conjunto de treinamento como, por exemplo, minimizar o erro médio quadrático estimado. Esses procedimentos visam determinar o melhor valor de h_j para uma determinada função núcleo e um dado número de observações no conjunto de treinamento. Para problemas de classificação, alguns critérios para seleção de h_j são apresentados em McLachlan (1992, Seção 9.4), discutindo suas qualidades e deficiências, tanto do ponto de vista teórico como prático, e sendo um exemplo desses critérios a validação cruzada mensurando a taxa de erro de classificação (veja Seção 2.4).

Fukunaga (1990, Capítulo 7) propõe uma função núcleo que depende de alguns parâmetros, cuja variação inclui as funções-núcleo normal e uniforme. O autor apresenta ainda uma ampla discussão sobre os efeitos da função núcleo, do parâmetro de suavização e do tamanho da amostra na construção do classificador.

Uma grande desvantagem do emprego da estimação por função-núcleo, é que todas as observações no conjunto de treinamento precisam ser armazenadas, pois todas são necessárias para estimar as probabilidades para as novas observações. Outro aspecto, é a necessidade de um conjunto de treinamento excessivamente grande para se obter um razoável grau de suavização das estimativas quando a dimensão de $\mathbf{x}$ for relativamente grande como, por exemplo, se tivermos $d > 5$ (veja Scott (1992, Seção 7.2)).

Uma referência específica sobre o emprego dos estimadores por função núcleo nos problemas de RPS é Hand (1982). Nessa referência, são discutidas desde a escolha da função núcleo, passando pelas questões relativas aos parâmetros de suavização, até a estimação do erro de classificação, considerando, em separado, vetores de características contínuos e aqueles com variáveis preditoras categorizadas.

2.3.2 Método dos Estimadores pelos k-Vizinhos Mais Próximos

Um método simples para estimar $f_j(\mathbf{x})$ é estabelecer uma vizinhança em torno de $\mathbf{x}$, com hipervolume denotado por $\mathcal{V}_j(\mathbf{x})$, e contar o número $\mathcal{N}_j(\mathbf{x})$ de observações do conjunto de treinamento da classe j contidos nessa vizinhança. Com esse procedimento, temos o seguinte estimador de densidade por função-núcleo

$$\widehat{f}_j(\mathbf{x}) = \frac{1}{n_j}\sum_{i=1}^{n_j}\frac{1}{\mathcal{V}_j(\mathbf{x})}I_{\{\mathcal{V}_j(\mathbf{x})\}}(\mathbf{x}_{j,i}) = \frac{\mathcal{N}_j(\mathbf{x})}{n_j\mathcal{V}_j(\mathbf{x})}, \tag{2.37}$$

onde $I_A(\cdot)$ é a função indicadora de A.

O hipervolume $\mathcal{V}_j(\mathbf{x})$ em (2.37) é constante sobre todo o espaço e isso pode trazer desvantagens na estimação. Em regiões de densidade baixa, isso poderia levar a estimativas com muitos "picos" desnecessários, enquanto que em regiões de alta densidade, poderia ocorrer uma excessiva "suavização", o que impediria a identificação das flutuações de f_j. Uma maneira de evitar essas desvantagens seria deixar $\mathcal{V}_j(\mathbf{x})$ variar sobre as observações: a idéia é fixar um valor k_{n_j} e determinar o hipervolume necessário para conter os k_{n_j} pontos mais próximos de $\mathbf{x}$, que denotamos por $\mathcal{V}_{(k_{n_j})j}(\mathbf{x})$. Esse procedimento define o *estimador pelos k vizinhos mais próximos*

$$\widehat{f}_j^{(kVP)}(\mathbf{x}) = \frac{k_{n_j}}{n_j\mathcal{V}_{(k_{n_j})j}(\mathbf{x})}\,, \tag{2.38}$$

onde, como descrito, k_{n_j} é fixado e $\mathcal{V}_{(k_{n_j})j}(\mathbf{x})$ é função de k_{n_j} e de $\mathbf{x}$. No estimador em (2.38), temos que $\mathcal{V}_{(k)j}(\mathbf{x})$ será grande em regiões de baixa densidade e pequeno nas regiões de alta densidade.

Para o estimador dado em (2.38), as condições

$$\lim_{n_j \to \infty} k_{n_j} = \infty \quad \text{e} \quad \lim_{n_j \to \infty} \frac{k_{n_j}}{n_j} = 0 \tag{2.39}$$

são necessárias e suficientes para $\widehat{f}_j^{(kVP)}(\mathbf{x})$ convergir em probabilidade para $f_j(\mathbf{x})$ em todos os pontos de continuidade de f_j. O estimador em (2.38), entretanto, não é uma densidade pois, como observado em Hand (1981, Seção 2.4), sua integral é infinita. Como o objetivo é estimar as densidades condicionais, para obter uma estimativa da regra de Bayes, é feita uma modificação na determinação das estimativas $\widehat{f}_j^{(kVP)}(\mathbf{x})$, que leva ao classificador conhecido como a *regra dos k vizinhos mais próximos*, que descrevemos a seguir.

Inicialmente, com todas as observações das M classes no conjunto de treinamento, é formado um único conjunto com n observações ($\sum_{j=1}^{M} n_j = n$). Seja $\mathcal{V}_{(k)}(\mathbf{x})$ o volume de uma hiperesfera em torno de $\mathbf{x}$ necessária para conter um número fixo k de pontos e suponha que entre os k pontos, k_j sejam da classe $\mathcal{C}_j$. É definido o estimador pelos k vizinhos mais próximos (modificado) para a densidade condicional dessa classe por

$$\widehat{f}_j^{(kVP)'}(\mathbf{x}) = \frac{k_j}{n_j \mathcal{V}_{(k)}(\mathbf{x})}\,. \tag{2.40}$$

O estimador em (2.40) é um também um estimador por função-núcleo. Para ver isso, considere que $\{\mathbf{x}_{(1)}\mathbf{x}_{(2)}\mathbf{x}_{(3)}, ..., \mathbf{x}_{(n)}\}$ são as observações do conjunto de treinamento em ordem crescente de acordo com a distância $d(\mathbf{x}_{j,i}, \mathbf{x})$. Lembrando que, a cada observação $\mathbf{x}_{(l)}$ está associada a variável indicadora de classe $y_{(l)}$ (veja Seção 1.3), podemos escrever

$$\widehat{f}_j^{(kVP)'}(\mathbf{x}) = \frac{1}{n} \sum_{l=1}^{n} \frac{w(\mathbf{x}_{(l)})}{\mathcal{V}_k(\mathbf{x})} I_{\{y_{(l)}=j\}}(\mathbf{x}_{(l)}), \tag{2.41}$$

onde $w(\mathbf{x}_{(l)}) = 1$ se $l \leq k$ e $w(\mathbf{x}_{(l)}) = 0$ se $l > k$.

Estimando as probabilidades a priori por $\widehat{P}(\mathcal{C}_j) = n_j/n$ (veja (2.19)), a densidade marginal (veja 2.1) é estimada por

$$\widehat{f}^{(kVP)'}(\mathbf{x}) = \sum_{j=1}^{M} (\frac{n_j}{n}) \frac{k_j}{n_j \mathcal{V}_{(k)}(\mathbf{x})} = \frac{k}{n\mathcal{V}_{(k)}(\mathbf{x})}\,. \tag{2.42}$$

Empregando (2.40) e (2.42), temos que as probabilidades a posteriori são estimadas por

$$\widehat{P}^{(kVP)'}(\mathcal{C}_j|\mathbf{x}) = \frac{(\frac{n_j}{n})\frac{k_j}{n_j \mathcal{V}_{(k)}(\mathbf{x})}}{\frac{k}{n\mathcal{V}_{(k)}(\mathbf{x})}} = \frac{k_j}{k}, \quad j = 1, 2, 3, ..., M. \tag{2.43}$$

Usando as probabilidades dadas em (2.43), a regra pelos k vizinhos mais próximos é definida por

$$r_{(n)}^{(kVP)}(\mathbf{x}) = t \quad \text{se} \quad k_t = \max_j \, k_j. \tag{2.44}$$

Da definição de $r_{(n)}^{(kVP)}$, vemos que, para $k = 1$, a regra aloca o objeto à classe a que pertence seu vizinho mais próximo. Para esse caso, a regra é denominada *regra do vizinho mais próximo* e a denotaremos por $r_{(n)}^{(VP)}$.

Na descrição da abordagem para definir $r_{(n)}^{(kVP)}$, vimos que é necessário estabelecer uma métrica para mensurar a distância entre $\mathbf{x}$ e cada uma das observações $\mathbf{x}_{j,i}$. Na prática, geralmente é empregada a distância Euclidiana, mas isso não é apropriado quando as variáveis preditoras são mensuradas em unidades muito distintas. Para o emprego da distância Euclidiana, portanto, primeiro deve ser feito um conveniente reescalonamento das variáveis. Se as matrizes de dispersão das classes forem aproximadamente similares, uma alternativa seria empregar a distância de Mahalanobis baseada na estimativa da

matriz combinada dessas matrizes como dado em (2.23). Para uma discussão sobre a escolha da métrica e referências sobre o assunto, indicamos Ripley (1996, Seção 6.2).

Muita pesquisa tem sido feita sobre o erro de classificação com o emprego de $r_{(n)}^{(kVP)}$. A maioria desses resultados é assintótica, ou seja, abordando o comportamento da probabilidade do erro quando $n \to \infty$. A seguir apresentaremos alguns desses resultados.

A probabilidade do erro de classificação associado ao emprego da regra $r_{(n)}^{(kVP)}$ será denotado por $e_{(n)}^{(kVP)}$. No caso de $k = 1$, escrevemos $e_{(n)}^{(VP)}$. Um dos primeiros resultados, devido a Cover e Hart (1967), é dado no teorema a seguir.

Teorema 2.3.2 *Para um problema com M classes, temos que* $\lim_{n\to\infty} E\{e_{(n)}^{(VP)}\} = e^{(VP)}$, *onde $e^{(VP)}$ satisfaz*

$$e^* \leq e^{(VP)} \leq e^*(2 - \frac{M}{M-1}e^*). \tag{2.45}$$

Prova: Hand (1981, Secão 2.4). □

Pode ser verificado que o termo à direita em (2.45) é dominado por $2e^*$; portanto, o Teorema 2.3.2 estabelece que a probabilidade do erro da regra $r_{(n)}^{(VP)}$ converge para um valor limitado superiormente pelo dobro do erro de Bayes. Desse resultado, temos que $e^{(VP)}$ é no máximo o dobro da probabilidade do erro de qualquer outro classificador quando $n \to \infty$. Também, invertendo a desigualdade à direita em (2.45), obtemos

$$e^* \geq \frac{M-1}{M} - \sqrt{\frac{M-1}{M}}\sqrt{\frac{M-1}{M} - e^{(VP)}}, \tag{2.46}$$

significando que, para um dado problema, qualquer classificador terá a probabilidade de

erro com pelo menos o valor à direita da desigualdade em (2.46)(veja Hand (1981, Secão 2.4)). Por outro lado, como $e^* \leq e^{(VP)}$, temos que valores pequenos para $e^{(VP)}$ indicam que as classes apresentam algum grau de separação no espaço do vetor de características.

No Teorema 2.3.2, o valor de k é mantido fixo quando n cresce. Quando k varia com n, explicitamente, $k = k(n) = k_n$, temos outros resultados sobre a convergência de $e_{(n)}^{(kVP)}$. Geralmente, nós não esperamos obter um classificador que atinja exatamente o erro de Bayes, mas é desejável que a probabilidade de erro do classificador se aproxime de e^* quando o conjunto de treinamento for suficientemente grande. O teorema a seguir aborda essa questão com respeito a $e_{(n)}^{(kVP)}$, considerando que k_n apresenta um determinado comportamento com relação ao aumento de n.

Teorema 2.3.3 *Se* $\mathbf{X}$ *tem uma densidade,* $k_n \to \infty$ *e* $k_n/n \to 0$*, então, para todo* $\epsilon \in (0,1)$*, existe* $n_0 > 0$ *tal que, para* $n > n_0$*,*

$$Pr(e_{(n)}^{(kVP)} - e^* > \epsilon) \leq \exp(-bn\epsilon^2), \tag{2.47}$$

onde $b > 0$ *é uma constante que depende da dimensão* d*.*

Prova: Devroye e Györfi (1985, Secão 10.5). □

O Teorema 2.3.3 estabelece a consistência da regra $r_{(n)}^{(kVP)}$ com relação ao erro de Bayes (veja Seção 4.4). Por esse teorema, com n suficientemente grande, a probabilidade da diferença entre $e_{(n)}^{(kVP)}$ e o erro de Bayes ser muito pequena será arbitrariamente próxima de 1. Teoricamente, a partir de um conjunto de treinamento com um número suficientemente grande de observações, a regra dos k vizinhos mais próximos pode "aprender" a decisão ótima.

Muitos outros resultados assintóticos para a regra $r_{(n)}^{(kVP)}$ são apresentados e rigorosamente provados em Devroye *et al.* (1996) para o caso de $M = 2$. Considerando valores de k par e ímpar, são discutidos resultados de convergência e as relações da magnitude de $e_{(n)}^{(kVP)}$ para esses dois casos de k (veja também Ripley (1996, Seção 6.2)).

Os bons resultados apresentados nos Teoremas 2.3.2 e 2.3.3 são de natureza assintótica, mas, na prática, estaremos lidando com n finito. Com um conjunto de treinamento finito, a regra $r_{(n)}^{(kVP)}$ pode ter uma taxa de erro muito maior que o dobro do erro de Bayes, principalmente, com o aumento da dimensão das observações(McLachlan (1992, Secão 9.7.4)). Em Fukunaga (1990, Secão 7.4), considerando a diferença entre o erro assintótico e o erro da regra $r_{(n)}^{(kVP)}$ com n finito, para $k = 1, 2$, são obtidas expressões que permitem analisar separadamente o efeito do tamanho de n e da dimensão d dos dados. Da análise dessas expressões, é concluído que o tamanho de n necessário para estimar o erro assintótico seria absurdamente grande quando d for relativamente alta. Outras referências sobre a discussão com respeito a n finito são dadas em McLachlan (1992, Secão 9.7.4).

Outra questão importante para o emprego da regra $r_{(n)}^{(kVP)}$ diz respeito à escolha de k. Nos resultados assintóticos, devemos ter $k_n \to \infty$ e $k_n/n \to 0$ quando $n \to \infty$. Como só podemos dispor de n finito, a escolha de k assume um papel importante na prática. Algumas propostas aparecem na literatura, uma das quais sugere que, para $M = 2$ e com aproximadamente o mesmo número de observações de treinamento dentro das classes, o valor de k deve ser próximo de $n^{2/8}$ ou $n^{3/8}$, dependendo de as matrizes de covariâncias das classes serem ou não muito diferentes (veja McLachlan (1992, Secão 9.7.5)). Na prática, na maioria das vezes, a escolha de k é feita através de validação cruzada, minimizando-se

a taxa do erro de classificação (veja Seção 2.4)

Na seção a seguir, apresentaremos alguns métodos para mensurar o desempenho de um classificador em termos dos erros de classificação. A proporção desses erros nos dá indicações sobre a probabilidade de má classificação do classificador considerado.

2.4 Estimação do Erro de Classificação

Nas seções anteriores, vimos que o objetivo é construir um classificador cujo risco seja próximo a e^*. Para o caso relativamente simples de duas classes com densidades condicionais normais homocedásticas, a taxa de erro e^* pode ser computada explicitamente (veja Ripley (1996, Seção 2.1)). Embora e^* esteja bem definido em termos das densidades condicionais, em geral, em situações mais complexas, com modelos paramétricos ou não-paramétricos, é muito difícil, às vezes impossível, obter expressões analíticas que o determinem.

Em problemas reais, é essencial obtermos uma estimativa da probabilidade de má classificação (pmc), o erro de classificação, para o classificador sendo empregado. Essa estimativa é uma medida do desempenho do classificador e serve como um critério de comparação entre possíveis classificadores para o problema. É importante observar que um estimador para a pmc é uma variável aleatória, uma vez que está baseado nas observações do conjunto de treinamento. Na prática, as distribuições das observações não são conhecidas e, por isso, é necessário empregar métodos não-paramétricos para estimar a pmc. Para um classificador r construído com um conjunto de treinamento $\mathcal{T}_{(n)}$, a pmc tem sido denotada

por $e_{(n)}^{(r)}$, mas, nas discussões a seguir, o índice (r) será omitido.

2.4.1 Método da Contagem dos Erros

Com esse método, a idéia central é estimar $e_{(n)}$ através da contagem das classificações erradas efetuadas por r em um conjunto de teste independente do conjunto de treinamento $\mathcal{T}_{(n)}$. Sendo $\{\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \mathbf{x}_{j,3}, ..., \mathbf{x}_{j,m_j};\ j = 1, 2, 3, ..., M;\ \sum_{j=1}^{M} m_j = m\}$ o conjunto de teste com m observações, o *estimador pela contagem dos erros* para $e_{(n)}$ é dado por

$$\widehat{e}_{(n,m)} = \frac{1}{m} \sum_{j=1}^{M} \sum_{i=1}^{m_j} I_{\{r(\mathbf{x}_{j,i}) \neq j\}}(\mathbf{x}_{j,i}). \tag{2.48}$$

Vemos que $\widehat{e}_{(n,m)}$ (mais formalmente $\widehat{e}_{(n,m)}^{(r)}$) é a proporção de classificações erradas das observações no conjunto de teste. Pode ser mostrado que a distribuição condicional de $m\widehat{e}_{(n,m)}$, dado $\mathcal{T}_{(n)}$, é Binomial$(m, e_{(n)})$ e, com isso, temos que

$$E\{\widehat{e}_{(n,m)}|\mathcal{T}_{(n)}\} = \frac{1}{m} E\{m\widehat{e}_{(n,m)}|\mathcal{T}_{(n)}\} = \frac{1}{m} m e_{(n)} = e_{(n)} \tag{2.49}$$

e

$$Var\{\widehat{e}_{(n,m)}|\mathcal{T}_{(n)}\} = \frac{1}{m^2} Var\{m\widehat{e}_{(n,m)}|\mathcal{T}_{(n)}\} = \frac{1}{m^2} m e_{(n)}(1 - e_{(n)}) = \frac{e_{(n)}(1 - e_{(n)})}{m}. \tag{2.50}$$

De (2.49), vemos que $\widehat{e}_{(n,m)}$ é um estimador não-tendencioso para $e_{(n)}$ e, por (2.50), que é consistente (veja Definição A.2.1). Pode ser mostrado também que, para todo $\epsilon > 0$ (veja Devroye *et al.* (1996, Seção 8.2)),

$$Pr\{|\widehat{e}_{(n,m)} - e_{(n)}| > \epsilon \,\Big|\, \mathcal{T}_{(n)}\} \leq 2\exp(-2m\epsilon^2)\ . \tag{2.51}$$

Um aspecto importante é que os resultados acima são válidos quaisquer que sejam as distribuições condicionais ou o classificador, garantindo que o estimador $\widehat{e}_{(n,m)}$ pode ser empregado em diversas circunstâncias. Agora, considerando o resultado dado em (2.51) e o fato de que

$$Var\{\widehat{e}_{(n,m)}|\mathcal{T}_{(n)}\} = \frac{e_{(n)}(1-e_{(n)})}{m} \leq \frac{1}{4m}\,, \tag{2.52}$$

vemos que a precisão das estimativas dependem diretamente do tamanho do conjunto de teste. É necessário que m seja muito grande para que a variância do estimador seja relativamente pequena ou que o intervalo de confiança seja relativamente pequeno.

Das análises acima, temos que $\widehat{e}_{(n,m)}$ apresenta muito boas qualidades como estimador de $e_{(n)}$. O problema é que o estimador necessita de um conjunto de teste muito grande e, em situações reais, entretanto, nem sempre é possível dispormos de um conjunto de teste nessas condições. Na prática, o conjunto de observações disponível é geralmente dividido em duas partes, uma para construir o classificador, o *conjunto de treino*, e a outra para teste, o *conjunto de teste*. Nessas circunstâncias, temos um dilema: tomar o conjunto de teste muito grande implica em diminuir o tamanho do conjunto de treino. A construção do classificador, no entanto, também exige um conjunto de treino de tamanho relativamente grande, para que ele seja capaz de absorver toda a variabilidade inerente ao vetor de características e, dessa forma, obtermos uma boa generalização. Como observado em Jain *et al.* (2000), não existem diretrizes satisfatórias para efetuar essa divisão do conjunto de treinamento de modo a satisfazer as exigências, a menos que o número de observações disponível seja extremamente grande.

Uma discussão sobre o tamanho do conjunto de teste necessário para estimar o erro de classificação e para comparar classificadores é apresentada em Guyon, Makhoul, Schwartz

e Vapnick (1998). Baseada na teoria estatística de estimação por intervalos de confiança, nesse artigo os autores definem o tamanho do conjunto de treinamento para se obterem estimativas do erro de classificação com um certo nível de significância. De forma análoga, os autores determinam o tamanho do conjunto de teste para comparar dois classificadores, através da estimativa de seus erros de classificação, numa estrutura de teste de hipóteses. O estudo também considera as situações onde as observações no conjunto de teste não são independentes. Numa aplicação da metodologia proposta à classificação de manuscritos, é determinado que o tamanho do conjunto de teste necessário é realmente muito grande, na ordem de 10.000 observações.

2.4.2 Estimação por Validação Cruzada

Vimos que a necessidade de um conjunto de teste independente do conjunto de treinamento é um obstáculo para o estimador pela contagem dos erros. Mesmo no caso de o conjunto de treinamento ser grande, poderia ser mais proveitoso usar todas as observações para construir o classificador com maior precisão. A motivação da estimação por validação cruzada, agora considerada, é usar o máximo possível de observações na construção do classificador.

Considere que o conjunto de treinamento é dividido em v partes. O classificador é construído empregando-se $(v-1)$ partes, sendo a parte restante usada como teste e o procedimento é repetido para cada uma das v partes. O caso de interesse é tomar $v = n$, ou seja, em cada repetição, o classificador é construído usando $(n-1)$ observações e empregado para classificar a observação restante. Temos, então, que cada uma das n observações é

empregada como teste. Nesse caso de tomar-se $v = n$, o procedimento é denominado na literatura *método "leave-one-out"*.

Em cada repetição, é determinado

$$\widehat{e}_{i(n-1,1)} = I_{\{r(\mathbf{x}_{j,i}) \neq j\}}(\mathbf{x}_{j,i}) \quad i = 1, 2, 3, \ldots, n, \tag{2.53}$$

e o estimador para o erro de classificação pelo método *"leave-one-out"* é dado por

$$\widehat{e}_{(n)}^{(CV)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{e}_{i(n-1,1)}. \tag{2.54}$$

Por sua definição em (2.53), vemos que cada $\widehat{e}_{i(n-1,1)}$ é uma varável aleatória com distribuição $Binomial(1, e_{(n-1)})$, e temos, portanto,

$$E\{\widehat{e}_{(n)}^{(CV)}\} = \frac{1}{n} \sum_{i=1}^{n} E\{\widehat{e}_{i(n-1,1)}\} = \frac{1}{n} n e_{(n-1)} = e_{(n-1)}. \tag{2.55}$$

De (2.55), vemos que $\widehat{e}_{(n)}^{(CV)}$ é um estimador não tendencioso para $e_{(n-1)}$, ou seja, para o erro de classificação do classificador construído com $n-1$ observações do conjunto de treinamento. Se, para o classificador considerado, $e_{(n)}$ convergir, então a diferença entre $e_{(n)}$ e $e_{(n-1)}$ será desprezível para n suficientemente grande.

Em Devroye *et al.* (1996, Capítulo 24), o estimador $\widehat{e}_{(n)}^{(CV)}$ é denominado de *estimador "deleted"*. Nessa referência, são apresentados resultados sobre limites superiores para $E\{(\widehat{e}_{(n)}^{(CV)} - e_{(n)})^2\}$, considerando-se alguns classificadores específicos. É observado que, uma das grandes desvantagens desse estimador, é a sua variância ser muito grande e são apresentados exemplos que ilustram esse fato (veja Seção 24.3 dessa referência).

Para a determinação de $\hat{e}_{(n)}^{(CV)}$, é necessário que o classificador seja estimado n vezes empregando-se $n-1$ observações. Isso, a princípio, representa um esforço computacional muito grande. Para alguns classificadores, no entanto, foram desenvolvidas formas que atualizam os parâmetros dos classificadores em cada repetição sem a necessidade de serem refeitos todos os cálculos. Para os classificadores baseados na distribuição normal, essas formas são apresentadas em McLachlan (1992, Seção 10.2.2) e, para os classificadores não-paramétricos, em Fukunaga (1990, Capítulo 7).

Apesar das suas desvantagens, o método *"leave-one-out"* é considerado como uma boa maneira de empregar o conjunto de treinamento, tanto para se construir o classificador como também para se estimar o erro de classificação. Mesmo no caso de um conjunto de treinamento grande, esse método ainda é mais apropriado por tomar "quase" todas as observações para construir o classificador.

2.5 Redução da Dimensionalidade

Nas seções anteriores, mencionamos que a dimensão dos dados é crucial na construção dos classificadores. Para estimar as densidades das classes, são necessários números de observações por classes extremamente grandes, se a dimensão d das observações for relativamente alta. Além disso, Scott (1992, Seção 7.1) observa que, para dados em $\Re^d$, a estrutura subjacente está quase sempre numa dimensão menor que d. Uma questão importante, portanto, é buscar a representação das observações em uma dimensão d', onde $d' < d$.

Uma abordagem para a redução de dimensão é substituir a variável original $\mathbf{X} \in \Re^d$ por uma projeção $\mathbf{Y} = T(\mathbf{X})$ em um subespaço de dimensão d', ou seja, $\mathbf{Y} \in \Re^{d'}$. Essa projeção deve ser obtida de maneira que as propriedades geométricas e estatísticas das variáveis originais sejam preservadas tanto quanto possível. Em problemas de classificação, por exemplo, as projeções deveriam ter pelo menos o mesmo poder de discriminação das variáveis originais. Como descrito no Capítulo 1, Seção 1.3, essa abordagem se enquadra nas questões de Extração de Características mas será abordada aqui simplesmente como uma questão de *Redução de Dimensão*.

A seguir, será descrito o método da *Análise de Componentes Principais* que é o mais empregado para abordar a questão da redução de dimensão.

2.5.1 Análise de Componentes Principais

O objetivo na Análise de Componentes Principais (ACP) é descrever a estrutura de variabilidade das variáveis originais através de suas combinações lineares. O método determina uma rotação dos eixos originais (dados pelas variáveis originais), para obter outros eixos que são ortogonais, de forma que as projeções dos pontos sobre esses novos eixos sejam não correlacionadas. Esses novos eixos coincidem com as direções de variação máxima no espaço das variáveis originais.

Considere um vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, ..., X_d)^T$, com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Sejam $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_d \geq 0$ os auto-valores e $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_d$ os respectivos auto-vetores de $\boldsymbol{\Sigma}$. Definimos a seguir as *componentes principais* para $\mathbf{X}$.

Definição 2.5.1 *Com a notação acima, a j-ésima componente principal para* $\mathbf{X}$ *é dada por*

$$Y_j = \mathbf{e}_j^T(\mathbf{X} - \boldsymbol{\mu}) \qquad j = 1, 2, 3, ..., d. \tag{2.56}$$

Note que, obtemos a transformação

$$\mathbf{X} \quad \longrightarrow \quad \mathbf{Y} = \mathbf{E}^T(\mathbf{X} - \boldsymbol{\mu}), \tag{2.57}$$

sendo $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_d\}$. Pela decomposição espectral, a matriz $\mathbf{E}$ satisfaz $\boldsymbol{\Sigma} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T$, onde $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_d)$. Desde que $\mathbf{E}^T\mathbf{E} = \mathbf{I}$, essa transformação é ortonormal, o que leva à preservação da distância Euclidiana pois

$$\|\mathbf{Y}\| = \mathbf{Y}^T\mathbf{Y} = \mathbf{X}^T\mathbf{E}\mathbf{E}^T\mathbf{X} = \mathbf{X}^T\mathbf{X} = \|\mathbf{X}\|. \tag{2.58}$$

A transformação para as componentes principais, dada em (2.57), é também denominada *transformação de Karhunen–Loéve*. Essa tranformação tem uma série de propriedades interessantes que são resumidas no teorema dado a seguir.

Teorema 2.5.1 *Para o vetor de componentes principais* $\mathbf{Y} = (Y_1, Y_2, Y_3, ..., Y_d)$*, temos que:*

(i) $Cov(\mathbf{Y}) = \boldsymbol{\Lambda}$;

(ii) $\sum_{j=1}^{d} Var(Y_j) = \sum_{j=1}^{d} Var(X_j) = tr(\boldsymbol{\Sigma})$;

(iii) $\prod_{j=1}^{d} Var(Y_j) = |\boldsymbol{\Sigma}|$;

(iv) *Para qualquer vetor* $\mathbf{a}_1$, *com* $\mathbf{a}_1^T\mathbf{a}_1 = 1$, $Var(\mathbf{a}_1^T\mathbf{X})$ *toma seu valor máximo* λ_1 *quando* $\mathbf{a}_1 = \mathbf{e}_1$;

(v) *Para qualquer vetor* $\mathbf{a}_j$, *com* $\mathbf{a}_j^T\mathbf{a}_j = 1$ *e tal que* $\mathbf{a}_j^T\mathbf{e}_i = 0 \quad (i = 1, 2, 3, .., j-1)$, $Var(\mathbf{a}_j^T\mathbf{X})$ *toma seu valor máximo* λ_j *quando* $\mathbf{a}_j = \mathbf{e}_j$.

Prova: Mardia *et al.* (1979, Seção 8.2), Teoremas 8.2.1, 8.2.2 e 8.2.3. □

Do Teorema 2.5.1, pelo ítem (i), as componentes principais são não-correlacionadas e que a variância da j-ésima componente principal é igual ao j-ésimo auto-valor, $j = 1, 2, 3, .., d$. Lembrando que a *Variância Total* é definida como o $tr(\mathbf{\Sigma})$ e a *Variância Generalizada* como $|\mathbf{\Sigma}|$, pelos itens (ii) e (iii), temos que essas medidas de variabilidade para as componentes principais são iguais às das variáveis originais. Os itens (iv) e (v) afirmam, respectivamente, que Y_1 é a combinação linear de $\mathbf{X}$ com máxima variância e que Y_j maximiza a variância das combinações lineares de $\mathbf{X}$ que sejam não-correlacionadas com as $j-1$ primeiras componentes principais. Dos resultados no teorema, salientamos dois aspectos importantes: as componentes principais são ortogonais entre si e preservam a variância total das variáveis originais.

Resultados adicionais podem ser estabelecidos se $\mathbf{X}$ tem distribuição normal. Suponha, por exemplo, que $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ então temos que $\mathbf{Y} \sim N(\mathbf{E}^T\boldsymbol{\mu}, \mathbf{\Lambda})$. Nesse caso, qualquer subconjunto não vazio das componentes principais tem distribuição normal e as mesmas são variáveis aleatórias independentes. Considerando isso, podem ser construídos os procedimentos clássicos de inferência, como teste de hipóteses e intervalos de confiança, para as componentes principais.

Em situações reais, os parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são desconhecidos e, portanto, é necessário estimá-los. Para isso dispomos de um conjunto de obervações (independentes) do vetor $\mathbf{X}$, $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n\}$, e os estimadores são dados por

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \text{e} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^T. \tag{2.59}$$

As componentes principais são então obtidas com base nos estimadores dados em (2.59). Sendo $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \widehat{\lambda}_3 \geq ... \geq \widehat{\lambda}_d$ os auto-valores de $\widehat{\boldsymbol{\Sigma}}$ e $\widehat{\mathbf{E}} = \{\widehat{\mathbf{e}}_1, \widehat{\mathbf{e}}_2, \widehat{\mathbf{e}}_3, ..., \widehat{\mathbf{e}}_d\}$ a matriz com os auto-vetores correspondentes, o vetor das componentes principais estimadas $\mathbf{y}_i = (y_{1i}, y_{2i}, y_{3i}, ..., y_{di})^T$ tem seus elementos determinados por

$$\mathbf{y}_i = \widehat{\mathbf{E}}^T(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}), \quad i = 1, 2, 3, ..., n. \tag{2.60}$$

Uma questão a ser levantada é a verificação de que as componentes principais estimadas são ainda a "melhor" redução da variância das observações $\mathbf{x}_i$. A resposta para essa questão é afirmativa, pode ser verificado que essencialmente não existe diferença entre as propriedades das componentes principais $\mathbf{Y}$ e as das componentes principais estimadas $\mathbf{y}_i$ (veja Seber (1984, Seção 5.2.4), Teorema 5.5 e os comentários sobre o teorema).

Pela definição das componentes principais, temos que $\mathbf{X} \in \Re^d$ é transformada em $\mathbf{Y} \in \Re^d$, ou seja, as componentes principais tem a mesma dimensão de vetor original. Para obter-se uma redução de dimensão, a idéia é selecionar um subconjunto das componentes que descrevam parte substancial da variância total das variáveis originais. Na prática, essa redução de dimensão é efetuada por tomar um $d' < d$ tal que

$$\frac{\sum_{j=1}^{d'}\widehat{\lambda}_j}{\sum_{j=1}^{d}\widehat{\lambda}_j} \geq v, \tag{2.61}$$

onde o valor de v é fixado, de maneira a manter as componentes principais estimadas que descrevem uma grande parte da variância total. Poderíamos, por exemplo, tomar $v = 0,95$ e então seriam selecionadas as componentes principais que descrevem pelo menos 95% da variância total. Usando essas idéias, temos, portanto,

$$\mathbf{y}_{i_{(d' \times 1)}} = \widehat{\mathbf{E}}^T_{(d')}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}), \quad i = 1, 2, 3, ..., n, \tag{2.62}$$

onde $\widehat{\mathbf{E}}_{(d')} = \{\widehat{\mathbf{e}}_1, \widehat{\mathbf{e}}_2, \widehat{\mathbf{e}}_3, ..., \widehat{\mathbf{e}}_{d'}\}$.

As componentes principais podem também ser empregadas como um meio para selecionar um subconjunto de variáveis diretamente no vetor de características. A idéia é basicamente descartar as variáveis que mais contribuem para as componentes principais a serem eliminadas na redução de dimensão. Para ver isso, lembrando que as componentes principais são da forma (omitindo o índice i)

$$y_j = \widehat{e}_{j1}z_1 + \widehat{e}_{j2}z_2 + \widehat{e}_{j3}z_3 + ... + \widehat{e}_{jd}z_d, \tag{2.63}$$

onde $z_l = x_l - \widehat{\mu}_l$, considere que, para um v fixo, são mantidas $d^{'}$ componentes principais. Na d-ésima componente principal, determina-se a variável que está associada com o maior coeficiente (em valor absoluto) e descarta-se essa variável. Esse procedimento é repetido com as componentes de ordem $d-1$, $d-2$,... até a componente de ordem $d^{'}+1$, sem considerar em cada repetição as variáveis que já tenham sido eliminadas previamente. Ao final do processo a dimensão do vetor de características estará reduzido para $d^{'}$.

A idéia do método para descartar variáveis no vetor de características está baseda na noção de que, se uma dada componente pode ser considerada sem importância, a variável dominante nessa componente poderia igualmente ser menos importante ou redundante. A redução de dimensão, portanto, se dá sobre o próprio vetor de características.

O emprego da ACP como um procedimento de redução de dimensão tem sido uma prática constante em problemas de Reconhecimento de Padrões. Deve ser enfatizado, no entanto, que o método não tem por objetivo realçar qualquer aspecto de separação de classes. Para os eixos dados pelas componentes principais, onde as observações originais são projetadas, não há garantias que sejam as direções que evidenciem a separação das classes (veja, por exemplo, Seber (1984, Seção 7.8)). É importante salientar, portanto, que o emprego de ACP para redução de dimensão pode obscurecer a separação das classes.

Por não considerar a informação relativa às classes das quais se originam as observações, a ACP é considerada uma técnica para redução de dimensão *não-supervisionada*.

Em muitas aplicações, as componentes principais são empregadas apenas como um mecanismo para eliminar a correlação existente entre as variáveis originais (veja, por exemplo, Cooley e MacEachern (1998)). Isto pode ser de grande utilidade no que diz respeito à estimação das matrizes de covariâncias, pois, nesse caso, seria necessário estimar matrizes diagonais (apenas d parâmetros) e, em procedimentos onde é necessária a inversão dessas matrizes, as questões computacionais seriam bastantes simplificadas.

Capítulo 3

Misturas Finitas de Densidades

Neste capítulo, definimos o modelo de mistura finita de densidades e discutimos suas propriedades. Abordamos a estimação dos parâmetros, com ênfase na estimação de Máxima Verossimilhança, e o emprego do algoritmo EM para obter as estimativas. Discutimos, também, a determinação da dimensão do modelo adequada aos dados. As misturas finitas de normais são abordadas de forma particularizada.

3.1 O Modelo de Mistura Finita de Densidades

A situação a ser considerada é a de um vetor aleatório $\mathbf{X}$, assumindo valores em $\Re^d$, cuja distribuição é representada por uma função da forma

$$p(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j f_j(\mathbf{x}), \quad \mathbf{x} \in \Re^d, \tag{3.1}$$

onde k é um inteiro, $\alpha_j \geq 0$, para $j = 1, 2, 3, \ldots, k$, $\sum_{j=1}^{k} \alpha_j = 1$ e as funções $f_j(\cdot)$ são funções densidade com respeito a uma mesma medida ν (de Lebesgue ou de contagem) sobre $\Re^d$. Em (3.1), estamos restringindo, portanto, que todas as $f_j(\cdot)$ sejam funções de probabilidade ($\mathbf{X}$ é discreto) ou todas sejam funções densidades de probabilidade ($\mathbf{X}$ é contínuo). Com essas definições, temos, portanto,

$$p(\mathbf{x}) \geq 0, \forall \mathbf{x} \quad \text{e} \quad \int_{\Re^d} p(\mathbf{x})\, d\nu = 1.$$

Para os dois casos considerados, as funções $f_j(\cdot)$ em (3.1) serão denominadas simplesmente *função densidade.*

Definição 3.1.1 *Um vetor aleatório* $\mathbf{X}$ *com função densidade* $p(\cdot)$ *da forma dada em (3.1), é dito ter uma distribuição de mistura finita de densidades. A função* $p(\cdot)$ *é denominada de mistura finita de densidades com* k *componentes, onde os parâmetros* $\alpha_1, \alpha_2, \ldots, \alpha_k$ *são denominados* proporções da mistura *e as densidades* $f_1(\cdot), f_2(\cdot), \ldots, f_k(\cdot)$ *denominadas componentes da mistura.*

Se as componentes $f_j(\cdot)$ pertencem a famílias paramétricas de distribuição, reescrevemos (3.1) como

$$p(\mathbf{x}; \mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j f_j(\mathbf{x}; \boldsymbol{\theta}_j), \quad \mathbf{x} \in \Re^d, \tag{3.2}$$

onde $\boldsymbol{\theta}_j \in \Theta_j \subseteq \Re^{m_j}$ são os parâmetros definindo cada uma das componentes $f_j(\cdot)$ e $\mathbf{\Phi}$ denota o vetor contendo todos os parâmetros distintos envolvidos na mistura de densidades, isto é,

$$\mathbf{\Phi} = (\alpha_1, \alpha_2, \ldots, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k).$$

O espaço dos parâmetros definindo a distribuição em (3.2) será denotado por Ω, ou seja,

$$\Omega = (\alpha_1, \alpha_2, \ldots, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k : \sum_{j=1}^{k} \alpha_j = 1, \alpha_j \geq 0, \boldsymbol{\theta}_j \in \Theta_j, j = 1, 2, 3, \ldots, k).$$

Na maioria das aplicações, as componentes da mistura são membros de uma mesma família paramétrica de distribuições e, para esses casos, nós denotaremos a mistura finita de densidades em (3.2) por

$$p(\mathbf{x}; \boldsymbol{\Phi}) = \sum_{j=1}^{k} \alpha_j f(\mathbf{x}; \boldsymbol{\theta}_j), \quad \mathbf{x} \in \Re^d. \tag{3.3}$$

É interessante observar que em (3.3), os parâmetros $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k$ são elementos em um mesmo espaço de parâmetros, que será denotado por Θ, ou seja $\Theta_j = \Theta$, $j = 1, 2, 3, \ldots, k$. Nesse caso, com as restrições sobre os $\alpha_{j's}$, podemos considerar que $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ define uma distribuição de probabilidade sobre Θ com

$$\alpha_j = Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_j), \quad j = 1, 2, 3, \ldots, k.$$

Com essa conjectura, se $G_{\boldsymbol{\alpha}}(\cdot)$ denota uma medida de probabilidade sobre Θ definida por $\boldsymbol{\alpha}$, então podemos estabelecer (3.3) formalmente como

$$p(\mathbf{x}; \boldsymbol{\Phi}) = \int_{\Theta} f(\mathbf{x}; \boldsymbol{\theta})\, dG_{\boldsymbol{\alpha}}(\boldsymbol{\theta}). \tag{3.4}$$

A formulação dada em (3.4) sugere uma generalização para uma forma de *mistura geral de densidades* (termo empregado em Titterington, Smith e Makov (1985)), por permitir que $G_{\boldsymbol{\alpha}}(\cdot)$ seja uma medida mais geral sobre Θ e não apenas uma medida discreta com suporte

finito como no caso que estamos considerando. Na forma dada em (3.4), a distribuição $G_{\boldsymbol{\alpha}}(\cdot)$ é denominda *distribuição de mistura*(*mixing distribution*) (Lindsay (1983)).

Na forma dada em (3.3), vemos uma justificativa para o emprego do termo *estimação semiparamétrica* ao modelar dados através de uma mistura finita de densidades. Embora as componentes sejam de famílias paramétricas (o aspecto paramétrico), os modelos de mistura não são restritos a uma forma funcional específica (o aspecto não-paramétrico). Como mencionado no Capítulo 1, Seção 1.5, com um número de componentes suficientemente grande e os parâmetros escolhidos corretamente, os modelos de mistura podem aproximar com precisão arbitrária uma ampla classe de densidades (veja, por exemplo, Priebe (1994)) e, dessa forma, podem ser empregados para modelar distribuições desconhecidas, como por exemplo, as distribuições condicionais das classes em RPS.

Outra forma de ver a questão do parágrafo acima, é considerarmos os estimadores por função-núcleo. Se, em (3.1), tomarmos

$$k = n, \quad \alpha_j = \frac{1}{n} \quad \text{e} \quad f_j(\mathbf{x}) = \frac{1}{h} K_0(\frac{\mathbf{x} - \mathbf{x}_j}{h}),$$

nós obtemos o estimador por função-núcleo

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K_0(\frac{\mathbf{x} - \mathbf{x}_j}{h}),$$

com base em um conjunto de n observações (veja Subseção 2.3.1). Dessa forma, as misturas finitas, podem ser vistas como um modelo completamente paramétrico, no caso de $k = 1$ representando uma única família paramétrica, e, no outro extremo, um modelo não-paramétrico, se $k = n$, representando o estimador por função-núcleo. Nesse sentido, um dos objetivos neste trabalho é determinar um valor para k de forma a preservar as vantagens da abordagem paramétrica e da não-paramétrica, como mencionado acima.

3.2 Distribuições Marginais

Em determinadas aplicações em RP, torna-se necessário obter alguma distribuição marginal do vetor de características. Um exemplo dessas aplicações, é na modelagem dos *pixels* de uma imagem para classificação de texturas: para um fixado tipo de vizinhança, modelamos a distribuição do *pixel* e sua vizinhança, um vetor aleatório $\mathbf{X} = (X_1, X_2, X_3, \ldots, X_d)^T$, e também a distribuição condicional $p_{X_d|X_1,X_2,X_3,\ldots,X_{d-1}}(x_d|x_1, x_2, x_3, \ldots, x_{d-1})$ do *pixel* dada a vizinhança, para a qual se faz necessário conhecermos a distribuição marginal de $(X_1, X_2, X_3, \ldots, X_{d-1})^T$ (veja, por exemplo, Popat e Picard (1997a)).

No teorema dado a seguir, vê-se que as distribuições marginais em misturas finitas de densidades com uma dada dimensão são também misturas finitas de mesma dimensão.

Teorema 3.2.1 *Seja* $\mathbf{X}$ *um vetor aleatório cuja distribuição é uma mistura finita de densidades com* k *componentes. Então, a distribuição de qualquer vetor aleatório formado com variáveis em* $\mathbf{X}$ *é também uma mistura finita de densidades com* k *componentes.*

Prova: Seja $\mathbf{X}_{(1)}$ um vetor aleatório cujas variáveis componentes formam um subconjunto das variáveis em $\mathbf{X}$. Então, a densidade de $\mathbf{X}_{(1)}$ é dada por

$$p_{\mathbf{X}_{(1)}}(\mathbf{x}_{(1)}) = \int p(\mathbf{x})\, d\nu(\mathbf{x}_{(-1)}), \tag{3.5}$$

onde a notação $d\nu(\mathbf{x}_{(-1)})$ significa tomar a integral com relação às componentes de $\mathbf{X}$ que não estão em $\mathbf{X}_{(1)}(\int(\cdot)\, d\mathbf{x}_{(-1)}$ ou $\sum_{\mathbf{x}_{(-1)}})$. Em (3.5), temos

$$p_{\mathbf{X}_{(1)}}(\mathbf{x}_{(1)}) \;\; = \;\; \int \{\sum_{j=1}^{k} \alpha_j f_j(\mathbf{x})\}\, d\nu(\mathbf{x}_{(-1)})$$

$$p_{\mathbf{X}_{(1)}}(\mathbf{x}_{(1)}) = \sum_{j=1}^{k} \alpha_j \int f_j(\mathbf{x})\, d\nu(\mathbf{x}_{(-1)}). \tag{3.6}$$

A passagem da integral da soma para a soma das integrais acima é justificada pelo Teorema da Aditividade, cada integral $\int \alpha_j f_j(\mathbf{x})\, d\nu(\mathbf{x}_{(2)})$ satisfaz as condições do teorema (veja Ash (1972, Seção 1.6)). As integrais em (3.6) dão as densidades marginais de $\mathbf{X}$ para as variáveis em $\mathbf{X}_{(1)}$ com relação a cada densidade componente $f_j(\mathbf{x})$. Temos, portanto, que

$$p_{\mathbf{X}_{(1)}}(\mathbf{x}_{(1)}) = \sum_{j=1}^{k} \alpha_j f_{(1)j}(\mathbf{x}_{(1)}), \tag{3.7}$$

onde $f_{(1)j}(\cdot)$ é a distribuição marginal para $\mathbf{X}_{(1)}$ com relação a $f_j(\cdot)$. □

Um caso particular do Teorema 3.2.1 ocorre quando as densidades componentes são $N_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Um vetor de dimensão q, formado por um subconjunto das variáveis em $\mathbf{X}$, pode ser definido tomando-se $\mathbf{X}_{(1)} = A\mathbf{X}$, onde A é uma matriz $q \times d$ cujas linhas tem 1 nas posições correspondentes as variáveis de interesse e 0 nas outras. Nesses termos, temos o corolário a seguir.

Corolário 3.2.1 *Se a distribuição de* $\mathbf{X}$ *é uma mistura de* k *densidades* $N_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2, 3, ..k$. *A distribuição de um vetor com* q *das variáveis em* $\mathbf{X}$ *é uma mistura de* k *densidades* $N_q(A\boldsymbol{\mu}_j, A\boldsymbol{\Sigma}_j A^T)$, $j = 1, 2, 3, ..., k$, *onde* A *é a matriz que determina as variáveis incluídas no vetor.*

Prova: Da teoria de probabilidade temos que as distribuições marginais para vetores com distribuição normal são também normais (veja Mardia *et al.* (1979, Teorema 3.1.1)). Além disso, se $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, temos também que $\mathbf{Y} = A\mathbf{X} \sim N_q(A\boldsymbol{\mu}, A\boldsymbol{\Sigma}A^T)$ (veja Mardia

et al. (1979, Teorema 3.2.1)). O corolário segue por empregar esses resultados no Teorema 3.2.1. □

O resultado estabelecido no teorema acima se aplica também em situações cujo interesse seja obter uma "redução" da dimensão de $\mathbf{X}$, eliminando-se variáveis redundantes na descrição dos objetos. Se o vetor de características é modelado como uma mistura finita de densidades, eliminadas as variáveis redundantes, a distribuição conjunta das variáveis restantes é ainda uma mistura finita de densidades. O corolário considera o caso particular de mistura finita de normais; para esse caso, a distribuição das variáveis mantidas para a classificação é também uma mistura finita de normais.

Na seção a seguir, será considerada a questão das condições para uma mistura finita de densidades ser *identificável*, isto é, possuir uma caracterização única a partir de seus parâmetros. Essa é uma questão crucial na estimação dos parâmetros pois, na ausência dessa condição, os procedimentos de estimação não serão exatos.

3.3 Misturas Finitas Identificáveis

A condição de que a mistura finita seja indentificável se faz necessária para que todos os parâmetros em $\mathbf{\Phi}$ possam ser estimados de forma única. O exemplo abaixo ilustra essa questão.

Exemplo 3.3.1 *Considere a família de distribuições Binomial*$(2, \theta)$, $0 < \theta < 1$, *e uma*

mistura de dois membros dessa família

$$p(x) = \alpha \binom{2}{x} \theta_1^x (1-\theta_1)^{2-x} + (1-\alpha) \binom{2}{x} \theta_2^x (1-\theta_2)^{2-x}, \qquad x \in \{0,1,2\}.$$

Então, em particular, temos

$$p(0) = \alpha(1-\theta_1)^2 + (1-\alpha)(1-\theta_2)^2$$

e

$$p(1) = 2\alpha\theta_1(1-\theta_1) + 2(1-\alpha)\theta_2(1-\theta_2).$$

Das equações acima, vemos que podem ser obtidos valores iguais para cada uma das probabilidades $p(0)$ e $p(1)$ com valores distintos de $(\alpha, \theta_1, \theta_2)$. Assim, o modelo em questão não pode ser representado unicamente em termos dos seus parâmetros.

Em geral, uma família paramétrica $\mathcal{F}$ de fdp's $f(\cdot;\mathbf{\Phi})$ é dita ser identificável se valores distintos de $\mathbf{\Phi}$ determinam membros distintos na família. Para famílias de misturas finitas de densidades, no entanto, é necessário uma definição mais específica.

Definição 3.3.1 *Seja $\mathcal{F} = \{f(\mathbf{x};\boldsymbol{\theta}) : \theta \in \Theta, \mathbf{x} \in \Re^d\}$ uma família paramétrica de densidades e*

$$\begin{aligned} \mathcal{P} \;=\; \Big\{ & p(\mathbf{x};\mathbf{\Phi}) : p(\mathbf{x};\mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j f(\mathbf{x};\boldsymbol{\theta}_j), \alpha_j \geq 0, \sum_{j=1}^{k} \alpha_j = 1, \\ & f(\mathbf{x};\boldsymbol{\theta}_j) \in \mathcal{F}, \mathbf{\Phi} = (\alpha_1, \alpha_2, \ldots, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k) \Big\} \end{aligned}$$

uma classe de misturas finitas de densidades. A classe $\mathcal{P}$ é dita identificável se, para quaisquer dois membros,

$$p(\mathbf{x};\mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j f(\mathbf{x};\boldsymbol{\theta}_j) \qquad e \qquad p(\mathbf{x};\mathbf{\Phi}') = \sum_{j=1}^{k'} \alpha_j' f(\mathbf{x};\boldsymbol{\theta}_j')$$

temos que $p(\mathbf{x};\mathbf{\Phi}) \equiv p(\mathbf{x};\mathbf{\Phi}^{'})$ *se, e somente se,* $k = k^{'}$ *e podemos permutar os índices das componentes de forma que* $\alpha_j = \alpha_j^{'}$ *e* $f(\mathbf{x};\boldsymbol{\theta}_j) = f(\mathbf{x};\boldsymbol{\theta}_j^{'}) \quad q.t.p., \quad j = 1, 2, 3, \ldots, k.$

Em Titterington *et al.* (1985), são discutidas as questões teóricas sobre as condições para uma mistura finita de densidades ser identificável e são apresentados resultados que estabelecem as condições necessárias e suficientes. Os resultados, em forma de teoremas, são dados a seguir.

Teorema 3.3.1 (Teicher (1963)). *Seja* $\mathcal{F} = \{F\}$ *uma família de funções de distribuição univariadas com transformação* $\phi(t)$ *definida para* $t \in S_\phi$ *(o domínio da definição de* ϕ*), tal que a transformação* $M : F \to \phi$ *é linear e injetiva. Suponha que exista uma ordenação* $(\preceq)$ *de* $\mathcal{F}$ *tal que* $F_1 \preceq F_2$ *implica: (i)* $S_{\phi_1} \subseteq S_{\phi_2}$*; e (ii) a existência de algum* $t_1 \in \overline{S}_{\phi_1}$ *(*t_1 *sendo independente de* ϕ_2*) tal que* $\lim_{t \to t_1} \phi_2(t)/\phi_1(t) = 0$*. Então, a classe* $\mathcal{P}$ *de todas as misturas finitas de* $\mathcal{F}$ *é identificável.*

Prova: Teicher (1963) □

O teorema acima permite aplicações diretas utilizando a função geratriz de momentos como a transformação $\phi(t)$. Empregando o Teorema 3.3.1, segue uma proposição que diz respeito às misturas de normais univariadas.

Proposição 3.3.1 *A classe de todas as misturas finitas de distribuições normais univariadas é identificável.*

O Teorema 3.3.2 a seguir permite estender para as distribuições multivariadas alguns dos resultados dados em Teicher (1963).

Teorema 3.3.2 (Yakowitz e Spragins (1968))*Considere $\mathcal{F}$ e $\mathcal{P}$ da forma dada na Definição 3.3.1.Uma condição necessária e suficiente para que $\mathcal{P}$ seja identificável é que $\mathcal{F}$ seja um conjunto linearmente independente sobre o corpo dos números reais $\Re$.*

Prova: Titterington *et al.* (1985) □

O resultado na proposição a seguir diz espeito às misturas de normais multivariadas. O resultado é decorrente do Teorema 3.3.2 e demonstrado em Yakowitz e Spragins (1968).

Proposição 3.3.2 *Uma família $\mathcal{F}$ de funções densidades normais d-dimensionais gera misturas finitas identificáveis.*

Em Titterington *et al.* (1985), além dos teoremas aqui apresentados, é feita uma revisão da literatura abordando a questão de misturas finitas identificáveis. Da revisão, os autores concluem que, a menos de casos especiais , principalmente envolvendo misturas de distribuições discretas com espaços amostrais finitos, as misturas finitas de densidades são, em geral, identificáveis

Em McLachlan e Basford (1988, Seção 1.5), são abordadas algumas das dificuldades que podem ocorrer quando as componentes $f_j(\cdot;\boldsymbol{\theta})$ pertencem a uma mesma família de distribuições. Nesse caso, $p(\cdot;\boldsymbol{\Phi})$ terá o mesmo valor se os índices j forem permutados em $\boldsymbol{\Phi}$, ou seja, embora a mistura seja identificável, temos que $\boldsymbol{\Phi}$ não o é. De fato, se todas as

$f_j(\cdot;\boldsymbol{\theta})$ pertencerem à mesma família de distribuições, então $p(\cdot;\boldsymbol{\Phi})$ será invariante para as $k!$ permutações dos índices em $\boldsymbol{\Phi}$. Duda e Hart (1973) ilustram esta questão com uma mistura de duas densidades normais com variâncias unitárias, ou seja,

$$p(x;\boldsymbol{\Phi}) = \alpha_1 \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x-\theta_1)^2] + \alpha_2 \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(x-\theta_2)^2].$$

Na mistura acima, com os valores dos parâmetros fixados, se permutarmos os índices em $\boldsymbol{\Phi} = (\alpha_1, \alpha_2, \theta_1, \theta_2)$ a densidade $p(x;\boldsymbol{\Phi})$ terá o mesmo valor em cada x.

Na prática, entretanto, o fato de $\boldsymbol{\Phi}$ não ser identificável pode ser totalmente superada por meio da imposição de uma apropriada restrição sobre $\boldsymbol{\Phi}$ (McLachlan e Basford (1988)). Como um exemplo desse tipo de abordagem, Aitkin e Rubin (1985) impõem a restrição $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$ na estimação dos parâmetros em modelos de mistura finita com componentes de uma mesma família de distribuições.

A seguir, será discutida a questão da estimação dos parâmetros em mistura finita de densidades. Na discussão, é suposto que o número de componentes no modelo é fixa e o objetivo é apenas estimar $\boldsymbol{\Phi}$.

3.4 Estimação dos Parâmetros

As duas principais abordagens para a estimação dos parâmetros em mistura finita de densidades são a *estimação de máxima verossimilhança* e a *estimação Bayesiana*. Nesta seção, serão descritas as duas abordagens, sendo a estimação de máxima verossimilhança mais detalhada por ser a abordagem empregada neste trabalho. Com relação a estimação

Bayesiana, serão apresentadas apenas as idéias básicas e indicadas referências complementares sobre o assunto.

3.4.1 Estimação de Máxima Verossimilhança

O método de estimação de Máxima Verossimilhança é a abordagem mais utilizada na estimação dos parâmetros em mistura finita de densidades (Redner e Walker (1984)). A idéia do método é a seguinte: dado um conjunto de observações independentes do modelo considerado (identicamente distribuídas), o método determina os estimadores dos parâmetros a partir da maximização da distribuição conjunta dessas observações como função dos parâmetros. A idéia, portanto, é que, sobre todo o espaço paramétrico, os valores dos parâmetros devem ser aqueles para os quais os valores observados são mais verossímeis sob o modelo considerado.

Definição 3.4.1 *Seja* $\mathbf{X}$ *um vetor aleatório com distribuição dada por* $p(\cdot;\mathbf{\Phi})$. *Dado* $\mathcal{A}_{(n)} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$ *um conjunto de* n *observações independentes de* $\mathbf{X}$, *a* Função de Verossimilhança (FV) *é definida como*

$$l(\mathbf{\Phi}) = l(\mathbf{\Phi}|\mathcal{A}_{(n)}) = \prod_{i=1}^{n} p(\mathbf{x}_i;\mathbf{\Phi}),$$

isto é, a distribuição conjunta das observações considerada como uma função de $\mathbf{\Phi}$ *e com* $\mathcal{A}_{(n)}$ *fixo.*

Considerando os objetivos deste trabalho, na definição acima, as observações em $\mathcal{A}_{(n)}$ correspondem aos exemplos de uma dada classe no conjunto de treinamento em um problema de RPS.

Para uma mistura finita de densidades $p(\cdot;\mathbf{\Phi})$, a FV, portanto, é dada por

$$l(\mathbf{\Phi}) = \prod_{i=1}^{n}\{\sum_{j=1}^{k}\alpha_j f_j(\mathbf{x}_i;\boldsymbol{\theta}_j)\}. \tag{3.8}$$

Definição 3.4.2 (Lehmann (1983))*Se existe um único valor* $\hat{\mathbf{\Phi}} \in \Omega$ *que maximiza a Função de Verossimilhança, então* $\hat{\mathbf{\Phi}}$ *é denominado* Estimador de Máxima Verossimilhança (EMV) *de* $\mathbf{\Phi}$

A Definição 3.4.2 será referida como a *definição clássica* do EMV nas discussões a seguir.

Pela Definição 3.4.2, se as densidades componentes forem diferenciáveis com respeito às componentes de $\mathbf{\Phi}$, então

$$\frac{\partial}{\partial\phi_l} l(\mathbf{\Phi}) = 0, \quad l = 1, 2, 3, ..., m, \tag{3.9}$$

onde ϕ_l denota as componentes de $\mathbf{\Phi}$, é uma condição necessária que deve ser satisfeita pela FV. Esse sistema de equações, formado pelas derivadas parciais da FV com relação às componentes de $\mathbf{\Phi}$, que recebe a denominação de *Equações de Verossimilhança*, pode ser utilizado para obter uma estimativa para $\mathbf{\Phi}$. Na prática, no entanto, é mais comum utilizar o logaritmo natural da FV, $L(\mathbf{\Phi}) = \ln(l(\mathbf{\Phi}))$, e resolver o sistema de equações

$$\frac{\partial}{\partial\phi_l} L(\mathbf{\Phi}) = \frac{\partial}{\partial\phi_l}\sum_{i=1}^{n}\ln\{\sum_{j=1}^{k}\alpha_j f_j(\mathbf{x}_i;\boldsymbol{\theta}_j)\} = 0, \quad l = 1, 2, 3, ..., m \tag{3.10}$$

uma vez que $L(\mathbf{\Phi})$, denominada *Função de log-Verossimilhança*, é uma transformação monótona de $l(\mathbf{\Phi})$.

Para muitos modelos paramétricos, os estimadores de máxima verossimilhança são relativamente simples de serem determinados (modelos para os quais a FV tem seu máximo global no interior do espaço dos parâmetros). Com modelos de mistura finita de densidades, no entanto, a determinação dos estimadores de máxima verossimilhança não é tão imediata. Dois problemas podem ocorrer na prática: (i) as equações de verossimilhança têm várias raízes; e (ii) a FV não é limitada superiormente.

O primeiro problema diz respeito a situações em que a FV atinge seu máximo local em diferentes valores de $\mathbf{\Phi}$. Na seção anterior, isso foi abordado no caso de misturas cujas densidades componentes pertencem a uma mesma família paramétrica de distribuições, onde o valor de $L(\mathbf{\Phi})$ não mudará se os pares $(\alpha_i, \boldsymbol{\theta}_i)$ e $(\alpha_j, \boldsymbol{\theta}_j)$ forem permutados em $\mathbf{\Phi}$. Como observado em Redner e Walker (1984), esse problema pode ser grave ou não, dependendo de haver um interese específico nas componentes da mistura ou apenas em uma aproximação da mistura. Nessa segunda situação, a questão não é crucial e pode ser superada na prática impondo restrições nos parâmetros, como mencionado na seção anterior. Em particular, em Ripley (1996, Seção 6.4), é observado que para os propósitos em RPS é necessário somente obter uma "boa" aproximação para a mistura finita de densidades.

Com relação ao segundo problema, existem situações onde $L(\mathbf{\Phi})$ não é limitada superiormente e, por isso, não existe o estimador de máxima verossimilhança para $\mathbf{\Phi}$. Essa situação é ilustrada no exemplo dado a seguir para uma mistura finita de densidades normais multivariadas (Hand (1981)).

Exemplo 3.4.1 *Considere a mistura de densidades*

$$p(\mathbf{x};\mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}_j|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\}.$$

Assim,

$$l(\mathbf{\Phi}) = \prod_{i=1}^{n} p(\mathbf{x}_i;\mathbf{\Phi}) = p(\mathbf{x}_1;\mathbf{\Phi}) \prod_{i=2}^{n} p(\mathbf{x}_i;\mathbf{\Phi}).$$

Fazendo-se $\prod_{i=2}^{n} p(\mathbf{x}_i;\mathbf{\Phi}) = c$, *temos*

$$l(\mathbf{\Phi}) = cp(\mathbf{x}_1;\mathbf{\Phi}) = c\sum_{j=1}^{k} \alpha_j f(\mathbf{x}_1;\boldsymbol{\theta}_j),$$

e, também,

$$l(\mathbf{\Phi}) \geq c\alpha_1 f(\mathbf{x}_1;\boldsymbol{\theta}_1) = c\alpha_1 \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}_1|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\}.$$

Agora, se tivermos que a média $\boldsymbol{\mu}_1$ *coincide com a observação* $\mathbf{x}_1$ *teremos,*

$$l(\mathbf{\Phi}) \geq c\alpha_1 \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}_1|^{\frac{1}{2}}}$$

e se , além disso, tivermos $|\mathbf{\Sigma}_1| \to 0$ *então* $l(\mathbf{\Phi}) \to \infty$.

Pela exposição dos problemas acima, para modelos de mistura finita de densidades, a equação de verossimilhança pode ter múltiplas raízes ou $L(\mathbf{\Phi})$ pode ser não limitada superiormente e, então, por sua definição, o estimador de máxima verossimilhança para $\mathbf{\Phi}$ pode não existir. Considerando essas dificuldades práticas mas, como será visto à frente, sem perder o objetivo essencial da estimação de máxima verossimilhança, Redner e Walker (1984) consideram a estimativa de máxima verossimilhança em termos de máximo local de $L(\mathbf{\Phi})$ para $\mathbf{\Phi} \in \Omega$.

Como enfatizado em Lehmann (1983), o objetivo essencial da estimação de máxima verossimilhança é a obtenção de uma estimativa $\widehat{\mathbf{\Phi}}_n = \widehat{\mathbf{\Phi}}_n(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n)$ de $\mathbf{\Phi}$ para cada n, como raiz das equações de verossimilhança, de forma a definir uma sequência de raízes $\{\widehat{\mathbf{\Phi}}_n\}$ que seja *consistente* e *assintoticamente eficiente* (veja Seção A.2). Sob algumas condições de regularidade, essa sequência de raízes existe e, com probabilidade tendendo a 1, essas raízes correspondem ao máximo local no interior de espaço dos parâmetros, mas não necessariamente ao máximo global. As condições para existência dessa sequência consistente de raízes será abordada à frente (veja, também, as discussões em McLachlan e Peel (2000, Seções 2.2 e 2.5)).

Pela exposição acima, admitir o maior máximo local de $L(\mathbf{\Phi})$ como o estimador de máxima verossimilhança não gera nenhuma incoerência se o objetivo é a determinação de estimadores de máxima verossimilhança que sejam consistentes. Dessa forma, neste trabalho procedemos como em Redner e Walker (1984), ou seja, admitimos a *"estimativa de máxima verossimilhança" para* $\mathbf{\Phi}$ *como sendo o valor* $\widehat{\mathbf{\Phi}} \in \Omega$ *para o qual a função* $L(\mathbf{\Phi})$ *atinge seu maior máximo local em* Ω.

A seguir, apresentaremos dois teoremas que resumem a existência dos resultados mencionados sobre a existência da sequência consistente de estimadores. Para a exposicão dos teoremas, é assumido que $p(\mathbf{x}; \mathbf{\Phi})$ é identificável e que $\alpha_j > 0, \quad j = 1, 2, 3, \ldots, k$. Com um pouco de abuso de notação, considere agora que $\mathbf{\Phi} = (\alpha_1, \alpha_2, \ldots, \alpha_{k-1}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$ é a coleção de todos os parâmetros com o redundante α_k omitido ($\alpha_k = 1 - \sum_{j=1}^{k-1} \alpha_j$) e que Ω é o espaço dos parâmetros modificado, ou seja

$$\Omega = \Big\{\alpha_1, \alpha_2, \ldots, \alpha_{k-1}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k : \sum_{j=1}^{k-1} \alpha_j < 1, \alpha_j > 0, \boldsymbol{\theta}_j \in \Theta_j, j = 1, 2, 3, \ldots, k\Big\}.$$

Estabelecemos inicialmente as condições de regularidade para o teorema. Ao estabelecer as condições e nas exposições a seguir, denotaremos por $\mathbf{\Phi}^*$ o verdadeiro valor do parâmetro $\mathbf{\Phi}$ e sugerimos Serfling (1980, Seções 4.1 e 4.2) para uma discussão sobre as condições. Veja Apêndice A para as definições necessárias.

Condição 1: Para todo $\mathbf{\Phi} \in \Omega$, para quase todo $\mathbf{x} \in \Re^d$ e para $i, j, k = 1, 2, 3, .., m$, as derivadas parciais de primeira, segunda e terceira ordem de $p(\mathbf{x}; \mathbf{\Phi})$ com relação às componentes de $\mathbf{\Phi}$ existem e satisfazem

$$\left|\frac{\partial p(\mathbf{x};\mathbf{\Phi})}{\partial \xi_i}\right| \leq f_i(\mathbf{x}), \quad \left|\frac{\partial p(\mathbf{x};\mathbf{\Phi})}{\partial \xi_i \xi_j}\right| \leq f_{ij}(\mathbf{x}), \quad \left|\frac{\partial p(\mathbf{x};\mathbf{\Phi})}{\partial \xi_i \xi_j \xi_k}\right| \leq f_{ijk}(\mathbf{x}),$$

onde $\xi_{i's}$ representam as componentes de $\mathbf{\Phi}$, as funções $f_i(\mathbf{x})$ e $f_{ij}(\mathbf{x})$ são integráveis e $f_{ijk}(\mathbf{x})$ satisfaz a $\int_{\Re^d} f_{ijk}(\mathbf{x})\, d\mu < \infty$.

Condição 2: A matriz de informação de Fisher em $\mathbf{\Phi}^*$, $I(\mathbf{\Phi}^*)$, é finita e definida positiva.

Teorema 3.4.1 *Considere que as condições 1 e 2 são satisfeitas e que seja dada uma vizinhança qualquer, suficientemente pequena, de* $\mathbf{\Phi}^*$ *em* Ω. *Então, com probabilidade* 1, *para* $n \to \infty$, $\exists\ \widehat{\mathbf{\Phi}}_n = \widehat{\mathbf{\Phi}}_n(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n)$ *único, solução das equações de verossimilhança nessa vizinhança, que maximiza localmente a função de log-verossimilhança* $L(\mathbf{\Phi})$. *Além disso,* $\sqrt{n}(\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}^*) \xrightarrow{D} N(\mathbf{0}, [I(\mathbf{\Phi}^*)]^{-1})$.

Prova: Peters e Walker (1978); Lehmann (1983). □

O Teorema 3.4.1 estabelece que, se as condições de regularidades são satisfeitas, existe uma única solução das equações de verossimilhança que é consistente, maximizando (pelo

menos) localmente a função de log-verossimilhança, e assintóticamente eficiente como estimador do verdadeiro valor do parâmetro (sobre as condições, veja Serfling (1980, Seções 4.1 e 4.2)). Esse teorema não nos diz, entretanto, se $\widehat{\mathbf{\Phi}}_n$ é realmente uma estimativa de máxima verossimilhança, ou seja, um ponto onde a função de verossimilhança atinge seu maior máximo local. Outra questão não respondida é que, se $\widehat{\mathbf{\Phi}}_n$ é um EMV, existem outras estimativas de máxima verossimilhança diferentes de $\widehat{\mathbf{\Phi}}_n$? O próximo teorema, embora de forma restrita, dá uma resposta a essas questões.

O teorema a seguir é uma versão mais fraca do resultado apresentado em Redner (1981), onde o autor considerou também famílias de distribuições não-identificáveis. Primeiramente, é necessário definir duas funções para estabelecer as condições do teorema: sendo $N_r(\mathbf{\Phi})$ uma bola fechada de raio $r > 0$ em torno de $\mathbf{\Phi}$, são definidas as seguintes funções

$$p(\mathbf{x}|\mathbf{\Phi}, r) = \sup_{\mathbf{\Phi}' \in N_r(\mathbf{\Phi})} p(\mathbf{x}|\mathbf{\Phi}') \quad \text{e} \quad p^*(\mathbf{x}|\mathbf{\Phi}, r) = \max\{1, p(\mathbf{x}|\mathbf{\Phi}, r)\}.$$

Condição 3: Para cada $\mathbf{\Phi} \in \Omega$ e r suficientemente pequeno

$$\int_{\Re^d} \ln p^*(\mathbf{x}|\mathbf{\Phi}, r) p(\mathbf{x}|\mathbf{\Phi}^*)\, d\mu < \infty.$$

Condição 4:

$$\int_{\Re^d} \ln p(\mathbf{x}|\mathbf{\Phi}^*) p(\mathbf{x}|\mathbf{\Phi}^*)\, d\mu < \infty.$$

Teorema 3.4.2 *Seja Ω_o um subconjunto compacto de Ω contendo $\mathbf{\Phi}^*$ no seu interior e considere o conjunto*

$$C = \{\mathbf{\Phi} \in \Omega_o : p(\mathbf{x}|\mathbf{\Phi}) = p(\mathbf{x}|\mathbf{\Phi}^*) \quad q.t.p.\}.$$

Se as condições 1,2,3 e 4 estão satisfeitas e D é qualquer subconjunto fechado de Ω_o sem interseção com C, então

$$Pr\left\{\lim_{n\to\infty}\sup_{\mathbf{\Phi}\in D}\frac{\prod_{i=1}^{n}p(\mathbf{x}_i|\mathbf{\Phi})}{\prod_{i=1}^{n}p(\mathbf{x}_i|\mathbf{\Phi}^*)}=0\right\}=1.$$

Prova: Redner (1981); Redner e Walker (1984). □

O Teorema 3.4.2 implica que, se Ω_o é qualquer subconjunto compacto de Ω contendo $\mathbf{\Phi}^*$ no seu interior então, com probabilidade 1, $\widehat{\mathbf{\Phi}}_n$ é o EMV em Ω_o para n suficientemente grande (veja também Redner (1981, Teoremas 2 e 4)) . Dado que a mistura é identificável, qualquer outra estimativa em Ω_o é obtida a partir de $\widehat{\mathbf{\Phi}}_n$ pela permutação dos índices e, por isso, levando à mesma densidade limitante $p(\mathbf{x}|\mathbf{\Phi}^*)$.

Do ponto de vista teórico, os Teoremas 3.4.1 e 3.4.2 são adequados para os problemas de estimação em mistura finita de densidades, ao assegurar a existência do estimador de máxima verossimilhança fortemente consistente e caracterizar seu comportamento assintótico como solução das equações de verossimilhança. O teorema, entretanto, não estabelece como lidar com as situações onde a função de verossimilhança tem vários máximos locais, por exemplo. No caso de várias raízes, não fica estabelecido qual delas tomar para obter uma sequência consistente. Lehmann (1999) sugere que teríamos uma solução se fosse conhecido um estimador consistente de $\mathbf{\Phi}$, $\widetilde{\mathbf{\Phi}}_n$, e, consequentemente, a raiz $\widehat{\mathbf{\Phi}}_n$ das equações de versossimilhança mais próxima de $\widetilde{\mathbf{\Phi}}_n$ também seria consistente. O obstáculo dessa abordagem é a necessidade de serem determinadas todas as raízes, o que pode ser uma tarefa difícil. Salientamos também, que os resultados nos teoremas dependem das condições estabelecidas. Com respeito a isso, em McLachlan e Basford (1988, Seção 1.8), os autores comentam que a forma dessas condições sugere que elas são válidas para

muitas famílias de distribuições. Para mais discussões sobres essas questões e referências adicionais sobre o assunto sugerimos, Lehmann (1983) e Lehmann (1999).

Na prática, por outro lado, a determinação dos estimadores de máxima verossimilhança para mistura finita de densidades envolvem várias dificuldades devido à complexidade da dependência da função de verossimilhança nos parâmetros. Geralmente, as equações não são lineares e não é possível obter soluções analíticas e, por isso, a saída é recorrer a procedimentos iterativos na busca das soluções. Em Redner e Walker (1984), são discutidos vários aspectos práticos da determinação de EMV consistentes, a relação da matriz de informação de Fisher com as questões numéricas envolvidas, bem como relatam resultados de estimação obtidos com outros métodos em mistura finita de densidades. Esses autores concluem que, apesar das dificuldades envolvidas, o método da máxima verossimilhança tem sido muito bem sucedido quando comparado a outros métodos de estimação.

3.4.2 Estimação Bayesiana

A idéia básica na inferência Bayesiana é que, dado um modelo com um parâmetro desconhecido $\mathbf{\Phi} \in \Omega$, antes de tomarmos uma amostra de observações do modelo, introduzimos no processo de estimação uma conjectura sobre $\mathbf{\Phi}$. A conjectura é expressa por meio de uma distribuição de probabilidade $p(\mathbf{\Phi})$ para $\mathbf{\Phi}$. Essa conjectura é então avaliada num segundo estágio, após ter sido observada uma amostra $\mathcal{A}_{(n)}$ do modelo, através da distribuição condicional $p(\mathbf{\Phi}|\mathcal{A}_{(n)})$. A distribuição $p(\mathbf{\Phi})$ é denominada distribuição *a priori* e $p(\mathbf{\Phi}|\mathcal{A}_{(n)})$ distribuição *a posteriori*.

Sendo $l(\mathbf{\Phi})$ a função de verossimilhança para $\mathbf{\Phi}$, as idéias acima são formalizadas com o

emprego do teorema de Bayes

$$p(\mathbf{\Phi}|\mathcal{A}_{(n)}) = \frac{l(\mathbf{\Phi})p(\mathbf{\Phi})}{\int_{\Omega} l(\mathbf{\Phi})p(\mathbf{\Phi})\, d\mathbf{\Phi}}. \tag{3.11}$$

Das idéias descritas acima, temos a questão de como escolher a distribuição a priori para $p(\mathbf{\Phi})$, distribuições essas que, em geral, também dependem de outros parâmetros, que são denominados *hiperparâmetros*. Para a escolha de $p(\mathbf{\Phi})$, geralmente é considerado o conceito de *famílias conjugadas*, propriedade de que, para algumas famílias de distribuição, tomando a priori pertencente à família, está assegurado que a posteriori também pertence à mesma família (veja Berger (1985, Capítulo 4)). O emprego de famílias conjugadas simplifica bastante o processo de inferência.

As inferências sobre $\mathbf{\Phi}$ são obtidas com o emprego de (3.11) que, em geral, tem forma bastante complicada. Mesmo trabalhando com famílias conjugadas, quando $l(\mathbf{\Phi})$ é proveniente de uma mistura finita de densidades, não é possível um tratamento analítico de (3.11) (Titterington *et al.* (1985)). Várias abordagens têm sido propostas para a resolução desses problemas e, atualmente, os métodos de simulação têm sido os mais utilizados na prática. Os métodos de simulação empregados são denominados *métodos de Monte Carlo via Cadeias de Markov*(MCMC)(*Markov Chain Monte Carlo*).

Uma referência básica para os métodos MCMC aplicados para estimação em misturas finitas de densidades é Gilks, Richardson e Spiegelhalter (1996, Capítulo 24). Nessa abordagem, a idéia básica é estabelecer as distribuições a priori e a posteriori convenientes e obter, através de dados simulados por meio dessas distribuições, as estimativas para os parâmetros da mistura (veja, também, Gamerman (1996)).

Como referência, citamos também Lavine e West (1992), onde os autores consideram uma mistura finita de normais para modelar problemas de RPS e RPNS, estabelecem as prioris convenientes e desenvolvem os passos necessários a estimação dos parâmetros. No contexto de estimação de densidades através de mistura finita de densidades, empregando a abordagem Bayesiana para a estimação dos parâmetros, citamos Roeder e Wasserman (1997), onde os autores apresentam um método que gera uma distribuição a posteriori para os parâmetros consistente e citam várias referências sobre o assunto.

Na seção a seguir, será considerada a questão da escolha da dimensão da mistura. A escolha do número de componentes é de fundamental importância para o emprego adequado dos modelos de misturas finitas.

3.5 A Estimação do Número de Componentes

Nas seções anteriores, os modelos de misturas finitas foram abordadas considerando a dimensão fixa. Nesta seção, abordaremos a questão de como estimar a dimensão k para o modelo. Essa é uma das questões mais críticas para o emprego desses modelos e tem resistido a uma abordagem estatística completamente satisfatória (Roeder (1990), Polymenis e Titterington (1998), McLachlan e Peel (2000, Seção 6.1)). Em geral, duas principais abordagens estatísticas têm sido adotadas para esse problema, o emprego de *Testes de Hipóteses* e a abordagem denominada *Seleção de Modelos*, que consiste em selecionar o modelo por otimizar uma dada função-critério.

3.5.1 Teste de Hipóteses para determinar k

A abordagem por Teste de Hipóteses consiste em formular hipóteses estatísticas relativas ao número de componentes para o modelo e efetuar testes sobre essas hipóteses, para determinar o menor valor de k compatível com as observações. Mais precisamente, para $k_0 < k_1$, são considerados as hipóteses estatísticas

$$H_o : k = k_0 \quad \text{e} \quad H_a : k = k_1,$$

que correspondem, respectivamente, a um modelo de dimensão k_0 e outro de dimensão k_1. Para testar as hipóteses acima, o procedimento óbvio seria empregar o *teste da razão das verossimilhanças generalizado*. A estatística de teste é da forma

$$\Lambda_n = \frac{L(\widehat{\mathbf{\Phi}}_{(0)})}{L(\widehat{\mathbf{\Phi}}_{(1)})},$$

onde $L(\widehat{\mathbf{\Phi}}_{(c)})$ é o máximo da FV para o modelo de dimensão k_c, $c = 0, 1$. É conhecido da teoria estatística que, sob algumas condições de regularidade, a variável $\lambda_n = -2ln\Lambda_n$ tem distribuição assintótica *Qui-quadrado* (χ^2) com graus de liberdade dados pela diferença entre o número de parâmetros nas duas hipóteses (Lehmann (1986)). Para misturas finitas de densidades, no entanto, as condições de regularidades não se verificam (McLachlan e Peel (2000, Seção 6.4)). Para contornar esse problema, várias alternativas têm sido propostas, muitas delas sugerem ajustes para os graus de liberdade numa tentativa de aproximar a distribuição de λ_n. Para uma discussão sobre algumas dessas propostas e outras referências, sugerimos McLachlan e Basford (1988, Seção 1.10) e McLachlan e Peel (2000, Capítulo 6). Em Soromenho (1994) é apresentado um estudo comparativo entre algumas dessas propostas.

Outra abordagem é utilizar o método de *bootstrap*. Nesse caso, a idéia é encontrar uma

aproximação da distribuição nula de λ_n, empregando-se a reamostragem dos dados, ou seja, gerando *amostras de bootstrap* e determinando em cada uma delas o valor da estatística λ_n. Com esses valores determinados, é construída uma distribuição empírica para λ_n que é empregada para testar as hipóteses. Sobre essa abordagem, veja McLachlan e Peel (1997), Polymenis e Titterington (1998) e McLachlan e Peel (2000, Seção 6.6).

3.5.2 Seleção de Modelos

A idéia dos métodos nesta abordagem, é estabelecer uma função-critério cuja otimização indique o verdadeiro número de componentes da modelo. Em geral, são da forma

$$\widehat{k} = arg\, min_k\{C(\widehat{\mathbf{\Phi}}_{(k)}), k = k_{min}, ..., k_{max}\},$$

onde $C(\widehat{\mathbf{\Phi}}_{(k)})$ é o valor da função-critério para modelo estimado com dimensão k. O valor $-L(\widehat{\mathbf{\Phi}}_{(k)})$ poderia ser visto como um critério, porém, como a FV é uma função não decrescente em k, esse critério escolheria sempre a maior dimensão possível e, dessa forma, não seria uma formalização da noção intuitiva de modelo mais "adequado". Dois critérios de seleção de modelos bastantes conhecidos na literatura, o *Critério de Informação de Akaike* e o *Critério de Informação Bayesiano*, podem ser empregados para obtermos estimativas de k.

O Critério de Informação de Akaike

A idéia subjacente ao Critério de Informação de Akaike (AIC) (de *Akaike's Information Criterion*), é estimar a informação de *Kullback-Leibler* do verdadeiro modelo com respeito

ao modelo estimado. Sendo $p(\mathbf{x}|\mathbf{\Phi}^*)$ a modelo verdadeiro e $p(\mathbf{x}|\widehat{\mathbf{\Phi}})$ o modelo estimado, a informação de Kullback-Leibler é dada por

$$\begin{aligned} I_{KL}(\mathbf{\Phi}^*, \widehat{\mathbf{\Phi}}) &= \int p(\mathbf{x}|\mathbf{\Phi}^*) \ln(\frac{p(\mathbf{x}|\mathbf{\Phi}^*)}{p(\mathbf{x}|\widehat{\mathbf{\Phi}})})\, d\mathbf{x} \\ &= \int p(\mathbf{x}|\mathbf{\Phi}^*) \ln p(\mathbf{x}|\mathbf{\Phi}^*)\, d\mathbf{x} - \int p(\mathbf{x}|\mathbf{\Phi}^*) \ln p(\mathbf{x}|\widehat{\mathbf{\Phi}})\, d\mathbf{x}, \end{aligned}$$

que é uma medida de divergência entre o modelo verdadeiro e o modelo estimado, sendo o objetivo, portanto, minimizar essa divergência. Na expressão acima, vemos que somente o segundo termo a direita da igualdade, a *log-verossimilhança esperada*,

$$\eta(\mathcal{A}_{(n)}, \mathbf{\Phi}^*) = \int p(\mathbf{x}|\mathbf{\Phi}^*) \ln p(\mathbf{x}|\widehat{\mathbf{\Phi}})\, d\mathbf{x},$$

é relevante na minimização de $I_{KL}(\mathbf{\Phi}^*, \widehat{\mathbf{\Phi}})$.

Na derivação do AIC, $\eta(\mathcal{A}_{(n)}, \mathbf{\Phi}^*)$ é estimada por

$$\widehat{\eta}(\mathcal{A}_{(n)}, \widehat{\mathbf{\Phi}}) = \frac{1}{n} \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\widehat{\mathbf{\Phi}}),$$

denominada *log-verossimilhança média*, que é um estimador consistente para essa quantidade. Assumindo as condições de regularidade, que asseguram a distribuição assintótica qui-quadrado para o teste da razão de verossimilhança generalizada, o seguinte forma para o critério é obtida

$$AIC(k) = -2L(\widehat{\mathbf{\Phi}}_{(k)}) + 2k.$$

Todo o desenvolvimento para a determinação do AIC é apresentado em Bozdogan (1987). No contexto de misturas finitas de densidades, a forma empregada é (veja McLachlan e Peel (2000, Seção 6.8))

$$AIC(k) = -2L(\widehat{\mathbf{\Phi}}_{(k)}) + 2v(k)$$

onde $v(k)$ é o número de parâmetros livres, no modelo com dimensão k. O termo $v(k)$ pode ser visto como um fator de "penalização" para o crescimento do número de componentes k no modelo.

Algumas considerações têm sido feitas sobre o AIC. Do ponto de vista teórico, o problema com esse critério é que, na sua formulação, é assumido que o modelo verdadeiro e o modelo conjecturado pertencem a uma mesma família paramétrica de distribuições e, além disso, assume as condições de regularidade da teoria assintótica de $\lambda_n = -2ln\Lambda_n$. Como mencionado, essas condições não se verificam no contexto de misturas finitas.

Na prática, tem sido observado que o AIC não é *consistente em ordem* (um critério é consistente em ordem se, quando $n \to \infty$, com probabilidade tendendo a 1 o critério é minimizado na dimensão verdadeira do modelo) e tende a superestimar a dimensão do modelo (Celeux e Soromenho (1996)). No contexto de mistura, isso significa que o AIC apresenta uma tendência a selecionar um modelo com um número de componentes superior ao do modelo verdadeiro.

O Critério de Informação de Bayesiano

O Critério de Informação Bayesiano (BIC) (de *Bayesian Information Criterion*) está baseado na teoria Bayesiana de seleção de modelos. Nessa teoria, são considerados vários possíveis modelos, com suas probabilidades a priori, e o objetivo é selecionar o modelo com a maior probabilidade a posteriori dadas as observações $\mathcal{A}_{(n)}$. Sendo $M_1, M_2, M_3, ..., M_k$ os modelos considerados e $p(M_k)$, $k = 1, 2, 3, ..., K$, as respectivas probabilidades a priori,

pelo Teorema de Bayes, a posteriori de M_k dado $\mathcal{A}_{(n)}$ é

$$p(M_k|\mathcal{A}_{(n)}) = \frac{p(\mathcal{A}_{(n)}|M_k)p(M_k)}{\sum_{t=1}^{K} p(\mathcal{A}_{(n)}|M_t)p(M_t)}.$$

Da expressão acima, vemos que, para a posteriori de M_k, é necessário determinar $p(\mathcal{A}_{(n)}|M_k)$. Quando existem parâmetros desconhecidos nos modelos, essa distribuição é obtida por integração sobre o espaço dos parâmetros, ou seja,

$$p(\mathcal{A}_{(n)}|M_k) = \int p(\mathcal{A}_{(n)}|\mathbf{\Phi}_{(k)}, M_k)p(\mathbf{\Phi}_{(k)}|M_k)\, d\mathbf{\Phi}_{(k)},$$

onde $p(\mathbf{\Phi}_{(k)}|M_k)$ é a distribuição a priori para $\mathbf{\Phi}_{(k)}$ (veja Kass e Raftery (1995)). É interessante observar que $p(\mathcal{A}_{(n)}|\mathbf{\Phi}_{(k)}, M_k)$ é a função de verossimilhança para o modelo M_k, com vetor de parâmeros $\mathbf{\Phi}_{(k)}$.

A quantidade $p(\mathcal{A}_{(n)}|M_k)$ recebe a denominação de *verossimilhança integrada*. Se as probabilidades a priori $p(M_k)$ forem todas iguais, o procedimento seleciona o modelo com a maior verossimilhança integrada.

A maior dificuldade com essa abordagem Bayesiana é avaliar a integral que define a verossimilhança integrada. A principal abordagem para aproximar essa integral é através da minimização do Critério de Informação Bayesiano, que é dado por,

$$BIC(k) = -2L(\widehat{\mathbf{\Phi}}_{(k)}) + v(k)\ln n,$$

onde, como antes, $v(k)$ é o número de parâmetros livres, no modelo de dimensão k. A aproximação dada pelo BIC foi desenvolvida em Schwarz (1978) e, por isso, na literatura também recebe a denominação de *critério de Schwarz*.

O BIC é considerado *consistente em ordem*, o que implica que assintoticamente tende a selecionar o modelo de dimensão correta (Celeux e Soromenho (1996)). Esse critério foi

desenvolvido sob condições de regularidade que não se verificam para modelos de mistura finita. Seu emprego na prática, entretanto, tem demonstrado resultados considerados, no mínimo, razoáveis. Em particular, esses bons resultados têm se verificado empregando-se esse critério para selecionar o número de componentes para uma mistura finita em estimação de densidades (Biernacki, Celeux e Govaert (2000)).

Para o emprego com mistura finitas, a forma do BIC sugere que esse critério tende a favorecer os modelos mais simples, com um menor número de componentes, que o AIC.

Na seção a seguir, descrevemos o algoritmo EM e a sua aplicação na determinação da solução das equações de verossimilhança.

3.6 Estimação de Máxima Verossimilhança via o algoritmo EM

O algoritmo EM, E de "Expectation" e M de "Maximization", é um método iterativo empregado para maximizar a função de verossimilhança. Nesta seção, descrevemos o algoritmo EM e discutimos suas principais propriedadades. No contexto de mistura finita de densidades, desenvolvemos os passos necessários à determinação das raízes das equações de verossimilhança. É suposto que as densidades envolvidas são diferenciáveis com relação às componentes dos seus vetores de parâmetros.

3.6.1 O Algoritmo EM

A abordagem que adotadamos na descrição do algoritmo EM, é conhecida na literatura como *estimação a partir de dados incompletos*. Essa denominação é devida ao trabalho de Dempester, Laird e Rubin (1977), que é a referência básica para o emprego desse algoritmo em misturas finitas de densidades.

A idéia subjacente em estimação com dados incompletos, é considerar os dados observados ($\mathbf{x}_i$) como *dados incompletos* e aumentá-los com a inclusão de variáveis latentes ($\mathbf{z}_i$), variáveis não observáveis diretamente, de modo que a distribuição dos *dados completos* ($\mathbf{y}_i = (\mathbf{x}_i, \mathbf{z}_i)$) simplifique as as análises a serem desenvolvidas. É observado em Wu (1983) que, em muitos problemas estatísticos de estimação, a maximização do problema em termos de dados completos é mais simples do que especificando-o como dados incompletos.

Para formalizar as idéias acima, considere que temos dois espaços amostrais $\mathcal{Y}$ e $\mathcal{X}$. Os dados incompletos, $\mathbf{x} \in \mathcal{X}$, são observados e consideramos uma variável latente $\mathbf{z}$ para formar os dados completos $\mathbf{y} = (\mathbf{x}, \mathbf{z}) \in \mathcal{Y}$. Seja $g(\mathbf{x}; \mathbf{\Phi})$ a densidade definida sobre $\mathcal{X}$, parametrizada por $\mathbf{\Phi} \in \Omega$, e suponha que existe uma densidade (conjunta) $f(\mathbf{y}; \mathbf{\Phi})$ sobre $\mathcal{Y}$. Dado $\mathbf{x} \in \mathcal{X}$, o objetivo é estimar $\mathbf{\Phi}$ de forma a maximizar a função de log-verossimilhança $L(\mathbf{\Phi}) = \ln g(\mathbf{x}; \mathbf{\Phi})$, para $\mathbf{\Phi} \in \Omega$ empregando as relações entre $f(\mathbf{y}; \mathbf{\Phi})$ e $g(\mathbf{x}; \mathbf{\Phi})$.

Visando o objetivo acima, considere a densidade condicional de $\mathbf{y}|\mathbf{x}$ dada por

$$k(\mathbf{y}|\mathbf{x}; \mathbf{\Phi}) = \frac{f(\mathbf{y}; \mathbf{\Phi})}{g(\mathbf{x}; \mathbf{\Phi})},$$

a partir da qual podemos escrever

$$\ln g(\mathbf{x};\mathbf{\Phi}) = \ln f(\mathbf{y};\mathbf{\Phi}) - \ln k(\mathbf{y}|\mathbf{x};\mathbf{\Phi}). \tag{3.12}$$

A idéia do algoritmo é tomar a esperança condicional em (3.12) com relação a variável desconhecida $\mathbf{z}$, dado $\mathbf{x}$ e um valor $\mathbf{\Phi}^{'} \in \Omega$, então, dessa forma, obtemos

$$E_{\mathbf{z}}\{\ln g(\mathbf{x};\mathbf{\Phi})|\mathbf{x},\mathbf{\Phi}^{'}\} = E_{\mathbf{z}}\{\ln f(\mathbf{y};\mathbf{\Phi})|\mathbf{x},\mathbf{\Phi}^{'}\} - E_{\mathbf{z}}\{\ln k(\mathbf{y}|\mathbf{x};\mathbf{\Phi})|\mathbf{x},\mathbf{\Phi}^{'}\}, \tag{3.13}$$

onde assumimos que as esperanças existem para todos os pares $(\mathbf{\Phi},\mathbf{\Phi}^{'})$.

Como em (3.13) temos que $E_{\mathbf{z}}\{\ln g(\mathbf{x};\mathbf{\Phi})|\mathbf{x},\mathbf{\Phi}^{'}\} = \ln g(\mathbf{x};\mathbf{\Phi})$, adotando a notação mais comumente empregada, escrevemos (3.13) como

$$L(\mathbf{\Phi}) = Q(\mathbf{\Phi}|\mathbf{\Phi}^{'}) - H(\mathbf{\Phi}|\mathbf{\Phi}^{'}), \tag{3.14}$$

onde

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{'}) = E_{\mathbf{z}}\{\ln f(\mathbf{y};\mathbf{\Phi})|\mathbf{x},\mathbf{\Phi}^{'}\}$$

e

$$H(\mathbf{\Phi}|\mathbf{\Phi}^{'}) = E_{\mathbf{z}}\{\ln k(\mathbf{y}|\mathbf{x};\mathbf{\Phi})|\mathbf{x},\mathbf{\Phi}^{'}\}.$$

Com a notação acima, o algoritmo EM é definido como segue: dado uma aproximação $\mathbf{\Phi}^{(s)}$ para um maximizador de $L(\mathbf{\Phi})$, obtemos a aproximação $\mathbf{\Phi}^{(s+1)}$ por:

1. Passo E: Determinar $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)})$.
2. Passo M: Escolher $\mathbf{\Phi}^{(s+1)} \in arg\,max_{\mathbf{\Phi}\in\Omega}\, Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)})$

A notação $arg\,max_{\mathbf{\Phi}\in\Omega} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)})$ significa o conjunto dos valores de $\mathbf{\Phi} \in \Omega$ que maximizam $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)})$ sobre Ω. A sequência de aproximações geradas pelo algoritmo EM será denotada por $\{\mathbf{\Phi}^{(s)}\} = \{\mathbf{\Phi}^{(s)} : s = 0, 1, 2, 3, \ldots\}$, onde $\mathbf{\Phi}^{(0)}$ é um valor inicial fornecido ao algoritmo.

A aplicação do algoritmo consiste em repetir, alternadamente, os passos E e M. Numa dada repetição do passo E, a aproximação $\mathbf{\Phi}^{(s)}$ é substituida pela aproximação $\mathbf{\Phi}^{(s+1)}$ determinada na última repetição do passo M.

Na definição acima, a forma de determinar $\mathbf{\Phi}^{(s+1)}$ garante que $Q(\mathbf{\Phi}^{(s+1)}|\mathbf{\Phi}^{(s)}) \geq Q(\mathbf{\Phi}^{(s)}|\mathbf{\Phi}^{(s)})$, além disso, empregando a desigualdade de Jensen para esperança condicional (Ash (1972, pag. 287)) pode ser mostrado que $H(\mathbf{\Phi}^{(s+1)}|\mathbf{\Phi}^{(s)}) \leq H(\mathbf{\Phi}^{(s)}|\mathbf{\Phi}^{(s)})$ (veja Dempester *et al.* (1977)). Esses dois resultados implicam que

$$\begin{aligned} L(\mathbf{\Phi}^{(s+1)}) - L(\mathbf{\Phi}^{(s)}) &= \{Q(\mathbf{\Phi}^{(s+1)}|\mathbf{\Phi}^{(s)}) - Q(\mathbf{\Phi}^{(s)}|\mathbf{\Phi}^{(s)})\} \\ &\quad + \{H(\mathbf{\Phi}^{(s)}|\mathbf{\Phi}^{(s)}) - H(\mathbf{\Phi}^{(s+1)}|\mathbf{\Phi}^{(s)})\} \geq 0, \end{aligned}$$

ou seja,

$$L(\mathbf{\Phi}^{(s+1)}) \geq L(\mathbf{\Phi}^{(s)}). \tag{3.15}$$

O resultado em (3.15) implica que as aproximações geradas pelo algoritmo EM, $\{\mathbf{\Phi}^{(s)}\}$, geram uma sequência $\{L(\mathbf{\Phi}^{(s)})\}$ monótona não-decrescente e, sendo limitada superiormente, essa sequência converge para algum L^* (Rudin (1976, pag. 55)). Essa é, portanto, uma propriedade fundamental do EM que o qualifica como um algoritmo para maximizar a função de log-verossimilhança (equivalentemente, a função de verossimilhança).

Os aspectos teóricos da convergência do algoritmo EM foram discutidos em Wu (1983), onde são abordadas duas questões fundamentais: (i) a caracterização da convergência de $L(\mathbf{\Phi}^{(s)})$, no sentido de verificar se é para um máximo global sobre Ω, ou para um máximo local ou, ainda, para um ponto estacionário; e (ii) a convergência da sequência de estimativas $\mathbf{\Phi}^{(s)}$ geradas pelo algoritmo. As propriedades estabelecidas nessa discussão, estão resumidas em um teorema estabelecido em Redner e Walker (1984) que apresentamos a seguir. Ressaltamos, entretanto, que os resultados estabelecidos no teorema são válidos para o algoritmo EM em geral, não apenas no contexto de mistura finita de densidades.

Teorema 3.6.1 *Para $\mathbf{\Phi}^{(0)} \in \Omega$, seja $\{\mathbf{\Phi}^{(s)}\}$ uma sequência em Ω gerada pelo algoritmo EM. Então, a função de log-verossimilhança, $L(\mathbf{\Phi})$, é monótona não-decrescente sobre $\{\mathbf{\Phi}^{(s)}\}$, convergindo para um limite L^* que pode ser infinito. Também, sendo $\mathcal{L}$ o conjunto dos pontos limites de $\{\mathbf{\Phi}^{(s)}\}$, temos que:*

(i) $\mathcal{L}$ *é um conjunto fechado em* Ω.

(ii) Se $\{\mathbf{\Phi}^{(s)}\}$ *está contida em um subconjunto compacto de* Ω, *então* $\mathcal{L}$ *é compacto.*

(iii) Se $\{\mathbf{\Phi}^{(s)}\}$ *está contida em um subconjunto compacto de* Ω *e* $\lim_{s\to\infty} \|\mathbf{\Phi}^{(s+1)} - \mathbf{\Phi}^{(s)}\| = 0$, *para uma norma sobre* Ω, *então* $\mathcal{L}$ *é conectado e compacto.*

(iv) Se $L(\mathbf{\Phi})$ *é contínua em* Ω *e* $\mathcal{L} \neq \emptyset$, *então,* L^* *é finito e* $L(\widehat{\mathbf{\Phi}}) = L^*$ *para* $\widehat{\mathbf{\Phi}} \in \mathcal{L}$.

(v) Se $Q(\mathbf{\Phi}|\mathbf{\Phi}')$ *e* $H(\mathbf{\Phi}|\mathbf{\Phi}')$ *são contínuas para* $\mathbf{\Phi}$ *e* $\mathbf{\Phi}'$ *em* Ω, *então, cada* $\widehat{\mathbf{\Phi}} \in \mathcal{L}$ *satisfaz* $\widehat{\mathbf{\Phi}} \in arg\, max_{\mathbf{\Phi}\in\Omega}\, Q(\mathbf{\Phi}|\mathbf{\Phi}')$.

(vi) Se $Q(\mathbf{\Phi}|\mathbf{\Phi}')$ *e* $H(\mathbf{\Phi}|\mathbf{\Phi}')$ *são contínuas para* $\mathbf{\Phi}$ *e* $\mathbf{\Phi}'$ *em* Ω *e diferenciáveis em* $\mathbf{\Phi}$,

para $\mathbf{\Phi} = \mathbf{\Phi}^{'} = \hat{\mathbf{\Phi}} \in \mathcal{L}$*, então,* $L(\mathbf{\Phi})$ *é diferenciável em* $\mathbf{\Phi} = \hat{\mathbf{\Phi}}$ *e as equações de verossimilhança são satisfeitas para* $\mathbf{\Phi} = \hat{\mathbf{\Phi}}$.

Prova: Wu (1983); Redner e Walker (1984). □

O Teorema 3.6.1 estabelece os resultados globais de convergência para a sequência gerada pelo algoritmo EM. As declarações em $(i), (ii)$ e (iii), são genéricas, válidas para qualquer sequência numérica. As declarações em $(iv), (v)$ e (vi), estão baseadas nos fatos já mencionados de que a função de log-verossimilhança é monótona não-decrescente sobre a sequência gerada pelo algoritmo. Para uma discussão mais aprofundada do ponto de vista técnico, sugerimos consultar Redner e Walker (1984), onde são apresentados outros resultados complementares. Como um exemplo, no caso onde as densidades componentes $f(\mathbf{x}; \boldsymbol{\theta}_j)$ são completamente especificadas, sendo necessário apenas estimar as proporções da mistura, os autores mostram que as estimativas geradas pelo EM convergem para o estimador de máxima verossimilhança.

Um aspecto interessante discutido em Wu (1983), é a questão de que a convergência de $L(\mathbf{\Phi}^{(s)})$ para um L^* não implica a convergência $\mathbf{\Phi}^{(s)}$ para um ponto $\mathbf{\Phi}^*$. Um exemplo sobre essa questão é apresentado em Boyles (1983), onde $L(\mathbf{\Phi}^{(s)})$ é limitada superiormente e, no entanto, $\mathbf{\Phi}^{(s)}$ converge para um círculo de raio unitário e não para um único ponto. A despeito do Teorema 3.6.1 caracterizar o conjunto de pontos limites da sequência $\{\mathbf{\Phi}^{(s)}\}$, ele não estabelece se essa sequência, no caso de convergir para um valor finito, converge para um estimador de máxima verossimilhança. Visando responder essa questão, Redner e Walker (1984) apresentam um teorema que dá as condições sob as quais se obtém localmente essa convergência. O teorema, que assume as condições dos Teoremas 3.4.1 e

3.4.2, para as quais nos referimos como *condições de regularidades*, é dado a seguir.

Teorema 3.6.2 *Suponha que as condições de regularidades são satisfeitas em* Ω. *Seja* Ω_o *um subconjunto compacto de* Ω, *que contém* $\mathbf{\Phi}^*$ *no seu interior, para o qual* $p(\mathbf{x};\mathbf{\Phi}) = p(\mathbf{x};\mathbf{\Phi}^*)$ *em q.t.p. em* $\mathbf{x}$, *para* $\mathbf{\Phi} \in \Omega_o$, *somente se* $\mathbf{\Phi} = \mathbf{\Phi}^*$. *Suponha ainda que, com probabilidade* 1, *a função* $Q(\mathbf{\Phi}|\mathbf{\Phi}^{'})$ *é contínua e diferenciável para* $\mathbf{\Phi}$ *e* $\mathbf{\Phi}^{'}$ *em* Ω_o *e , além disso, que* $L(\mathbf{\Phi})$ *é diferenciável em* $\mathbf{\Phi} \in \Omega_o$, *sempre que* n *for suficientemente grande. Também, para* $\mathbf{\Phi}^{(0)} \in \Omega_o$, *seja* $\{\mathbf{\Phi}^{(s)}\}$ *a sequência gerada pelo algoritmo EM em* Ω_o, *ou seja, uma sequência que satisfaz*

$$\mathbf{\Phi}^{(s+1)} \in arg\,max_{\mathbf{\Phi}\in\Omega_o} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)}).$$

Então, se n *é suficientemente grande, com probabilidade* 1, *a única estimativa de máxima verossimilhança consistente,* $\hat{\mathbf{\Phi}}_n$, *é bem definida em* Ω_o *e* $\hat{\mathbf{\Phi}}_n = \lim_{s\to\infty} \mathbf{\Phi}^{(s)}$, *se* $\mathbf{\Phi}^{(0)}$ *está suficientemente próximo de* $\hat{\mathbf{\Phi}}_n$.

Prova: Redner e Walker (1984). □

As suposições fundamentais no teorema acima é o fato de que o valor inicial deve estar numa vizinhança de $\hat{\mathbf{\Phi}}_n$, em um subconjunto compacto de Ω, e a continuidade de $L(\mathbf{\Phi})$. Dado que a sequência $\{\mathbf{\Phi}^{(s)}\}$ está nessa vizinhança, pelo item (iv) do Teorema 3.6.1, cada ponto limite desta sequência também estará na vizinhança e, desde que $\hat{\mathbf{\Phi}}_n$ é a solução única das equações de verossimilhança, decorre o resultado de que a sequência converge para o estimador de máxima verossimilhança, considerando a continuidade de $L(\mathbf{\Phi})$.

A utilização do Teorema 3.6.2 na prática pode não ser tão imediato, não é uma questão simples determinar um valor inicial que saibamos estar próximo da solução desejada. Uma

situação onde haveria alguma orientação, seria no caso onde houvesse uma amostra suficientemente grande e com as observações identificadas quanto a procedência com relação as densidades componentes. Sendo conhecidas as distribuições das componentes, poderíamos dar como o valor inicial, $\mathbf{\Phi}^{(0)}$, as estimativas obtidas separadamente para cada componente, onde os valores iniciais das proporções da mistura seriam dados pelas respectivas proporções amostrais das observações para cada componente.

Resultados importantes para ao algoritmo EM, se verificam para misturas finitas cujas componentes são membros da família exponencial de distribuições (veja Seção A.3 para definições). Para esses casos, cada sucessiva aproximação $\mathbf{\Phi}^{(s+1)}$ do estimador de máxima verossimilhança é unica e explicitamente determinada a partir da aproximação anterior $\mathbf{\Phi}^{(s)}$, quase sempre de uma forma contínua. Pode ser mostrado também, veja Seção 3.6.3, que as aproximação $\Phi_j^{(s+1)}$ das componentes de $\mathbf{\Phi}^{(s+1)}$ são funções das estatísticas suficientes para as distribuições em questão (veja a Definição A.3.1 para estatística suficiente).

O teorema a seguir, trata das condições de convergência da sequência gerada pelo EM para mistura finitas com componentes na família exponencial. Sob algumas condições, o teorema estabelece que, com n suficientemente grande, para a sequência $\{\mathbf{\Phi}^{(s)}\}$ de aproximações do EM, temos que $\mathbf{\Phi}^{(s)} \to \widehat{\mathbf{\Phi}}_n$, a uma taxa linear, quando $s \to \infty$.

Teorema 3.6.3 *Suponha que a matriz de informação de Fisher seja definida positiva em* $\mathbf{\Phi}^*$ *e que temos todos* $\alpha_{i's}^* > 0$. *Para* $\mathbf{\Phi}^{(0)} \in \Omega$ *seja* $\{\mathbf{\Phi}^{(s)}\}$ *a sequência gerada pelo algoritmo EM em* Ω. *Então, para* n *suficientemente grande, com probabilidade* 1, *a única solução consistente* $\widehat{\mathbf{\Phi}}_n$ *das equações de máxima verossimilhança é bem definida e existe*

uma norma $\|\cdot\|$ *sobre* Ω*, na qual* $\{\boldsymbol{\Phi}^{(s)}\}$ *converge linearmente para* $\widehat{\boldsymbol{\Phi}}_n$*, sempre que* $\boldsymbol{\Phi}^{(0)}$ *for suficientemente próximo de* $\widehat{\boldsymbol{\Phi}}_n$*, isto é, existe uma constante* λ*,* $0 \leq \lambda < 1$*, para a qual*

$$\|\boldsymbol{\Phi}^{(s+1)} - \widehat{\boldsymbol{\Phi}}_n\| \leq \lambda \|\boldsymbol{\Phi}^{(s)} - \widehat{\boldsymbol{\Phi}}_n\| \quad s = 0, 1, 2, 3, \ldots,$$

sempre que $\boldsymbol{\Phi}^{(0)}$ *for suficientemente próximo de* $\widehat{\boldsymbol{\Phi}}_n$*.*

Prova: Redner e Walker (1984). □

A prova do teorema baseia-se no fato de que a condição de regularidade 1 está satisfeita e, se necessário, por restringir Ω para uma pequena vizinhança de $\boldsymbol{\Phi}^*$, a condição 2 também está satisfeita para o tipo de mistura sob consideração. Com essas considerações, o Teorema 3.4.1 se aplica. Impondo a condição que temos $\boldsymbol{\Phi}^{(0)}$ suficientemente próximo de $\widehat{\boldsymbol{\Phi}}_n$ é demonstrado a existência da norma $\|\cdot\|$ e da constante λ.

Uma discussão sobre estimação de máxima verossimilhança com dados incompletos da família exponencial, incluindo mistura finita de densidades, é dada em Sundberg (1974).

Uma ampla discussão quanto ao desempenho do algoritmo EM em aplicações práticas, bem como comparações com outros métodos para aproximação numérica da estimativa de máxima verossimilhança, é apresentada em Redner e Walker (1984). O método de Newton, o método de "scoring" e os métodos de gradiente conjugado são alguns dos métodos comparados ao EM. Da análise de suas comparações e das referências consideradas na discussão, esses autores concluem que por sua simplicidade, sua propriedade de não decrescer a função de verossimilhança, aliadas às suas boas propriedades de convergência, o algoritmo EM se constitue numa opcão muito boa para determinação do EMV em misturas finitas de densidades.

Por ser uma questão importante na prática, mencionamos que na literatura há relatos considerando que a convergência do algoritmo EM é relativamente lenta. Com relação a essa questão, há várias propostas de extensões do EM visando acelerar sua taxa de convergência, mantendo sua simplicidade e estabilidade. Embora muito importante, esta questão não será discutida aqui, para isso, indicamos o trabalho de Meng e van Dyk (1997) para uma discussão sobre o assunto. Outra referências são Jamshidian e Jennrich (1997) e Liu, Rubin e Wu (1998).

A literatura sobre o algoritmo EM é bastante extensa e dispersa em várias publicações científicas. Em Moon (1996), no entanto, estão relacionadas mais de 50 referências que abordam desde a teoria do EM, até as suas aplicações nas mais diversas áreas.

3.6.2 O Algoritmo EM para Mistura Finita de Densidades

Considere que dispomos de um conjunto $\mathcal{A}_{(n)} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}$, de observações independentes de um vetor aleatório $\mathbf{X}$, cuja distribuição é dada por uma mistura finita de densidades

$$p(\mathbf{x}; \boldsymbol{\Phi}) = \sum_{j=1}^{k} \alpha_j f_j(\mathbf{x}; \boldsymbol{\theta}_j), \quad \mathbf{x} \in \Re^d, \tag{3.16}$$

onde $\boldsymbol{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$, com $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_k)$ e $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_k)$. Nas aplicações que consideramos neste trabalho, as observações em $\mathcal{A}_{(n)}$ correspondem aos exemplos de uma dada classe.

Como mencionado na seção anterior, a abordagem a ser adotada é modelar o problema no contexto de dados incompletos. Para esse fim, considere agora um conjunto de ve-

tores aleatórios $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \ldots, \mathbf{z}_n$, com $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3}, \ldots, z_{ik})$, onde as componentes são definidas como variáveis indicadoras da forma

$$z_{ij} = \begin{cases} 1 & \text{se} \quad \mathbf{x}_i \sim f_j(\cdot; \boldsymbol{\theta}_j) \\ 0 & \text{se} \quad \mathbf{x}_i \sim f_l(\cdot; \boldsymbol{\theta}_l), \quad l \neq j. \end{cases}$$

Assumimos que $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \ldots, \mathbf{z}_n$ são independentes e identicamentes distribuidos com distribuição Multinomial(1, $\boldsymbol{\alpha}$), onde $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_k)$. É assumido, também, que $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n$ dado $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \ldots, \mathbf{z}_n$, respectivamente, são condicionalmente independentes. No contexto sendo considerado, as variáveis $\mathbf{z}_i$ representam as variáveis latentes não observáveis com informação sobre a distribuição de cada $\mathbf{x}_i$. Em analogia com exposição da seção anterior, os $\mathbf{x}_{i's}$ são os dados incompletos e $\mathbf{y}_i = (\mathbf{x}_i, \mathbf{z}_i)$ são os dados completos, para $i = 1, 2, 3, ..., n$.

A abordagem considerada aqui para $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3}, \ldots, z_{ik})$, às vezes recebe a denominação de *modelo multinomial oculto*. Se não há independência e assumindo que essas variáveis formam uma cadeia de Markov, então o modelo recebe a denominação de *modelo oculto de cadeia de Markov*, um modelo muito utilizado para modelar os problemas em reconhecimento de fala.

Com as pressuposições assumidas, temos que a densidade conjunta de $\mathbf{y}_i$ é dada por

$$\begin{aligned} f(\mathbf{y}_i; \boldsymbol{\Phi}) &= f(\mathbf{x}_i | \mathbf{z}_i; \boldsymbol{\Phi}) f(\mathbf{z}_i; \boldsymbol{\Phi}) \\ &= \prod_{j=1}^{k} (f_j(\mathbf{x}_i; \boldsymbol{\theta}_j))^{z_{ij}} f(\mathbf{z}_i; \boldsymbol{\Phi}) \end{aligned} \tag{3.17}$$

Em (3.17), pela distribuição assumida para $\mathbf{z}_i$, temos

$$f(\mathbf{z}_i;\boldsymbol{\Phi}) = f(\mathbf{z}_i;\boldsymbol{\alpha}) = \prod_{j=1}^{k} \alpha_j^{z_{ij}}. \tag{3.18}$$

Substituindo (3.18) em (3.17) obtemos

$$f(\mathbf{y}_i;\boldsymbol{\Phi}) = \prod_{j=1}^{k} \alpha_j^{z_{ij}} (f_j(\mathbf{x}_i;\boldsymbol{\theta}_j))^{z_{ij}}. \tag{3.19}$$

Considerando agora a independência dos dados completos e (3.19), temos que a função de verossimilhança é dada por

$$\begin{aligned} l(\boldsymbol{\Phi}) &= \prod_{i=1}^{n} f(\mathbf{y}_i;\boldsymbol{\Phi}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k} \alpha_j^{z_{ij}} (f_j(\mathbf{x}_i;\boldsymbol{\theta}_j))^{z_{ij}} \end{aligned} \tag{3.20}$$

De (3.20), vemos que a função de log-verossimilhança para os dados completos é dada por

$$L(\boldsymbol{\Phi}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \ln \alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \ln f_j(\mathbf{x}_i;\boldsymbol{\theta}_j). \tag{3.21}$$

No Passo E do algoritmo EM, devemos determinar

$$\begin{aligned} Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)}) &= E_{\mathbf{z}}\{\ln l(\boldsymbol{\Phi})|\mathcal{A}_{(n)}, \boldsymbol{\Phi}^{(s)}\} \\ &= E_{\mathbf{z}}\{L(\boldsymbol{\Phi})|\mathcal{A}_{(n)}, \boldsymbol{\Phi}^{(s)}\}, \end{aligned} \tag{3.22}$$

onde, como na seção anterior, $\boldsymbol{\Phi}^{(s)} = (\alpha_1^{(s)}, \alpha_2^{(s)}, \ldots, \alpha_k^{(s)}, \boldsymbol{\theta}_1^{(s)}, \boldsymbol{\theta}_2^{(s)}, \ldots, \boldsymbol{\theta}_k^{(s)})$ representa uma dada aproximação de $\boldsymbol{\Phi}$.

Como $\mathcal{A}_{(n)}$ é fixado e a esperança condicional é tomada somente para as variáveis z_{ij}, então, (3.22) torna-se

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s)}) = \sum_{i=1}^{n}\sum_{j=1}^{k} E[z_{ij}|\mathcal{A}_{(n)}, \mathbf{\Phi}^{(s)}] \ln \alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{k} E[z_{ij}|\mathcal{A}_{(n)}, \mathbf{\Phi}^{(s)}] \ln f_j(\mathbf{x}_i; \boldsymbol{\theta}_j). \quad (3.23)$$

De (3.23), vemos que é necessário determinar

$$\tau_{ij}^{(s)} \stackrel{\text{def}}{=} E[z_{ij}|\mathcal{A}_{(n)}, \mathbf{\Phi}^{(s)}] = Pr[z_{ij} = 1|\mathcal{A}_{(n)}, \mathbf{\Phi}^{(s)}] = Pr[z_{ij} = 1|\mathbf{x}_i, \mathbf{\Phi}^{(s)}], \quad (3.24)$$

onde a segunda igualdade decorre da definição de z_{ij} e, a última, devido a independência dos $\mathbf{z}_{i's}$. Considerando a distribuição dada em (3.18), temos que

$$Pr[z_{ij} = 1|\mathbf{\Phi}^{(s)}] = Pr[z_{ij} = 1, z_{il} = 0\,\forall\, l \neq j|\mathbf{\Phi}^{(s)}] = \alpha_j^{(s)}. \quad (3.25)$$

Usando (3.17), vemos que

$$f(\mathbf{x}_i|\mathbf{z}_i; \mathbf{\Phi}^{(s)}) = \prod_{j=1}^{k} (f_j(\mathbf{x}_i; \boldsymbol{\theta}_j^{(s)}))^{z_{ij}}, \quad (3.26)$$

e, portanto, temos que

$$f(\mathbf{x}_i|z_{ij} = 1; \mathbf{\Phi}^{(s)}) = f_j(\mathbf{x}_i; \boldsymbol{\theta}_j^{(s)}). \quad (3.27)$$

Empregando o teorema de Bayes, vemos que

$$\tau_{ij}^{(s)} = Pr[z_{ij} = 1|\mathbf{x}_i, \mathbf{\Phi}^{(s)}] = \frac{Pr[z_{ij} = 1|\mathbf{\Phi}^{(s)}] f(\mathbf{x}_i|z_{ij} = 1; \mathbf{\Phi}^{(s)})}{p(\mathbf{x}_i; \mathbf{\Phi}^{(s)})}. \quad (3.28)$$

Usando (3.16), (3.25) e (3.27) em (3.28), obtemos

$$\tau_{ij}^{(s)} = \frac{\alpha_j^{(s)} f_j(\mathbf{x}_i; \boldsymbol{\theta}_j^{(s)})}{\sum_{t=1}^{k} \alpha_t^{(s)} f_t(\mathbf{x}_i; \boldsymbol{\theta}_t^{(s)})}. \quad (3.29)$$

Da expressão acima, vemos que $\tau_{ij}^{(s)}$ representa uma estimativa da probabilidade de $\mathbf{x}_i$ pertencer a uma população cuja distribuição é dada por $f_j(\cdot;\boldsymbol{\theta}_j^{(s)})$, com base em uma dada estimativa $\boldsymbol{\Phi}^{(s)}$ do vetor de parâmetros $\boldsymbol{\Phi}$.

Agora, usando (3.29) em (3.23), podemos escrever

$$\begin{aligned} Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)}) &= \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}\ln\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}\ln f_j(\mathbf{x}_i;\boldsymbol{\theta}_j) \\ &= Q_1(\boldsymbol{\alpha}) + Q_2(\boldsymbol{\theta}), \end{aligned} \tag{3.30}$$

onde

$$Q_1(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}\ln\alpha_j \tag{3.31}$$

e

$$Q_2(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}\ln f_j(\mathbf{x}_i;\boldsymbol{\theta}_j) \tag{3.32}$$

Como $Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)})$ pode ser escrita como a soma de uma função de $\boldsymbol{\alpha}$ e uma função de $\boldsymbol{\theta}$, o problema de maximização pode ser considerado, separadamente, para $\boldsymbol{\alpha}$ e para $\boldsymbol{\theta}$. De (3.30), temos que o passo de maximização de $Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)})$ dá origem a dois problemas separados de maximização

$$\frac{\partial Q_1(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} = \mathbf{0} \quad \text{e} \quad \frac{\partial Q_2(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0}.$$

Como em $Q_1(\boldsymbol{\alpha})$ temos que $\ln\alpha_1, \ln\alpha_2, \ldots, \ln\alpha_k$ estão linearmente relacionados, o primeiro problema de maximização tem uma solução única que é explicitamente determinada, independente da forma funcional das componentes da mistura. No problema de maximização com relação a $\boldsymbol{\theta}$, vemos que a solução depende da forma funcional das componentes

$f_j(\mathbf{x}_i;\boldsymbol{\theta}_j)$. É interessante observar que, se as componentes de $\boldsymbol{\theta}$ forem variáveis mutuamentes independentes, a maximização se divide em k problemas, cada um envolvendo somente um $\boldsymbol{\theta}_j$, ou seja, devemos resolver

$$\frac{\partial Q_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \mathbf{0} \qquad j = 1, 2, 3, ..., k.$$

As aproximações $\alpha_j^{(s)}$:

Para determinar as aproximações $\alpha_j^{(s)}$, devido a restrição $\sum_{j=1}^{k} \alpha_j = 1$, empregamos o método dos multiplicadores de Lagrange para a resolução do problema de maximização envolvendo $\boldsymbol{\alpha}$. Escrevemos

$$M(\boldsymbol{\alpha}) = Q_1(\boldsymbol{\alpha}) + \lambda G(\boldsymbol{\alpha}), \tag{3.33}$$

onde $G(\boldsymbol{\alpha}) = 1 - \sum_{j=1}^{k} \alpha_j$. Tomando as derivadas parciais em (3.33) com relação a α_j e λ, temos, respectivamente,

$$\frac{\partial M(\boldsymbol{\alpha})}{\partial \alpha_j} = \frac{1}{\alpha_j} \sum_{i=1}^{n} \tau_{ij}^{(s)} - \lambda \quad \text{e} \quad \frac{\partial M(\boldsymbol{\alpha})}{\partial \lambda} = 1 - \sum_{j=1}^{k} \alpha_j.$$

Igualando a zero cada uma das derivas parciais acima, obtemos

$$\left.\frac{\partial M(\boldsymbol{\alpha})}{\partial \alpha_j}\right|_{\alpha_j = \alpha_j^{(s+1)}} = 0 \quad \Rightarrow \quad \alpha_j^{(s+1)} = \frac{1}{\lambda} \sum_{i=1}^{n} \tau_{ij}^{(s)} \tag{3.34}$$

e

$$\left.\frac{\partial M(\boldsymbol{\alpha})}{\partial \lambda}\right|_{\alpha_j = \alpha_j^{(s+1)}} = 0 \quad \Rightarrow \quad \sum_{j=1}^{k} \alpha_j^{(s+1)} = 1 \tag{3.35}$$

Combinando (3.34) e (3.35), chegamos a

$$1 = \sum_{j=1}^{k} \alpha_j^{(s+1)} = \frac{1}{\lambda} \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij}^{(s)}. \tag{3.36}$$

Agora, considerando que $\sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij}^{(s)} = \sum_{i=1}^{n} 1 = n$, em (3.36) temos $\lambda = n$ e, substituindo em (3.34), chegamos a

$$\alpha_j^{(s+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ij}^{(s)}, \qquad j = 1, 2, ..., k. \tag{3.37}$$

Dos resultados obtidos até aqui, temos que, dada uma aproximação $\mathbf{\Phi}^{(s)}$, determinamos $\tau_{ij}^{(s)}$ através de (3.29) e as aproximações $\alpha_j^{(s+1)}$ são determinadas por meio de (3.37). As aproximações $\boldsymbol{\theta}_j^{(s+1)}$ são determinadas através da solução de

$$\sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij}^{(s)} \frac{\partial}{\partial \boldsymbol{\theta}} \ln f_j(\mathbf{x}_i; \boldsymbol{\theta}_j) = \mathbf{0}, \tag{3.38}$$

ou, no caso em que $\boldsymbol{\theta}_1^{(s)}, \boldsymbol{\theta}_2^{(s)}, \boldsymbol{\theta}_3^{(s)}, \ldots, \boldsymbol{\theta}_k^{(s)}$ são mutuamente independentes, pela solução de

$$\sum_{i=1}^{n} \tau_{ij}^{(s)} \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln f_j(\mathbf{x}_i; \boldsymbol{\theta}_j) = \mathbf{0}, \quad j = 1, 2, 3, ..., k. \tag{3.39}$$

Da solução em (3.37), vemos que na determinação de $\alpha_j^{(s+1)}$ estão satisfeitas as restrições de serem não negativos e somarem 1. Para o problema em (3.38), ou em (3.39), não podemos garantir, a princípio, uma solução única e bem definida. Para esse problema, no entanto, para famílias de distribuições com o estimador de máxima verossimilhança determinado explicitamente e de forma única, cada aproximação $\boldsymbol{\theta}^{(s+1)}$ é também determinada explicitamente de forma única. Como será visto a seguir, isso se verifica quando as componentes da mistura são membros da família exponencial.

3.6.3 O EM para componentes na Família Exponencial

Em muitas aplicações, as densidades componentes da mistura finita são membros da família exponencial de distribuições. No Apêndice A, definimos a família exponencial e comentamos algumas de suas principais propriedades. Se as densidades componentes são membros da família exponencial, com parametrização de valor médio, obtemos uma solução explícita em (3.39). Para ver isso, considere as componentes na parametrização de valor médio com parâmetro $\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)$

$$f_j(\mathbf{x}_i;\boldsymbol{\theta}_j) = g(\mathbf{x}_i;\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)) = \frac{1}{a(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))} b_j(\mathbf{x}_i)\exp\{\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)^T\mathbf{t}_j(\mathbf{x}_i)\}, \tag{3.40}$$

onde temos que $\boldsymbol{\theta}_j = E(t(\mathbf{X})|\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)$.

Substituindo (3.40) em (3.39), temos que

$$\begin{aligned}\sum_{i=1}^{n}\tau_{ij}^{(s)}\frac{\partial}{\partial\boldsymbol{\theta}_j}\ln\{\frac{1}{a(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))}b_j(\mathbf{x}_i)\exp\{\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)^T\mathbf{t}_j(\mathbf{x}_i)\} &= \mathbf{0}\\ \sum_{i=1}^{n}\tau_{ij}^{(s)}\{\frac{a^{'}(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))}{a(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))}+\mathbf{t}_j(\mathbf{x}_i)\} &= \mathbf{0}, \quad j=1,2,3,...,k,\end{aligned} \tag{3.41}$$

onde $a^{'}(\cdot)$ representa a derivada de $a(\cdot)$.

Agora, temos que,

$$\begin{aligned}\frac{a^{'}(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))}{a(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))} &= \frac{1}{a(\boldsymbol{\phi}_j(\boldsymbol{\theta}_j))}\int \mathbf{t}_j(\mathbf{x}_i)b_j(\mathbf{x}_i)\exp\{\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)^T\mathbf{t}_j(\mathbf{x}_i)\}\,d\mu\\ &= E[\mathbf{t}_j(\mathbf{X})|\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)_j)] = \boldsymbol{\theta}_j^{(s)},\end{aligned} \tag{3.42}$$

onde, na última igualdade, estamos considerando que é dada a aproximação $\boldsymbol{\theta}_j^{(s)}$ para o parâmetro $\boldsymbol{\theta}_j$.

Substituindo (3.42) em (3.41), chegamos a

$$\sum_{i=1}^{n} \tau_{ij}^{(s)} \boldsymbol{\theta}_j^{(s)} + \sum_{i=1}^{n} \tau_{ij}^{(s)} \mathbf{t}_j(\mathbf{x}_i) = \mathbf{0} \tag{3.43}$$

e obtemos a solução

$$\boldsymbol{\theta}_j^{(s+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(s)} \mathbf{t}_j(\mathbf{x}_i)}{\sum_{i=1}^{n} \tau_{ij}^{(s)}}. \tag{3.44}$$

Como mencionada anteriormente, de (3.44), vemos que cada aproximação $\boldsymbol{\theta}_j^{(s+1)}$ é função da estatística $\mathbf{t}_j$ a qual é estatística suficiente para $\boldsymbol{\theta}_j$. Em Redner e Walker (1984) são apresentados alguns exemplos de mistura finita de densidades da família exponencial, com discussões sobre as propriedades da sequência gerada pelo algoritmo EM em cada um dos exemplos.

É importante salientar que, a maioria das distribuições paramétricas empregadas na prática, são membros da família exponencial. As distribuições binomial, de Poisson, normal e a gama, são alguns exemplos de distribuições da família exponencial.

3.7 Mistura Finita de Densidades Normais

Como mencionado anteriormente, o modelo de mistura finita de normais é o mais empregado na prática. Nesta seção, discutimos algumas questões relativas a mistura finita de normais e desenvolvemos os passos do algoritmo EM para determinar as estimativas de máxima verossimilhança dos parâmetros que, nesse caso, são as proporções da mistura, os vetores de médias e as matrizes de covariâncias das componentes.

3.7.1 Mistura Finita de Normais

O modelo de mistura finita considerado é da forma

$$p(\mathbf{x};\mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}_j|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\}, \tag{3.45}$$

onde temos $\mathbf{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$, sendo $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_k)$ e $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_k)$ com $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)$. Na formulação em (3.45), temos que os vetores de médias $\boldsymbol{\mu}_j \in \Re^d$ e as matrizes de covariâncias $\mathbf{\Sigma}_j$ são matrizes $d \times d$ simétricas e definidas positivas.

Já vimos anteriormente, que a função de verossimilhança para mistura finita de normais com matrizes de covariâncias diferentes pode não ser limitada superiormente e, nesse caso, no sentido da definição clássica, o EMV não existe. Apesar dessa dificuldade, foi visto que existe um maximizador local para $L(\mathbf{\Phi})$, o EMV admitido neste trabalho, que é consistente e assintoticamente eficiente. Outra dificuldade é que, para mistura de normais, existem diversos maximizadores locais, as equações de verossimilhança têm várias raízes. A abordagem para lidar com essa dificuldade adicional, tem sido impor restrições sobre Ω, o espaço dos parâmetros.

Para mistura de normais univariadas ($d = 1$), Hathaway (1985) impôs a restrição de $\min_{i\neq j}(\frac{\sigma_i}{\sigma_j}) \geq c > 0$, com $c \in (0, 1]$, e mostrou que existe um maximizador global de $L(\mathbf{\Phi})$ sobre o espaço dos parâmetros restringido, assumindo que o conjunto de observações contém, pelo menos, $k + 1$ observações distintas e sob as condições de $\alpha_j \neq 0, \forall j$ e $(\mu_i, \sigma_i^2) \neq (\mu_j, \sigma_j^2), \forall i \neq j$. Para o caso multivariado, em Hathaway (1985) também é indicado que um máximo global restrito pode ser obtido impondo a restrição de que os auto-valores de $\mathbf{\Sigma}_i\mathbf{\Sigma}_j^{-1}$ $(1 \leq i \neq j \leq k)$ sejam iguais ou maiores que uma constante $c > 0$.

Em Redner (1981) é observado que, as condições para a consistência de uma maximizador global restrito de $L(\mathbf{\Phi})$, são satisfeitas por uma mistura de normais, onde a restrição é um subconjunto compacto de Ω. Esta restrição tem aplicações práticas, por exemplo, ao modelar os pixels de uma imagem, as médias e as variâncias das observações estarão restritas a um subconjunto compacto.

Outra forma de restrição, é impor que as matrizes de covariâncias das componentes sejam todas iguais. No modelo em (3.45), com $\mathbf{\Sigma}_j = \mathbf{\Sigma}$, $j = 1, 2, 3, ...k$, denominado *mistura finita de normais homocedásticas*, a idéia é "manter as matrizes de covariâncias afastadas da fronteira dos parâmetros" e, nesse caso, o EMV, no sentido da definição clássica, existe e é fortemente consistente (Basford e McLachlan (1985)). O emprego do modelo homocedástico, no entanto, nem sempre é adequado na prática e, para uma discussão sobre esse assunto, sugerimos Basford e McLachlan (1985) e McLachlan e Basford (1988, Seção 2.2).

Com visto anteriormente, para misturas finitas em geral, a existência dos resultados mencionados aqui, dependem das condições de regularidade. Em McLachlan e Basford (1988, Seção 2.1), os autores argumentam que essas condições, por serem relativamente fracas, possivelmente são satisfeitas por misturas finitas de densidades normais.

3.7.2 O Algoritmo EM para as Componentes Normais

Na seção anterior, vimos que a determinação das aproximações $\alpha_j^{(s+1)}$ são dadas explicitamente, porém, as aproximações $\boldsymbol{\theta}_j^{(s+1)}$ dependem da forma funcional das componentes $f_j(\cdot; \boldsymbol{\theta}_j)$. Descrevemos a seguir, as formas para as aproximações dos parâmetros para as

misturas finitas de normais.

As aproximações para α_j:

As aproximações $\alpha_j^{(s+1)}$, $j = 1, 2, 3, ..., k$, foram obtidas em (3.37), sendo dadas por

$$\alpha_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}^{(s)}, \tag{3.46}$$

onde os $\tau_{ij}^{(s)}$ são dados em (3.29). Substituindo a forma das $f_j(\cdot; \boldsymbol{\theta}_j)$ com distribuição normal na expressão para $\tau_{ij}^{(s)}$ e cancelando a constante $(2\pi)^{-\frac{d}{2}}$, obtemos

$$\tau_{ij}^{(s)} = \frac{\alpha_j^{(s)} |\boldsymbol{\Sigma}_j^{(s)}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T \boldsymbol{\Sigma}_j^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})\}}{\sum_{t=1}^{k} \alpha_t^{(s)} |\boldsymbol{\Sigma}_t^{(s)}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})^T \boldsymbol{\Sigma}_t^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})\}}. \tag{3.47}$$

Na mistura finita de normais, os parâmetros $\boldsymbol{\theta}_j$ são mutuamente independentes e, dessa forma, as aproximações $\boldsymbol{\theta}_j^{(s+1)}$ são determinadas pelas soluções das equações dadas em (3.39), ou seja

$$\sum_{i=1}^{n} \tau_{ij}^{(s)} \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln f(\mathbf{x}_i; \boldsymbol{\theta}_j) = \mathbf{0}, \quad j = 1, 2, 3, ..., k.$$

As aproximações para $\boldsymbol{\mu}_j$:

Das equações acima, vemos que é necessário derivar o logaritmo natural das densidades normais com relação a $\boldsymbol{\mu}_j$ e com relação a $\boldsymbol{\Sigma}_j$, uma vez que $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Então, para esse caso, temos que

$$\ln f(\mathbf{x}_i; \boldsymbol{\theta}_j) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j). \tag{3.48}$$

Usando o fato de que

$$-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j) = -\frac{1}{2}\{\mathbf{x}_i^T\boldsymbol{\Sigma}_j^{-1}\mathbf{x}_i - 2\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\mathbf{x}_i + \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\},$$

obtemos que a derivada de (3.48) com relação a $\boldsymbol{\mu}_j$ é dada por

$$\frac{\partial}{\partial\boldsymbol{\mu}_j}\ln f(\mathbf{x}_i;\boldsymbol{\theta}_j) = \boldsymbol{\Sigma}_j^{-1}\mathbf{x}_i - \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j. \tag{3.49}$$

Usando (3.49), obtemos para (3.39)

$$\sum_{i=1}^{n}\tau_{ij}^{(s)}\{\boldsymbol{\Sigma}_j^{-1}\mathbf{x}_i - \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\} = \mathbf{0}, \quad j=1,2,3,..,k. \tag{3.50}$$

Rearranjando os termos em (3.50) e multiplicando à esquerda por $\boldsymbol{\Sigma}_j$ os dois membros , chegamos a

$$\boldsymbol{\mu}_j^{(s+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(s)}\mathbf{x}_i}{\sum_{i=1}^{n}\tau_{ij}^{(s)}}, \quad j=1,2,3,...,k. \tag{3.51}$$

Observamos em (3.51) que as aproximações $\boldsymbol{\mu}_j^{(s+1)}$ são dadas por uma soma ponderada das observações, onde o fator de ponderação para cada $\mathbf{x}_i$ é a estimativa da probabilidade da observação ter distribuição $f(\mathbf{x}_i;\boldsymbol{\theta}_j)$.

As aproximações para $\boldsymbol{\Sigma}_j$:

Para obter as aproximações $\boldsymbol{\Sigma}_j^{(s+1)}$, primeiramente, usamos os seguintes resultados em (3.48): $\mathbf{x}^TA\mathbf{x} = tr(A\mathbf{x}\mathbf{x}^T)$ e $|B| = |B^{-1}|^{-1}$ para B não singular. Então, podemos escrever (3.48) como

$$\ln f(\mathbf{x}_i;\boldsymbol{\theta}_j) = -\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_j^{-1}| - \frac{1}{2}tr[\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)(\mathbf{x}_i-\boldsymbol{\mu}_j)^T]. \tag{3.52}$$

Desenvolvemos, agora, os passos para determinação de $\boldsymbol{\Sigma}_j^{(s+1)}$ como dado em Mardia *et al.* (1979, Seção 4.2), derivando as expressões com relação a $\boldsymbol{\Sigma}_j^{-1}$, empregando os seguintes resultados: $\frac{\partial}{\partial A}\ln(|A|) = 2A^{-1} - diag[A^{-1}]$ e $\frac{\partial}{\partial A}tr(A\mathbf{x}\mathbf{x}^T) = 2\mathbf{x}\mathbf{x}^T - diag[\mathbf{x}\mathbf{x}^T]$ (veja Graybill (1969)).

Usando os resultados acima para derivar (3.52) e rearranjando os termos, temos que

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_j^{-1}} \ln f(\mathbf{x}_i;\boldsymbol{\theta}_j) = 2[\boldsymbol{\Sigma}_j - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T] - diag[\boldsymbol{\Sigma}_j - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T]. \quad (3.53)$$

Agora, substituindo (3.53) em (3.39), obtemos

$$\sum_{i=1}^{n} \tau_{ij}^{(s)} \{2[\boldsymbol{\Sigma}_j - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T] - diag[\boldsymbol{\Sigma}_j - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T]\} = \mathbf{0}, \quad (3.54)$$

para $j = 1, 2, 3, ..., k$.

Fazendo $M_j = (\sum_{i=1}^{n} \tau_{ij}^{(s)})[\boldsymbol{\Sigma}_j - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T]$ em (3.54) e observando o fato de que $2M_j - diag[M_j] = \mathbf{0} \implies M_j = \mathbf{0}$, de (3.54) obtemos

$$\sum_{i=1}^{n} \tau_{ij}^{(s)} \boldsymbol{\Sigma}_j - \sum_{i=1}^{n} \tau_{ij}^{(s)} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T = \mathbf{0}. \quad (3.55)$$

A aproximação $\boldsymbol{\Sigma}_j^{(s+1)}$ é determinada pela solução de (3.55) usando a aproximação $\boldsymbol{\mu}_j^{(s+1)}$ como estimativa para $\boldsymbol{\mu}_j$. Dessa forma, chegamos a solução

$$\boldsymbol{\Sigma}_j^{(s+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(s)} (\mathbf{x}_i - \boldsymbol{\mu}_j^{(s+1)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s+1)})^T}{\sum_{i=1}^{n} \tau_{ij}^{(s)}}, \quad j = 1, 2, 3, ..., k. \quad (3.56)$$

Em (3.56) vemos, novamente, que as aproximações $\boldsymbol{\Sigma}_j^{(s+1)}$ são dadas por uma soma ponderada pelas estimativas de probabilidades $\tau_{ij}^{(s)}$.

Um problema com relação a (3.56), é que não há garantias de que a sequência de matrizes $\{\boldsymbol{\Sigma}_j^{(s)}\}$ gerada pelo algoritmo EM permanecerá limita inferiormente, no sentido de não convergir para uma solução singular. Pode ser mostrado, no entanto, que se um número suficiente de observações rotuladas forem incluídas no conjunto de observações esse problema será evitado. Em outras palavras, se tivermos n_j observações rotuladas como sendo provenientes da componente $f(\cdot;\boldsymbol{\theta}_j)$ e sendo $n_j > d$, então, com probabilidade 1, a sequência $\{\boldsymbol{\Sigma}_j^{(s)}\}$ é limitada inferiormente por uma matriz definida positiva e, assim, não terá matrizes singulares como pontos limites (veja Redner e Walker (1984)).

Componentes Homocedásticas, as aproximações para $\boldsymbol{\Sigma}$:

No caso das componentes serem homocedásticas, as aproximações $\alpha_j^{(s+1)}$ e $\boldsymbol{\mu}_j^{(s+1)}$ são dadas também, respectivamente, por (3.46) e (3.51), sendo que os valores de $\tau_{ij}^{(s)}$ são determinados substituindo as aproximações $\boldsymbol{\Sigma}_j^{(s)}$ pelas aproximações $\boldsymbol{\Sigma}^{(s)}$ em (3.47). Efetuando essa substituição e simplificando, obtemos

$$\tau_{ij}^{(s)} = \frac{\alpha_j^{(s)} \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T \boldsymbol{\Sigma}^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})\}}{\sum_{t=1}^{k} \alpha_t^{(s)} \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})^T \boldsymbol{\Sigma}^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})\}}. \tag{3.57}$$

Para a determinação de $\boldsymbol{\Sigma}^{(s+1)}$ devemos resolver

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{(s)} \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln f(\mathbf{x}_i;\boldsymbol{\theta}_j) = \mathbf{0}. \tag{3.58}$$

Dos resultados obtidos em (3.53), temos que

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{(s)} \{2[\boldsymbol{\Sigma} - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T] - diag[\boldsymbol{\Sigma} - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T]\} = \mathbf{0}. \tag{3.59}$$

Como anteriormente, fazendo $M = (\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)})[\boldsymbol{\Sigma} - (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T]$, substituindo em (3.59) e usando o fato de que $2M - diag[M] = \mathbf{0} \implies M = \mathbf{0}$, chegamos a

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}\boldsymbol{\Sigma} - \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T = \mathbf{0}. \tag{3.60}$$

Lembrando que $\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)} = n$ e usando a aproximação $\boldsymbol{\mu}_j^{(s+1)}$ como estimativa de $\boldsymbol{\mu}_j$, obtemos a solução

$$\boldsymbol{\Sigma}^{(s+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T}{n}. \tag{3.61}$$

Resumo do Algoritmo:

Apresentamos a seguir um resumo dos passos do algoritmo EM para as misturas finitas de normais

1. Atribuir os valores iniciais $\alpha_j^{(0)}$, $\boldsymbol{\mu}_j^{(0)}$, $\boldsymbol{\Sigma}_j^{(0)}$ ou $\boldsymbol{\Sigma}^{(0)}$. para os parâmetros da mistura;

2. Determinar as estimativas das probabilidades a posteriori $\tau_{ij}^{(s)}$, $s = 0, 1, 2, ...$, dadas por

$$\tau_{ij}^{(s)} = \frac{\alpha_j^{(s)}|\boldsymbol{\Sigma}_j^{(s)}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T\boldsymbol{\Sigma}_j^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})\}}{\sum_{t=1}^{k}\alpha_t^{(s)}|\boldsymbol{\Sigma}_t^{(s)}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})^T\boldsymbol{\Sigma}_t^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})\}}$$

ou, no caso homocedástico, por

$$\tau_{ij}^{(s)} = \frac{\alpha_j^{(s)}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T\boldsymbol{\Sigma}^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})\}}{\sum_{t=1}^{k}\alpha_t^{(s)}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})^T\boldsymbol{\Sigma}^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})\}},$$

para $i = 1, 2, 3..., n$ e $j = 1, 2, 3..., k$;

3. Atualizar as aproximações dos parâmetros, com $s = 1, 2, 3, ...$ e $j = 1, 2, 3, ..., k$, empregando

$$\alpha_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ij}^{(s)},$$

$$\boldsymbol{\mu}_j^{(s+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(s)}\mathbf{x}_i}{\sum_{i=1}^{n}\tau_{ij}^{(s)}},$$

$$\boldsymbol{\Sigma}_j^{(s+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(s)}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s+1)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s+1)})^T}{\sum_{i=1}^{n}\tau_{ij}^{(s)}}$$

ou, para o caso homcedástico,

$$\boldsymbol{\Sigma}^{(s+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_{ij}^{(s)}(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T}{n};$$

4. Repetir alternadamente os passos 2 e 3 até que um critério de convergência estabelecido seja atingido.

Da descrição do algoritmo acima temos duas questões: como determinar os valores iniciais em $\boldsymbol{\Phi}^{(0)}$ e qual o critério de parada das iterações a ser adotado.

Para a escolha dos valores iniciais algumas propostas aparecem na literatura. Uma proposta, por exemplo, é selecionar aleatoriamente k observações dentro do conjunto $\mathcal{A}_{(n)}$ para usá-los como $\boldsymbol{\mu}_j^{(0)}$, tomar $\boldsymbol{\Sigma}_j^{(0)}$ (ou $\boldsymbol{\Sigma}^{(0)}$) como sendo a matriz identidade e fazer $\alpha_j^{(0)} = \frac{1}{k}$ para todo j. Uma discussão mais aprofundada sobre essa questão, com várias referências sobre o assunto, é apresentada em McLachlan e Basford (1988, Seção 1.7) e em McLachlan e Peel (2000, Seção 2.12). Essa questão será abordada no próximo capítulo visando as aplicações de interesse neste trabalho.

Com respeito ao critério de convergência, um procedimento seria parar as iterações quando, em duas iterações sucessivas, a maior das diferenças entre os elementos que compõem as aproximações é menor que um valor fixado suficientemente pequeno. Em muitas aplicações, a diferença entre as estimativas sucessivas da função de log-verossimilhança é considerada como o critério de parada, ou seja, se $|L(\mathbf{\Phi}^{(s+1)}) - L(\mathbf{\Phi}^{(s)})| < \epsilon$, para um ϵ muito pequeno, por exemplo, da ordem de 10^{-6}. Em particular, a diferença entre as estimativas de $L(\mathbf{\Phi}^{(s)})$ é o critério que adotamos neste trabalho.

A seguir será considerada a determinação das aproximações para os parâmetros das distribuições marginais em misturas finitas de normais.

3.7.3 O EM para as Marginais em Mistura de Normais

Foi visto na Seção 3.2 que as distribuições marginais em uma mistura finita de normais, é também uma mistura finita de normais. Se $\mathbf{X} \sim p(\mathbf{x};\mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j f(\mathbf{x};\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)$ e $\mathbf{Y} = A\mathbf{X}$, com A da forma dada no Corolário 3.2.1, sabemos que $\mathbf{Y} \sim h(\mathbf{y};\mathbf{\Phi}) = \sum_{j=1}^{k} \alpha_j f(\mathbf{y};\boldsymbol{\mu}_{\mathbf{Y}j},\mathbf{\Sigma}_{\mathbf{Y}j})$, onde $\boldsymbol{\mu}_{\mathbf{Y}j} = A\boldsymbol{\mu}_j$ e $\mathbf{\Sigma}_{\mathbf{Y}j} = A\mathbf{\Sigma}_j A^T$.

Como a distribuição de $\mathbf{Y}$ é uma mistura de normais, temos que

$$\tau^{(s)}_{\mathbf{Y}ij} = \frac{\alpha_j^{(s)} |\mathbf{\Sigma}^{(s)}_{\mathbf{Y}j}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}^{(s)}_{\mathbf{Y}j})^T \mathbf{\Sigma}^{(s)-1}_{\mathbf{Y}j}(\mathbf{y}_i - \boldsymbol{\mu}^{(s)}_{\mathbf{Y}j})\}}{\sum_{t=1}^{k} \alpha_t^{(s)} |\mathbf{\Sigma}^{(s)}_{\mathbf{Y}t}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}^{(s)}_{\mathbf{Y}t})^T \mathbf{\Sigma}^{(s)-1}_{\mathbf{Y}t}(\mathbf{y}_i - \boldsymbol{\mu}^{(s)}_{\mathbf{Y}t})\}}, \tag{3.62}$$

onde $\mathbf{y}_i = A\mathbf{x}_i$, $i = 1, 2, 3, ..., n$.

Das definições, temos a igualdade

$$|\mathbf{\Sigma}_{\mathbf{Y}j}^{(s)}|^{-\frac{1}{2}} = |A\mathbf{\Sigma}_j^{(s)}A^T|^{-\frac{1}{2}} = |A||\mathbf{\Sigma}_j^{(s)}|^{-\frac{1}{2}} \tag{3.63}$$

e, também,

$$\begin{aligned}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{Y}j}^{(s)})^T\mathbf{\Sigma}_{\mathbf{Y}j}^{(s)-1}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{Y}j}^{(s)}) &= (\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T A^T (A^T)^{-1}\mathbf{\Sigma}_j^{(s)-1}A^{-1}A(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)}) \\ &= (\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T\mathbf{\Sigma}_j^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)}). \end{aligned} \tag{3.64}$$

Substituindo (3.63) e (3.64) em (3.62), chegamos a

$$\tau_{\mathbf{Y}ij}^{(s)} = \frac{|A|\alpha_j^{(s)}|\mathbf{\Sigma}_j^{(s)}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s)})^T\mathbf{\Sigma}_j^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{X}j}^{(s)})\}}{|A|\sum_{t=1}^k \alpha_t^{(s)}|\mathbf{\Sigma}_t^{(s)}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})^T\mathbf{\Sigma}_t^{(s)-1}(\mathbf{x}_i - \boldsymbol{\mu}_t^{(s)})\}} = \tau_{ij}^{(s)}. \tag{3.65}$$

Usando o resultado acima, temos que

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{Y}j}^{(s+1)} &= \frac{\sum_{i=1}^n \tau_{ij}^{(s)}\mathbf{y}_i}{\sum_{i=1}^n \tau_{ij}^{(s)}} = \frac{\sum_{i=1}^n \tau_{ij}^{(s)}A\mathbf{x}_i}{\sum_{i=1}^n \tau_{ij}^{(s)}} \\ &= A\{\frac{\sum_{i=1}^n \tau_{ij}^{(s)}\mathbf{x}_i}{\sum_{i=1}^n \tau_{ij}^{(s)}}\} \\ &= A\boldsymbol{\mu}_j^{(s+1)}, \quad j = 1, 2, 3, ..., k. \end{aligned} \tag{3.66}$$

Analogamente,

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{Y}j}^{(s+1)} &= \frac{\sum_{i=1}^n \tau_{ij}^{(s)}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{Y}j}^{(s+1)})(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{Y}j}^{(s+1)})^T}{\sum_{i=1}^n \tau_{ij}^{(s)}} \\ &= \frac{\sum_{i=1}^n \tau_{ij}^{(s)}A(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s+1)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(s+1)})^T A^T}{\sum_{i=1}^n \tau_{ij}^{(s)}} \\ &= A\{\frac{\sum_{i=1}^n \tau_{ij}^{(s)}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{Y}j}^{(s+1)})(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{Y}j}^{(s+1)})^T}{\sum_{i=1}^n \tau_{ij}^{(s)}}\}A^T \\ &= A\mathbf{\Sigma}_j^{(s+1)}A^T, \quad j = 1, 2, 3, ..., k. \end{aligned} \tag{3.67}$$

Dos desenvolvimentos acima, temos também que, no caso de mistura de normais homocedásticas, as aproximações para $\mathbf{\Sigma_Y} = A\mathbf{\Sigma}A^T$ serão dadas por

$$\mathbf{\Sigma_Y} = A\mathbf{\Sigma}^{(s+1)}A^T \tag{3.68}$$

Das expressões dadas em (3.66), (3.67) e (3.68), vemos que as aproximações para os parâmetros das distribuições marginais podem ser determinadas a partir das aproximações para as estimativas dos parâmetros da distribuição do vetor original. Esses resultados implicam que, para as misturas finitas de normais, não é necessário executar o algoritmo EM separadamente para as distribuições marginais, em cada iteração com a distribuição original, podem ser determinadas as aproximações para as marginais através das expressões acima.

Capítulo 4

Classificação com Misturas Finitas de Densidades

Neste capítulo, propomos uma versão empírica do classificador de Bayes em que modelamos as distribuições condicionais das classes através de misturas finitas de densidades. São discutidas suas propriedades e abordados alguns aspectos teóricos que justificam o procedimento proposto.

4.1 A Modelagem das Densidades Condicionais

Como mencionado em capítulos anteriores, neste trabalho estamos interessados na aprendizagem informativa para os problemas de RPS. Nessa abordagem, faz-se necessário esti-

mar a distribuição do vetor de características em cada uma das classes definidas para o problema, as distribuições condicionais das classes (veja as Seções 1.4 e 2.1). Em muitas aplicações, não há qualquer justificativa para o emprego de distribuições paramétricas para modelar essas distribuições condicionais. Para esses casos, como discutido na Seção 2.3, podemos utilizar os métodos não-paramétricos. A nossa proposta, no entanto, é utilizar as misturas finitas de densidades, em particular *misturas finitas de normais*, para estimar as distribuições das classes e empregar essas estimativas para obter uma versão empírica da regra de Bayes.

Foi comentado no Capítulo 1, Seção 1.5, que os modelos de misturas finitas têm sido muito empregados na abordagem baseada em modelos para os problemas de RPNS, onde a suposição é que cada observação é proveniente de uma distribuição dentre um conjunto de distribuições distintas. Nesse contexto, emprega-se um modelo de mistura finita, onde cada componente modela a distribuição de um grupo e os coeficientes são as probabilidades de a observação provir do respectivo grupo. Foi acrescentado, no entanto, que as misturas finitas constituem uma classe de modelos com flexibilidade suficiente para modelar distribuições desconhecidas arbitrariamente complexas (Jain *et al.* (2000)). Com essa capacidade de modelagem, as misturas finitas são convenientes para modelar as distribuições condicionais desconhecidas nos problemas de RPS (Ripley (1996)).

Algumas vantagens do emprego das misturas finitas na estimação das densidades condicionais são imediatas. Com relação a estimação paramétrica, as misturas não restrigem as densidades condicionais a uma forma funcional rígida, no sentido de que as misturas finitas geram uma classe muito ampla de formas funcionais. Para o emprego de um fixado modelo de mistura finita, não é necessário reter todas as observações, como em muitos

métodos não-paramétricos, mas apenas as estimativas dos parâmetros do modelo, ou seja, as misturas *reduzem* os dados. De certa forma, as misturas finitas combinam a vantagem de menor complexidade dos modelos paramétricos com a vantagem da capacidade de modelagem dos métodos não-paramétricos.

Na decisão para empregar as misturas finitas de normais, foi levada em consideração, principalmente, a capacidade de modelagem da família de distribuições normais e sua simplicidade com respeito a estimação dos parâmetros. Em relação a essa segunda questão, foi visto no Capítulo 3 que as dificuldades inerentes a estimação dos parâmetro para as misturas finitas são bastantes simplificadas quando as componentes são densidades normais. Com essas considerações e a capacidade de modelagem das misturas em geral, nós acreditamos que empregando misturas finitas de normais é possível modelar as distribuições condicionais de forma bastante satisfatória, tanto no que diz respeito a eficiência da modelagem como em relação às questões de implementação computacional.

Sendo feita uma modelagem adequada das distribuições condicionais e com boas estimativas das probabilidades a priori das classes, espera-se obter uma versão empírica da regra de Bayes que, em algum sentido, aproxime o erro de Bayes. Essas considerações levam a questão da *consistência* da regra empírica obtida, o que será discutido na Seção 4.5.

Uma das dificuldades para estimar densidades em espaços de dimensão alta é que, em conjuntos de observações de tamanho moderado, ocorre a ausência de pontos correspondentes a extensas regiões nesses espaços. Esse comportamento para os dados amostrais, é reflexo da grande dispersão das distribuições em dimensões altas, fenômeno denominado *a maldição da dimensionalidade* (veja as discussões em (Silverman (1986, Capítulo 4) e em Scott (1992, Capítulo 7)). Uma das consequências desse fenômeno, é a necessidade

de amostras extremamente grandes para estimar essas distribuições. Na prática, com amostras de tamanho moderado, temos a questão da localização das componentes da mistura no espaço das características visando amenizar esse problema de representatividade no conjunto de dados. Fixado o número de componentes para a mistura, nós aplicamos um método de agrupamento buscando identificar as regiões no espaço das características que necessitam ser cobertas pelas componentes da mistura e, em cada uma das regiões definidas pelo agrupamento, é colocada uma componente da mistura. Com esse procedimento, torna-se crucial a seleção de um número de componentes capazes de descrever de forma adequada as observações, questão essa que será discutida na Seção 4.3.

Em McLachlan e Peel (2000, Seção 1.1), é observado que os modelos de mistura podem modelar distribuições bastantes complexas através de uma escolha apropriada de suas componentes para representar, de forma precisa, as regiões locais do suporte da verdadeira distribuição. O agrupamento tem esse objetivo, detectar as regiões críticas do espaço das características que sejam importantes para a definição do modelo. As caudas das distribuições condicionais, por exemplo, assumem um papel importante pois, em teoria, são as regiões onde ocorrem as interseções dessas distribuições e, em virtude disso, devem ser bem estimadas. Nós acreditamos que com o procedimento proposto, é possivel estimar de forma eficiente a distribuição em toda sua extensão.

Na determinação das estimativas dos parâmetros da mistura, empregamos o algoritmo EM, da forma discutida na Seção 3.7. No emprego do EM, uma questão crucial é como determinar os valores iniciais para os parâmetros no processo de iteração do algoritmo. Para esses valores iniciais, nós utilizamos as estimativas de Máxima Verossimilhança dos parâmetros obtidas a partir das observações em cada um dos grupos formados no processo

de agrupamento. Com esse procedimento, espera-se que os valores iniciais fornecidos dessa forma possam contribuir para acelerar a convergência do algoritmo.

O método *k-Means* é empregado para efetuar o agrupamento. Essa escolha se deve a conhecida eficiência desse método (Ripley (1996, Capítulo 9)) e ao fato de que suas propriedades já são bastante discutidas na literatura (veja, por exemplo, Pollard (1981)).

4.2 A Estimação da Regra de Bayes

De acordo com a abordagem proposta, considerando um problema com M classes, devemos estimar as densidades condicionais $f_j(\cdot)$, as probabilidades a priori $P(\mathcal{C}_j)$ e as probabilidades a posteriori $P(\mathcal{C}_j|\mathbf{x})$, $j = 1, 2, 3, ..., M$ (veja Seção 2.1). Para esse fim, o conjunto de treinamento é dividido em subconjuntos $\mathcal{T}_{(n_j)} = \{\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \mathbf{x}_{j,3}, ..., \mathbf{x}_{j,n_j}\}$, sendo as observações em cada um deles utilizadas separadamente para estimar suas correspondentes distribuições. Pela nossa proposta, para $f_j(\cdot)$, $P(\mathcal{C}_j)$ e $P(\mathcal{C}_j|\mathbf{x})$ temos, respectivamente, os seguintes estimadores

$$\widehat{f}_j(\mathbf{x}) = f_j(\mathbf{x}; \widehat{\boldsymbol{\theta}}_j) = \sum_{l=1}^{k_j} \widehat{\alpha}_{jl}\, g(\mathbf{x}; \widehat{\boldsymbol{\mu}}_{jl}, \widehat{\boldsymbol{\Sigma}}_{jl})\,, \tag{4.1}$$

$$\widehat{P}(\mathcal{C}_j) = \frac{n_j}{n}\,, \tag{4.2}$$

$$\widehat{P}(\mathcal{C}_j|\mathbf{x}) = \frac{\widehat{f}_j(\mathbf{x})\widehat{P}(\mathcal{C}_j)}{\sum_{t=1}^{M} \widehat{f}_t(\mathbf{x})\widehat{P}(\mathcal{C}_t)}\,, \tag{4.3}$$

onde $g(\cdot; \widehat{\boldsymbol{\mu}}_{jl}, \widehat{\boldsymbol{\Sigma}}_{jl})$ são as densidades $N(\widehat{\boldsymbol{\mu}}_{jl}, \widehat{\boldsymbol{\Sigma}}_{jl})$. Esses estimadores são empregados para

determinar uma versão empírica da regra de Bayes, dada por

$$r_{(n)}^{(mf)}(\mathbf{x}) = t \quad \text{se} \quad \widehat{P}(\mathcal{C}_t|\mathbf{x}) = \max_j \widehat{P}(\mathcal{C}_j|\mathbf{x}), \tag{4.4}$$

onde, no caso de o máximo ocorrer para mais de uma das classes, o objeto com observação $\mathbf{x}$ é classificado em qualquer uma das classes empatadas (veja Seção 2.1). Na implementação da regra, no caso de m classes empatadas, o objeto é alocado aleatoriamente a qualquer uma dessas classes com probabilidade $\frac{1}{m}$.

Para o emprego da regra em (4.4), é necessário obter as estimativas dos vetores de parâmetros $\widehat{\boldsymbol{\theta}}_j = (\widehat{\alpha}_1^j, \widehat{\alpha}_2^j, ..., \widehat{\alpha}_{k_j}^j, \widehat{\boldsymbol{\mu}}_1^j, \widehat{\boldsymbol{\mu}}_2^j, ..., \widehat{\boldsymbol{\mu}}_{k_j}^j, \widehat{\boldsymbol{\Sigma}}_1^j, \widehat{\boldsymbol{\Sigma}}_2^j, ..., \widehat{\boldsymbol{\Sigma}}_{k_j}^j)$, e selecionar as dimensões k_j, $j = 1, 2, 3, ..., M$. Fixado k_j, as estimativas $\widehat{\boldsymbol{\theta}}_j$ são obtidas empregando-se o algoritmo EM como mencionado na Seção 4.2, utilizando-se as observações do conjunto de treinamento $\mathcal{T}_j$ (veja também a Seção 3.7).

Com relação a k_j, a proposta é selecionar a dimensão do modelo para cada classe através do emprego do Critério de Informação de Akaike (AIC) e do Critério de Informação Bayesiano (BIC). Na seção a seguir essa questão será abordada.

4.3 A Seleção dos Números de Componentes

Como mencionado, uma questão crucial para o emprego das misturas finitas para modelar distribuições desconhecidas, é determinar a dimensão do modelo adequada aos dados. Um dos obstáculos para isso, é que não dispomos de um procedimento estatístico completamente satisfatório para determinar essa dimensão (Roeder (1992), Polymenis e Tit-

terington (1998)). Considerando essa dificuldade, nós propomos determinar a dimensão direcionada pelas próprias observações nos conjuntos de treinamentos $\mathcal{T}_j$, efetivando isso através da otimização das funções-critério estabelecidas pelo AIC e BIC (veja a Seção 3.5 para a discussão desses critérios).

Os critérios AIC e BIC foram desenvolvidos sob condições de regularidades que não são satisfeitas pelos modelos de misturas finitas. A motivação para empregá-los em nossa abordagem, no entanto, é o fato de que resultados na literatura indicam desempenhos satisfatórios desses critérios em selecionar a dimensão para misturas finitas. Em estimação bayesiana de densidades univariadas, Roeder e Wasserman (1997) usam uma mistura finita de normais, estimando a dimensão com o BIC e obtêm uma a posteriori consistente. Solka, Wegman, Priebe, Poston e Rogers (1998) empregam um modelo de mistura adaptativa para estimar densidades e concluem que o AIC é conveniente para avaliar a complexidade do modelo. Leroux (1992) mostrou que, no caso em que a distribuição a ser estimada é uma mistura, selecionando a dimensão do modelo estimado baseada no AIC ou BIC, assintoticamente, essa dimensão não será menor que a verdadeira dimensão (veja a Seção 4.5).

Na prática, quando utilizados para selecionar a dimensão de misturas finitas, os critérios AIC e BIC apresentam tendências divergentes: o AIC tende a selecionar modelos de dimensão maior do que aqueles selecionados pelo BIC (veja Seção 3.5). Em situações onde ocorra essa divergência, uma alternativa é adotar um modelo dado por uma ponderação entre os modelos selecionados com o emprego desses critérios, ou seja,

$$\widehat{f}_j(\mathbf{x}) = \frac{1}{2}\{\widehat{f}_j^{(A)}(\mathbf{x}) + \widehat{f}_j^{(B)}(\mathbf{x})\}, \tag{4.5}$$

onde $\widehat{f}_j^{(A)}(\cdot)$ e $\widehat{f}_j^{(B)}(\cdot)$ são os modelos de mistura finita de normais com suas dimensões determinadas, respectivamente, pelos critérios AIC e BIC.

Algumas questões podem ser consideradas com relação a proposta em (4.5). Quando se trata de estimação de densidades, não existe um estimador não-tendencioso (veja Definição A.2.1) que satisfaça razoáveis condições de regularidade (veja a discussão em Scott (1992, Seção 2.3)). Com essa observação e considerando o fato de que a estimativa $\widehat{f}_j(\mathbf{x})$ é uma variável aleatória, a precisão da estimação da densidade pode ser mensurada pontualmente através do *Erro Quadrático Médio* (EQM), ou seja,

$$\begin{aligned} EQM\{\widehat{f}_j(\mathbf{x})\} &\stackrel{\text{def}}{=} E\{\widehat{f}_j(\mathbf{x}) - f_j(\mathbf{x})\}^2 \\ &= E\{\widehat{f}_j(\mathbf{x}) - E[\widehat{f}_j(\mathbf{x})]\}^2 + \{E[\widehat{f}_j(\mathbf{x})] - f_j(\mathbf{x})\}^2. \end{aligned} \tag{4.6}$$

É importante observar que as esperanças E e as estimativas $\widehat{f}_j(\mathbf{x})$ em (4.6), são condicionadas ao conjunto de treinamento $\mathcal{T}_j$, ou seja, são realizações para um particular conjunto $\mathcal{T}_j$. As notações $E_{\mathcal{T}_j}$ e $\widehat{f}_j(\mathbf{x}|\mathcal{T}_j)$ não foram empregadas visando simplificar a apresentação.

No segundo membro em (4.6), o primeiro termo é a variância e o segundo o viés ao quadrado de $\widehat{f}_j(\mathbf{x})$, que denotaremos, respectivamente, por $Var[\widehat{f}_j(\mathbf{x})]$ e $v[\widehat{f}_j(\mathbf{x})]$. A variância reflete a sensiblidade da estimativa $\widehat{f}_j(\mathbf{x})$ com relação ao conjunto de treinamento; uma menor sensibilidade significa maior estabilidade da estimativa contra as variações amostrais. O viés reflete a sensibilidade com relação a f_j; ele representa a precisão com que, em média, a estimativa aproxima o valor de $f_j(\mathbf{x})$. Vemos, então, que o desejável é que a variância e o quadrado do viés sejam pequenos, uma vez que ambos contribuem com o mesmo peso para o EQM.

Usando a forma proposta em (4.5) para $\widehat{f}_j(\mathbf{x})$, obtemos

$$Var[\widehat{f}_j(\mathbf{x})] = \frac{1}{4}\{Var[\widehat{f}_j^{(A)}(\mathbf{x})] + Var[\widehat{f}_j^{(B)}(\mathbf{x})]\} + \frac{2}{4}Cov[\widehat{f}_j^{(A)}(\mathbf{x}), \widehat{f}_j^{(B)}(\mathbf{x})]. \quad (4.7)$$

e

$$v[\widehat{f}_j(\mathbf{x})] = \frac{1}{2}\{v[\widehat{f}_j^{(A)}(\mathbf{x})] + v[\widehat{f}_j^{(B)}(\mathbf{x})]\}. \quad (4.8)$$

Analisando (4.8), temos que

$$\min\left(v[\widehat{f}_j^{(A)}(\mathbf{x})], v[\widehat{f}_j^{(B)}(\mathbf{x})]\right) \leq v[\widehat{f}_j(\mathbf{x})] \leq \max\left(v[\widehat{f}_j^{(A)}(\mathbf{x})], v[\widehat{f}_j^{(B)}(\mathbf{x})]\right). \quad (4.9)$$

Se $v[\widehat{f}_j^{(A)}(\mathbf{x})] = v[\widehat{f}_j^{(B)}(\mathbf{x})]$, então $v[\widehat{f}_j(\mathbf{x})]$ será igual a esse valor. Se por um lado o $v[\widehat{f}_j(\mathbf{x})]$ nunca será maior que o "pior caso", por outro lado, nunca será inferior ao "melhor caso". Dessa forma, o procedimento de ponderar as estimativas é conservativo com relação ao viés, no sentido de que terá um viés que não será superior ao maior dos vieses correspondentes as estimativas $\widehat{f}_j^{(A)}(\mathbf{x})$ e $\widehat{f}_j^{(B)}(\mathbf{x})$.

Com respeito a variância, de (4.7), podemos escrever,

$$Var[\widehat{f}_j(\mathbf{x})] = \frac{1}{4}\{Var[\widehat{f}_j^{(A)}(\mathbf{x})] + Var[\widehat{f}_j^{(B)}(\mathbf{x})]\} + \frac{2}{4}\rho\sqrt{Var[\widehat{f}_j^{(A)}(\mathbf{x})]Var[\widehat{f}_j^{(B)}(\mathbf{x})]}, \quad (4.10)$$

onde ρ é o coeficiente de correlação entre $\widehat{f}_j^{(A)}(\mathbf{x})$ e $\widehat{f}_j^{(B)}(\mathbf{x})$. Analisando (4.10), temos que $Var[\widehat{f}_j(\mathbf{x})]$ será menor ou igual a maior das variâncias correspondentes as estimativas individuais, $\widehat{f}_j^{(A)}(\mathbf{x})$ e $\widehat{f}_j^{(B)}(\mathbf{x})$, qualquer que seja o valor de ρ. Para ver isso, considere que $m = \max\left(v[\widehat{f}_j^{(A)}(\mathbf{x})], v[\widehat{f}_j^{(B)}(\mathbf{x})]\right)$, então

$$Var[\widehat{f}_j(\mathbf{x})] \leq \frac{1}{4}\{m + m\} + \frac{2}{4}\rho\sqrt{m^2} = \frac{2}{4}(1+\rho)m. \quad (4.11)$$

Como o valor máximo do último termo em (4.11) ocorre para $\rho = 1$, temos, portanto,

$$Var[\widehat{f}_j(\mathbf{x})] \leq \max\left(v[\widehat{f}_j^{(A)}(\mathbf{x})], v[\widehat{f}_j^{(B)}(\mathbf{x})]\right). \tag{4.12}$$

Das considerações acima com respeito a (4.7) e (4.8), temos que a ponderação proposta em (4.5) pode levar a uma redução no EQM das estimativas com respeito as estimativas individuais dadas por $\widehat{f}_j^{(A)}(\cdot)$ e $\widehat{f}_j^{(B)}(\cdot)$.

Na seção a seguir discutiremos alguns aspectos relativos a *consistência* da regra de classificação proposta neste capítulo.

4.4 Consistência da Regra de Classificação

Nesta seção, $\widehat{P}_j$ denotará as probabilidades a priori estimadas das classes. Ressaltamos que todos os estimadores são funções dos tamanhos dos respectivos conjuntos de treinamento, ou seja, $\widehat{P}_j = \widehat{P}_{j(n_j)}$ e $\widehat{f}_j(\mathbf{x}) = \widehat{f}_{j(n_j)}(\mathbf{x})$, sendo os índices n_j omitidos para simplificar a notação.

O objetivo nesta seção é discutir se o erro de classificação da regra proposta em (4.4), denotado por $e_{(n)}^{(mf)}$, se aproxima, em alguma sentido, do erro de Bayes e^*. Isso é uma maneira de avaliar a regra pois, como discutido na Seção 2.1, e^* é o menor valor para a probabilidade do erro de classificação que pode ser atingido por qualquer classificador.

Já foi comentado que, em geral, nós não esperamos que uma regra empírica atinja exatamente o erro de Bayes, mas seria desejável obter um erro de classificação para essa regra

que estivesse próximo de e^* com alta probabilidade. Essa idéia é formulada na definição de *consistência* a seguir.

Definição 4.4.1 *Uma regra de classificação, com erro de classificação* $e_{(n)}$*, é* consistente*, ou* consistente para o risco de Bayes*, se* $\forall\, \epsilon > 0$ *temos*

$$\lim_{n\to\infty} Pr(e_{(n)} - e^* > \epsilon) = 0.$$

Pela Definição 4.4.1, consistência significa a convergência em probabilidade de $e_{(n)}$ para e^*. Para uma regra consistente, portanto, fica garantido que, aumentando a quantidade de observações no conjunto de treinamento, a probabilidade de que o erro de classificação esteja dentro de uma distância muito pequena do valor ótimo será arbitrariamente próxima de 1. Intuitivamente, temos uma probabilidade alta de que a regra "aprenda" a decisão ótima a partir de um conjunto de treinamento suficientemente grande.

Um primeiro resultado, fundamental para a nossa discussão, é estabelecido no Teorema 4.4.1 a seguir.

Teorema 4.4.1 *Seja* $e_{(n)}^{(mf)}$ *o erro de classificação da regra* $r_{(n)}^{(mf)}$ *em (4.4). Sendo* e^* *o erro de Bayes, então, temos que*

$$0 \leq e_{(n)}^{(mf)} - e^* \leq \sum_{j=1}^{M} \int |P_j f_j(\mathbf{x}) - \widehat{P}_j \widehat{f}_j(\mathbf{x})|\, d\nu.$$

Prova: Em Devroye e Györfi (1985, Capítulo 10, Teorema 1) é provado que a desigualdade do tipo acima se verifica para qualquer regra que, empregando as densidades condicionais

e as probabilidades a priori estimadas, seleciona a classe com a maior probabilidade a posteriori estimada. Em particular, se aplica à regra $r_{(n)}^{(mf)}$. □

Do Teorema 4.4.1, temos que a diferença entre o erro de classificação de $r_{(n)}^{(mf)}$ e o erro de Bayes é dominada pela soma das integrais de $|P_j f_j(\mathbf{x}) - \widehat{P}_j \widehat{f}_j(\mathbf{x})|$. Desse resultado, vemos que, como M é fixo, para que a regra seja consistente, basta que tenhamos as convergências $\widehat{P}_j \widehat{f}_j(\mathbf{x}) \to P_j f_j(\mathbf{x})$ em norma $L_1(d\nu)$ para cada j.

Em Priebe (1994), é observado que, no emprego de modelos de mistura finita para estimar densidades, são necessárias algumas suposições bastantes restritivas para se obter consistência no procedimento de estimação, em particular, a suposição de que a densidade a ser estimada seja uma mistura do mesmo tipo da empregada no estimador. Também é observado, no entanto, que permitindo o número de componentes do estimador crescer de forma adequada com relação ao número de observações, a suposição da verdadeira densidade ser uma mistura pode ser relaxada. Essa questão é abordada a seguir.

Na apresentação dos teoremas a seguir, o índice j é omitido, uma vez que diz respeito as estimativas para uma classe genérica.

O Teorema 4.4.2 a seguir, estabelece as condições para a convergência em norma L_1 para a estimação de densidades através de modelos de mistura. Para esse fim, seja $\mathcal{C}$ a classe das densidades contínuas e limitadas, $\mathcal{C}_0$ o conjunto das funções que se anulam em ∞, a entropia definida por $H(\alpha, \beta) = \int \alpha(x) \ln \beta(x)\, dx$ e considere a seguinte classe de densidades

$$\mathcal{F} = \Big\{ f \in \mathcal{C} \Big| \, f \in \mathcal{C}_0 \quad \text{e} \quad H(f, f) < \infty \Big\}.$$

Teorema 4.4.2 *Seja $f \in \mathcal{F}$ e considere um conjunto de treinamento $\mathcal{T}_{(n)}$ com observações de f. Seja $\widehat{f}_{(n)}$ o estimador de densidades definido por uma mistura de densidades normais com $k_{(n)}$ componentes baseado em $\mathcal{T}_{(n)}$. Então, quando $n \to \infty$ e $k_{(n)} \to \infty$, com $k_{(n)}/n \to 0$, temos que $\widehat{f}_{(n)} \to f$ em norma L_1 ($\widehat{f}_{(n)} \xrightarrow{L_1} f$) .*

Prova: Priebe (1993) e Priebe (1994). □

O Teorema 4.4.2 é provado em Priebe (1994) para uma mistura adaptativa de normais, onde as componentes vão sendo acrescentadas ao modelo recursivamente de acordo com novas observações que chegam ao conjunto de treinamento. A importância desse teorema aqui, no entanto, é garantir que podemos obter uma sequência de estimadores $\{\widehat{f}_{(n)}\}$ convergente em L_1 para f, desde que a sequência $\{k_{(n)}\}$ seja "adequadamente" escolhida.

Teorema 4.4.3 *Se $\widehat{f}_{(n)} \xrightarrow{L_1} f$, então $\widehat{P}_{(n)}\widehat{f}_{(n)} \xrightarrow{L_1} Pf$.*

Prova: Podemos escrever

$$|\widehat{P}_{(n)}\widehat{f}_{(n)} - Pf| = |\widehat{P}_{(n)}\widehat{f}_{(n)} - P\widehat{f}_{(n)} + P\widehat{f}_{(n)} - Pf| \leq |\widehat{P}_{(n)} - P||\widehat{f}_{(n)}| + |\widehat{P}_{(n)}||\widehat{f}_{(n)} - f|$$

Integrando e utilizando a desigualdade acima, temos

$$\int |\widehat{P}_{(n)}\widehat{f}_{(n)} - Pf|\, d\nu \leq |\widehat{P}_{(n)} - P| \int |\widehat{f}_{(n)}|\, d\nu + |\widehat{P}_{(n)}| \int |\widehat{f}_{(n)} - f|\, d\nu.$$

Então, como

$$\int |\widehat{f}_{(n)} - f|\, d\nu \xrightarrow{n\to\infty} 0$$

e que, pela Lei Forte dos Grandes Números, $\widehat{P}_{(n)} \to P$, temos, portanto,

$$\int |\widehat{P}_{(n)}\widehat{f}_{(n)} - Pf|\, d\nu \xrightarrow{n\to\infty} 0$$

□

Usando os resultados nos Teoremas 4.4.2 e 4.4.3 no Teorema 4.4.1, vemos que, se o número de componentes para misturas modelando cada uma das classes for selecionado de forma adequada ao crescimento de n, a regra em (4.4) será consistente.

Suponha agora que $f_{1(n)}, f_{2(n)} \xrightarrow{L_1} f$ e $g_{(n)} = \frac{1}{2}(f_{1(n)} + f_{2(n)})$. Podemos escrever

$$\begin{aligned} |g_{(n)} - f| &= |\frac{1}{2}f_{1(n)} + \frac{1}{2}f_{2(n)} - f| \\ &= |\frac{1}{2}(f_{1(n)} - f) + \frac{1}{2}(f_{2(n)} - f)| \\ &= \frac{1}{2}|(f_{1(n)} - f) + (f_{2(n)} - f)| \\ &\leq \frac{1}{2}|f_{1(n)} - f| + \frac{1}{2}|f_{2(n)} - f| \end{aligned}$$

e, por isso, temos que

$$\int |g_{(n)} - f|\, d\nu \leq \frac{1}{2}\int |f_{1(n)} - f|\, d\nu + \frac{1}{2}\int |f_{2(n)} - f|\, d\nu,$$

de onde concluimos que

$$\int |g_{(n)} - f|\, d\nu \xrightarrow{n\to\infty} 0. \tag{4.13}$$

Empregando o resultado em (4.13), vemos que, sob as condições de convergência do estimador com modelos de mistura, sendo empregada a ponderação proposta em (4.5), a regra em (4.4) ainda será consistente.

A questão agora, é discutir a forma empregada para selecionar a dimensão das misturas em cada uma das classes. No Capitulo 3, foi mencionado que, para uma família de densidades $\{g(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, o modelo de mistura geral é da forma,

$$f(\mathbf{x}) = \int_{\Theta} g(\mathbf{x};\boldsymbol{\theta})\, dG(\boldsymbol{\theta}), \tag{4.14}$$

onde G é a distribuição de mistura.

A estimação de Máxima Verossimilhança pode ser empregada para obter-se uma estimativa para a distribuição de mistura, uma vez que é provado a existência de um estimador de máxima verossimilhança $\widehat{G}$ para G (Lindsay (1983)). É observado em Leroux (1992), no entanto, que para uma mistura finita com k componentes, embora $\widehat{G}$ forneça uma estimativa de k, ela pode incluir mais componentes que o necessário para um bom ajuste dos dados. Em virtude disso, um procedimento que "penalize" esse ajuste excessivo seria mais adequado que a estimação de máxima verossimilhança sem restrição, pois a eliminação de componentes desnecessárias pode levar a uma estimação mais precisa dos parâmetros na mistura finita.

Em Leroux (1992) é proposto selecionar o valor $\widehat{k}_{(n)}$ para k de forma a maximizar

$$L(\widehat{\mathbf{\Phi}}_{(k_{(n)})}) - a_{k_{(n)}}, \tag{4.15}$$

onde $L(\widehat{\mathbf{\Phi}}_{(k_{(n)})})$ é a função de verossimilhança para a mistura finita com $k_{(n)}$ componentes e $a_{k_{(n)}}$ é um termo de penalidade, satisfazendo $a_{k_{(n)}} \geq a_{k_{(n)}+1}$, que desencoraja a seleção de um modelo com um número excessivo de componentes. Vemos, então, que o procedimento adotado aqui é da forma da proposta em (4.15). No nosso caso, para o AIC temos $a_{k_{(n)}} = dim(\widehat{\mathbf{\Phi}}_{(k_{(n)})})$ e, para o BIC, $a_{k_{(n)}} = \frac{1}{2}\ln(n) dim(\widehat{\mathbf{\Phi}}_{(k_{(n)})})$, onde $dim(\widehat{\mathbf{\Phi}}_{(k_{(n)})}) = (k_{(n)} - 1) + k_{(n)}[d + \frac{d(d+1)}{2}]$, sendo d a dimensão do vetor de características.

Um procedimento de estimação como o proposto em (4.15) é denominado *estimação de máxima verossimilhança penalizada.*

O Teorema 4.4.4 a seguir, estabele que, sob algumas condições, sendo G^* a verdadeira

distribuição de mistura, o emprego do procedimento proposto em (4.15) leva a seleção de uma distribuição $\widehat{G}_{k_{(n)}}$ consistente para G^*, no sentido de que $\widehat{G}_{k_{(n)}}(\boldsymbol{\theta}) \to G^*(\boldsymbol{\theta})$ para todo $\boldsymbol{\theta}$ ponto de continuidade de G^* quando $n \to \infty$, ou seja, *converge em distribuição.*

Teorema 4.4.4 *Seja f da forma dada em (4.14), identificável e que satisfaz algumas condições de regularidade. Seja G^* a distribuição de mistura de f com k^* componentes ($k^* = \infty$ se G^* não é uma distribuição finita). Se, para cada $k_{(n)} < k^*$, $a_{k_{(n)}} \geq a_{k_{(n)}+1}$ para todo n e $\limsup_n a_{k_{(n)}}/n = 0$ com probabilidade 1, então $\liminf_{n\to\infty} \widehat{k}_{(n)} \geq k^*$ ($\widehat{k}_{(n)} \to \infty$ se $k^* = \infty$) e $\widehat{G}_{k_{(n)}} \to G^*$ em distribuição quando $n \to \infty$.*

Prova: Leroux (1992, Teorema 4). □

As condições de regularidade mencionadas no Teorema 4.4.4 são estabelecidas em Leroux (1992) e, em particular, essas condições são satisfeitas pelas misturas finitas de normais. A importância do resultado estabelecida nesse teorema, é garantir que o estimador $\widehat{k}_{(n)}$ obtido maximizando (4.15), no limite, não subestima o número de componentes da distribuição verdadeira, inclusive, no caso da mistura não ser finita, em que $\widehat{k}_{(n)} \to \infty$.

Dos resultados apresentados nesta seção, concluimos que é possível obter estimativas de densidades consistentes através do emprego de misturas finitas de normais como estimador e o procedimento adotado para selecionar as dimensões dos modelos. Em consequência, a regra de classificação proposta será consistente. As condições sob as quais se obtém a convergências, a princípio, podem parecer muito restritivas, no entanto, elas são satisfeitas por uma classe muito ampla de distribuições. Os modelos de mistura, por exemplo, geram uma classe muita rica em distribuições, o que pode ser visto analisando (4.14):

se G é discreta com suporte finito, então, $f(\cdot)$ é uma mistura finita de densidades; se G é degenerada com probabilidade 1 para algum $\boldsymbol{\theta}$, temos que $f(\cdot)$ é uma distribuição paramétrica ordinária, ou seja, a classe dos modelos de mistura contém as diversas distribuições paramétricas conhecidas.

4.5 A Implementação do Classificador

Nesta seção, discutimos alguns aspectos práticos para a implementação da regra de classificação. Apresentamos, a seguir, os passos necessários para essa implementação.

1. Dividir o conjunto de treinamento $\mathcal{T}_{(n)}$ nos subconjuntos $\mathcal{T}_{(n_j)}$, $j = 1, 2, 3, ..., M$;

2. Para cada conjunto $\mathcal{T}_{(n_j)}$, empregar as observações para ajustar os modelos

$$f_j^{(A)}(\mathbf{x}) = \sum_{l=1}^{k_j^{(A)}} \widehat{\alpha}_{jl}^{(A)}\, g(\mathbf{x}; \widehat{\boldsymbol{\mu}}_{jl}^{(A)}, \widehat{\boldsymbol{\Sigma}}_{jl}^{(A)}) \quad \text{e} \quad f_j^{(B)}(\mathbf{x}) = \sum_{l=1}^{k_j^{(B)}} \widehat{\alpha}_{jl}^{(B)}\, g(\mathbf{x}; \widehat{\boldsymbol{\mu}}_{jl}^{(B)}, \widehat{\boldsymbol{\Sigma}}_{jl}^{(B)}),$$

onde $g(\cdot; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ é a densidade $N(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$. As dimensões $k_j^{(A)}$ e $k_j^{(B)}$ são selecionadas maximizando, respectivamente,

$$AIC(k_j^{(A)}) = 2L(\widehat{\boldsymbol{\Phi}}_{(k_j^{(A)})}) - 2\{(k_j^{(A)} - 1) + k_j^{(A)}[d + \frac{d(d+1)}{2}]\}$$

e

$$BIC(k_j^{(B)}) = 2L(\widehat{\boldsymbol{\Phi}}_{(k_j^{(B)})}) - \{(k_j^{(B)} - 1) + k_j^{(B)}[d + \frac{d(d+1)}{2}]\} \ln(n_j),$$

onde $L(\cdot)$ são as funções de log-verossimilhança e d a dimensão do vetor de características. A estimação dos parâmetros é feita através do algoritmo EM;

3. Se $k_j^{(A)} = k_j^{(B)} = k_j$, então a densidade estimada para classe j é dada por

$$f_j(\mathbf{x}) = \sum_{l=1}^{k_j} \widehat{\alpha}_{jl}\, g(\mathbf{x}; \widehat{\boldsymbol{\mu}}_{jl}, \widehat{\boldsymbol{\Sigma}}_{jl}),$$

   e, no caso de $k_j^{(A)} \neq k_j^{(B)}$, por

$$f_j(\mathbf{x}) = \frac{1}{2}\{f_j^{(A)}(\mathbf{x}) + f_j^{(B)}(\mathbf{x})\};$$

4. Para cada $\mathbf{x}_0$ a ser classificado, determinar as probabilidades a posteriori

$$\widehat{P}(\mathcal{C}_j|\mathbf{x}_0) = \frac{\widehat{f}_j(\mathbf{x}_0)\widehat{P}(\mathcal{C}_j)}{\sum_{t=1}^{M} \widehat{f}_t(\mathbf{x}_0)\widehat{P}(\mathcal{C}_t)},$$

   onde $\widehat{P}(\mathcal{C}_j) = \frac{n_j}{n}$, $j = 1, 2, 3, ..., M$;

5. Classificar $\mathbf{x}_0$ na classe $\mathcal{C}_t$ para a qual $\widehat{P}(\mathcal{C}_t|\mathbf{x}_0) = \max_j \widehat{P}(\mathcal{C}_j|\mathbf{x}_0)$. Se m classes atingirem o máximo, atribuir a probabilidade de $\frac{1}{m}$ para cada uma dessas classes e selecionar uma delas segundo essas probabilidades.

Discutiremos agora, alguns aspectos relativos a implementação do passo 2 acima, com respeito a inicialização do algoritmo EM para estimar os parâmetros em cada um dos modelos. Para cada $k_j^{(A)}$ e $k_j^{(B)}$ testados, é feito um agrupamento com esses números de grupos para as observações em $\mathcal{T}_{(n_j)}$ empregando-se o algoritmo k-Means. Com as observações em cada um dos grupos formados em cada um dos $\mathcal{T}_{(n_j)}$, são determinadas as estimativas de máxima verossimilhança para os parâmetros $\widehat{\boldsymbol{\mu}}_{jl}^{(A)}$, $\widehat{\boldsymbol{\Sigma}}_{jl}^{(A)}$), $\widehat{\boldsymbol{\mu}}_{jl}^{(B)}$ e $\widehat{\boldsymbol{\Sigma}}_{jl}^{(B)}$, sendo os coeficientes da mistura, $\widehat{\alpha}_{jl}^{(A)}$ e $\widehat{\alpha}_{jl}^{(B)}$, estimados pelas proporções de observações nos grupos correspondentes. Essas estimativas são utilizadas como os valores iniciais para o EM.

Mencionamos no Capítulo 3, que uma das dificuldades com o emprego do algoritmo EM é a sua dependência com relação aos valores de inicialização. Visando amenizar essa dificuldade, nós propomos que sejam feitas várias repetições do algoritmo k-Means, com inicializações aleatórias para os centros dos grupos e um fixado número de iterações. Dessa forma, são apresentadas diferentes inicializações para o EM e, após a convergência com cada uma delas, segundo o critério de parada estabelecido, selecionamos como estimativa final aquela que apresentar o maior valor para a função de log-verossimilhança.

Capítulo 5

Estudos de Simulação e Aplicação

Neste capítulo, apresentamos alguns resultados de aplicações do procedimento de classificação proposto no Capítulo 4. São empregados dados simulados com estruturas de classes distintas com respeito a separação das classes, a distribuição nas classes, a dimensão do vetor de características e o tamanho do conjunto de treinamento. É feita uma aplicação em um problema de classificação de assinaturas.

5.1 Problemas com Dados Simulados

No Capítulo 4, propomos uma regra de classificação como sendo uma versão empírica do classificador de Bayes, em que estimamos as densidades condicionais através de misturas finitas de densidades normais. Para averiguar o desempenho da regra proposta, foi

desenvolvido um experimento computacional, onde foram simulados conjuntos de dados com estruturas de classes bastante distintas. Foram consideradas diferenças com relação a forma da fronteira de decisão, as densidades condicionais das classes, a dimensão do vetor de características e o tamanho do conjunto de treinamento. Os dados simulados foram submetidos para serem classificados pela regra proposta e pelos métodos de classificação considerados no Capítulo 2, a saber, Análise Discriminante Linear(ADL), Análise Discriminante Quadrática (ADQ), o método Estimadores por Função Núcleo (EFN), o método dos Vizinhos Mais Próximos (VP) e o dos k-Vizinhos Mais Próximos (kVP). O método proposto será denominado *Misturas Finitas Ponderadas* (MFP).

As situações simuladas, num total de 6 distintas estruturas, são denominadas de Problemas. Todos os problemas foram simulados considerando duas classes ($M = 2$). Nos Problemas 1, 2, 3 e 4, a dimensão das observações foi mantida fixa em $d = 2$, enquanto nos outros problemas essa dimensão variava, sendo $d = 2$, 3, 6 e 10 no Problema 5 e $d = 2$, 3, 5 e 10 no Problema 6.

Em todas os problemas simulados, o experimento foi realizado em duas etapas: (1) foram simuladas observações para selecionar a dimensão do modelo para o MFP e o ajuste dos parâmetros no EFN e kVP; e (2) definidos os parâmetros, foram simuladas observações para estimar os parâmetros necessários para os métodos, a fase de *treinamento*, e outras diferentes observações para serem classificadas, a fase de *teste*.

Na etapa (1), em cada problema, para o método MFP foram gerados conjuntos de 1000 observações e determinados os valores dos critérios de seleção $AIC(k_j)$ e $BIC(k_j)$, com $k_{j_{min}} \leq k_j \leq k_{j_{max}}$, sendo esse procedimento repetido 11 vezes (por tratar-se de dados simulados, não havia muita variabilidade na seleção dos critérios). Em cada uma dessas

repetições, foi determinada o número de componentes que maximizava os critérios de seleção e, como número de componentes para o modelo, foram escolhidos para $k_j^{(A)}$ e $k_j^{(B)}$ os valores que mais vezes foram selecionados como maximizantes, respectivamente, de $AIC(k_j)$ e $BIC(k_j)$ nessas repetições. Quando $k_j^{(A)} = k_j^{(B)} = k_j$, o modelo empregado no treinamento e teste foi uma mistura de normais com k_j componentes e, no caso de $k_j^{(A)} \neq k_j^{(B)}$, na fase de treinamento foram estimados dois modelos, um com $k_j^{(A)}$ e outro com $k_j^{(B)}$ componentes, e na fase de teste foi empregado o modelo ponderado (veja as Subseções 4.3 e 4.5). Consideramos $k_{j_{max}} = 10$ de acordo com as observações em Roeder e Wasserman (1997).

Com os métodos EFN e kVP, foi feita uma simulação a parte para determinar o valor de seus parâmetros, respectivamente, o parâmetro de suavização e o número de vizinhos (veja Seções 2.3.1 e 2.3.2). Para cada problema, foram geradas observações e selecionou-se o valor para esses parâmetros baseado na classificação dessas observações, escolhendo-se aqueles valores que produziam um menor erro de classificação estimado. Para o método EFN, a função núcleo foi mantida fixa, empregando-se a densidade normal multivariada com a matriz de covariâncias estimada para cada classe. Com o kVP, a distância de Mahalanobis, baseada na matriz de covariâncias combinadas amostral, foi empregada como métrica para mensurar a distância entre observações.

Na etapa (2), o número de observações na fase de treinamento (n_{TRE}) variava segundo o problema considerado e, na fase de teste, as observações a serem classificadas (n_{TES}) foi mantido fixo com $n_{TES} = 1000$. Nos Problemas 1, 2, 3 e 4, foram considerados conjuntos de dados com $n_{TRE} = 200$, 500 e 1000, enquanto nos Problemas 5 e 6 foram $n_{TRE} = 500$ e 1000. Em cada Problema e para cada uma das situações consideradas, as fases de

treinamento e de teste foram repetidas 100 vezes, determinando-se a estimativa do erro de classificação em cada repetição, sendo essa estimativa dada por

$$\widehat{e} = P_1\widehat{e}_1 + P_2\widehat{e}_2,$$

onde P e $\widehat{e}_j$ são, respectivamente, a probabilidade a priori e a proporção de classificações erradas nos conjuntos de teste para a classe j, $j = 1, 2$. Obteve-se, portanto, 100 observações da estimativa do erro de classificação para cada um dos métodos considerados.

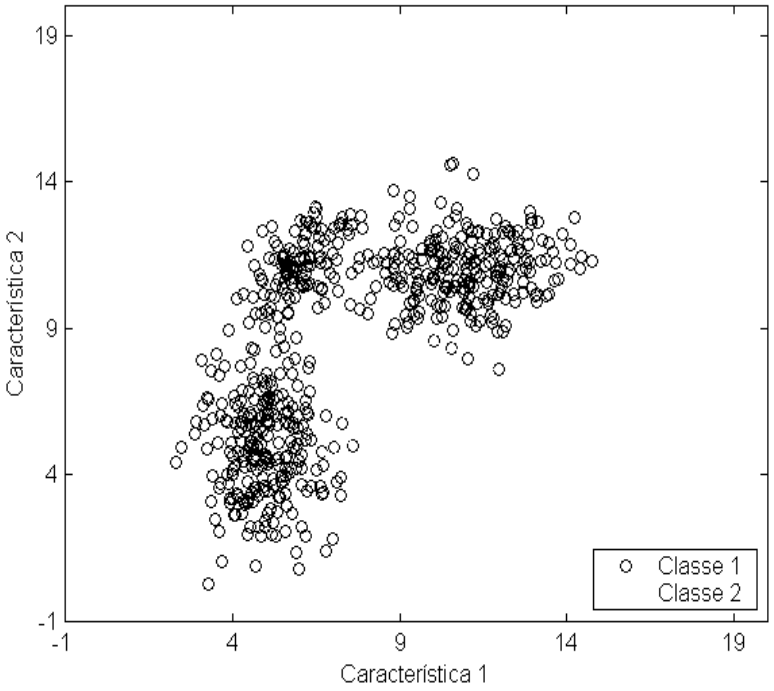
A simulações foram realizadas empregando procedimentos do Sistema de Programas SAS (*The SAS System*). A implementação do método MFP e a geração das observações, foram efetuadas no Procedimento IML, onde a função $normal(\cdot)$ foi empregada para gerar as observações normais univariadas independentes, utilizando como semente o sistema *clock* do microcomputador. Para os outros métodos de classificação empregamos as implementações dadas no Procedimento DISCRIM e, para o método k-Means, o Procedimento FASTCLUS.

Nas subseções a seguir, são descritas as estruturas simuladas e os resultados dos experimentos para cada um dos problemas.

5.1.1 Problema 1

Neste problema, para uma das classes, denominada Classe 1, foi simulada uma distribuição com forma alongada e aproximadamente não convexa. Para a outra classe, a Classe 2, a distribuição simulada foi descontínua, composta de duas subclasses separadas pela distribuição da Classe 1. As observações para essas classes foram geradas, respectivamente,

Figura 5.1: Distribuição das Classes no Problema 1



com probabilidades a priori $P_1 = 0,6$ e $P_2 = 0,4$. Na Figura 5.1, é apresentado um exemplo com 1000 pontos das observações simuladas neste problema, 603 da Classe 1 e 397 da Classe 2. Como pode ser visto dessa figura, uma das dificuldades neste problema é a forma da fronteira de decisão, de um lado a fronteira poderia ser linear e do outro lado, no entanto, é certamente não linear.

Para este problema a dimensão foi fixa, $d = 2$, e variamos o tamanho do conjunto de treinamento, fazendo $n_{TRE} = 200$, 500 e 1000.

As distribuições foram simuladas através de misturas de normais. Na Classe 1, o modelo

foi uma mistura de três normais com os valores dos parâmetros dados por

$$\alpha_{11} = 0,4 \quad \boldsymbol{\mu}_{11} = \begin{pmatrix} 5 \\ 5 \end{pmatrix} \quad \boldsymbol{\Sigma}_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\alpha_{12} = 0,2 \quad \boldsymbol{\mu}_{12} = \begin{pmatrix} 6 \\ 11 \end{pmatrix} \quad \boldsymbol{\Sigma}_{12} = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$$

$$\alpha_{13} = 0,4 \quad \boldsymbol{\mu}_{13} = \begin{pmatrix} 11 \\ 11 \end{pmatrix} \quad \boldsymbol{\Sigma}_{13} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Na Classe 2, o modelo adotado foi uma mistura de duas normais cujos valores dos parâmetros foram

$$\alpha_{21} = 0,5 \quad \boldsymbol{\mu}_{21} = \begin{pmatrix} 10 \\ 6 \end{pmatrix} \quad \boldsymbol{\Sigma}_{21} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\alpha_{22} = 0,2 \quad \boldsymbol{\mu}_{22} = \begin{pmatrix} 6 \\ 16 \end{pmatrix} \quad \boldsymbol{\Sigma}_{22} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Na etapa 1 do experimento, os critérios AIC e BIC selecionaram o mesmo número de componentes em cada classe, $k_1 = 3$ e $k_2 = 2$, coincindindo com o número de componentes empregados para gerar os dados. Os melhores valores para os parâmetros do EFN e kVP foram, respectivamente, 0,1 e 9.

Na Tabela 5.1, são apresentados os resultados para as estimativas do erro de classificação no conjunto de teste, segundo os tamanhos dos conjuntos de treinamento, para cada um dos métodos. São apresentadas a média, o desvio-padrão (entre parênteses) e o Intervalo de Confiança para média (IC), ao nível de 95%, das estimativas do erro para cada um

Tabela 5.1: Problema 1 – Estimativas do Erro de Classificação (%)

Método	n_{TRE}		
	200	500	1000
ADL	31,30(9,03) [29,50;33,09]	30,30(6,12) [29,09;31,52]	29,56(4,80) [28,61;30,51]
ADQ	9,42(1,58) [9,11;9,73]	9,49(1,11) [9,27;9,72]	9,37(1,57) [9,12;9,62]
EFN	1,07(0,35) [1,00;1,14]	0,71(0,31) [0,65;0,77]*	0,46(0,42) [0,38;0,55]*
kVP	<u>0,70</u>(0,23) [0,65;0,74]*	0,67(0,25) [0,61;0,72]*	0,60(0,31) [0,54;0,66]
MFP	0,71(0,25) [0,66;0,77]*	<u>0,64</u>(0,23) [0,59;0,69]*	0,57(0,27) [0,52;0,63]*
VP	0,94(0,34) [0,87;1,01]	0,65(0,35) [0,59;0,73]*	<u>0,42</u>(0,57) [0,30;0,53]*

Os valores são: média (desvio-padrão) [IC a 95%]

dos métodos. Para cada n_{TRE}, a menor média das estimativas do erro é apresentada sublinhada.

Da Tabela 5.1, vemos que, para os métodos paramétricos, os decréscimos com o aumento de n_{TRE} não são significativos e que esses métodos apresentaram as maiores médias. Com os métodos não-paraméricos, à exceção do kVP, verificou-se decréscimos significativos nas médias das estimativas do erro ao aumentar n_{TRE}. O método MFP não foi afetado significativamente com o aumento do conjunto de treinamento.

Analisando os resultados para cada n_{TRE}, vemos que, com n_{TRE}=200, a menor média foi obtida pelo kVP, entretanto, não apresentando diferença siginificativa com relação a média do MFP. Para n_{TRE}=500, o MFP alcançou a menor média, sem diferença significativa com as médias dos métodos EFN, kVP e VP. No caso de n_{TRE}=1000, a menor média foi

a do VP, porém, sem diferença significativa para as médias do kVP e do MFP.

Para cada n_{TRE} considerado, estão assinaladas com " * " as médias estatisticamente menores. Vemos, então, que, para todos os casos, a média das estimativas do erro de classificação do método MFP foi estatisticamente menor ao nível de significância adotado. É interessante observar, que as distribuições neste problema são misturas cujas componentes, em princípio, deveriam ser bem estimadas e, de certa forma, isso explica o bom desempenho do MFP. Por outro lado, o ADL não é adequado a este problema, isso se verificando pelos resultados apresentados por esse método.

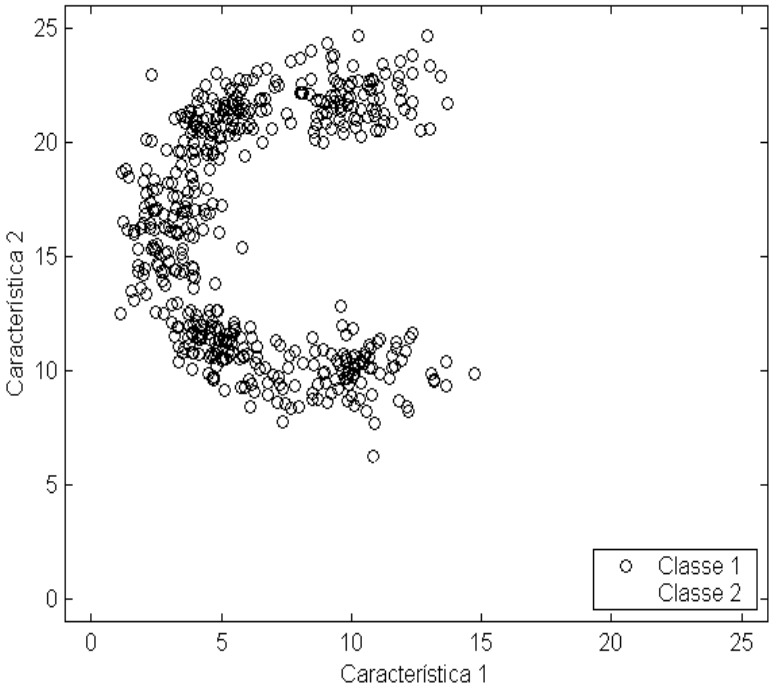
5.1.2 Problema 2

Para este problema, onde temos $d = 2$, as observações foram simuladas de maneira que as formas geométricas das distribuições nas classes não fossem convexas, com as classes relativamente bem separadas, porém, com uma fronteira de decisão estritamente não linear. Foram consideradas probabilidades a priori iguais para as classes, ou seja, $P_1 = P_2 = 0,5$. Um exemplo com 1000 pontos para essas distribuições é apresentada na Figura 5.2, onde 494 são da Classe 1 e 506 da Classe 2. Dessa figura, vemos que, apesar da boa separação das classes, a forma das distribuições pode causar dificuldades para estimar as densidades condicionais.

Como no Problema 1, aqui variamos o tamanho do conjunto de treinamento, considerando-se $n_{TRE} = 200$, 500 e 1000.

Para gerar as observações neste problema, para cada classe, foi adotado um modelo de

Figura 5.2: Distribuição das Classes no Problema 2



mistura de cinco normais com proporções de mistura iguais. Para a Classe 1, os valores empregados para parâmetros foram,

$$\alpha_{11} = \alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{15} = 0,20$$

$$\boldsymbol{\mu}_{11} = \begin{pmatrix} 10 \\ 10 \end{pmatrix} \quad \boldsymbol{\Sigma}_{11} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_{12} = \begin{pmatrix} 5 \\ 11 \end{pmatrix} \quad \boldsymbol{\Sigma}_{12} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_{13} = \begin{pmatrix} 3 \\ 16 \end{pmatrix} \quad \boldsymbol{\Sigma}_{13} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\boldsymbol{\mu}_{14} = \begin{pmatrix} 5 \\ 21 \end{pmatrix} \quad \boldsymbol{\Sigma}_{14} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_{15} = \begin{pmatrix} 10 \\ 22 \end{pmatrix} \quad \boldsymbol{\Sigma}_{15} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Para Classe 2, os valores dos parâmetros foram dados por

$$\alpha_{11} = \alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{15} = 0,20$$

$$\boldsymbol{\mu}_{21} = \begin{pmatrix} 11 \\ 4 \end{pmatrix} \quad \boldsymbol{\Sigma}_{21} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_{22} = \begin{pmatrix} 16 \\ 5 \end{pmatrix} \quad \boldsymbol{\Sigma}_{22} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_{23} = \begin{pmatrix} 18 \\ 10 \end{pmatrix} \quad \boldsymbol{\Sigma}_{23} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\boldsymbol{\mu}_{24} = \begin{pmatrix} 16 \\ 15 \end{pmatrix} \quad \boldsymbol{\Sigma}_{24} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$$\boldsymbol{\mu}_{25} = \begin{pmatrix} 11 \\ 16 \end{pmatrix} \quad \boldsymbol{\Sigma}_{25} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Os critérios AIC e BIC selecionaram número de componentes diferentes. Para as duas classes, o AIC selecionou $k_1^{(A)} = k_2^{(A)} = 6$ e o BIC $k_1^{(B)} = k_2^{(B)} = 5$, estes últimos coincindindo com o número de componentes empregados para gerar os dados. Como os critérios selecionaram dimensões diferentes para os modelos, estimou-se os modelos

ponderados para cada uma das classes. Para o EFN e o kVP, os valores 0,15 e 10 foram, respectivamente, os melhores para os parâmetros desses métodos.

A Tabela 5.2 mostra os resultados referentes as estimativas do erro de classificação, segundo os tamanhos do conjunto de treinamento, para cada método. Como no Problema 1, as informações são a média, o desvio-padrão e o IC para média (95%). Estão assinalados com " * " os resultados cujas médias não apresentam diferenças significativas com relação a menor das médias, que aparece sublinhadas, para cada n_{TRE} considerado.

Da Tabela 5.2, os resultados sugerem que os métodos paramétricos não foram afetados pelo aumento do tamanho do conjunto de treinamento, uma vez que, para esses métodos, não se verificam diferenças significativas entre as médias para os diferentes valores de n_{TRE}. Com os métodos não-paramétricos, as médias decresceram significativamente com o aumento de n_{TRE}, exceto para o kVP quando n_{TRE} aumenta de 500 para 1000. Para o método MFP, a difereça entre as médias com n_{TRE}=500 e n_{TRE}=1000 não é significativa.

Para cada n_{TRE}, os métdos ADL e ADQ tiveram as médias mais altas, sem diferenças significativas entre suas médias. Para n_{TRE}=200 e n_{TRE}=500, os métodos EFN, MFP e VP obtiveram as médias significativamente menores que os demais métodos. Para n_{TRE}=1000, a menor média foi obtida pelos métodos EFN e VP, porém, não se verificando diferença significativa para a média do MFP.

Vimos que BIC selecionou o modelo com a mesma dimensão do modelo empregado para gerar as observações e o AIC adicionou uma componente a mais. Para fins de análise, na Tabela 5.3, são apresentados os resultados do método MFP com os modelos selecionados pelos dois critérios, que denotamos por MFP_A para a seleção do AIC e por MFP_B para

Tabela 5.2: Problema 2 – Estimativas do Erro de Classificação (%)

Método	n_{TRE}		
	200	500	1000
ADL	12,58(1,11) [12,36;12,80]	12,59(1,09) [12,37;12,81]	12,50(0,97) [12,30;12,69]
ADQ	12,58(1,14) [12,36;12,81]	12,55(1,00) [12,33;12,77]	12,49(0,98) [12,29;12,68]
EFN	0,31(0,23) [0,26;0,36]*	0,18(0,15) [0,15;0,21]*	0,11(0,15) [0,08;0,14]*
kVP	0,60(0,36) [0,53;0,68]	0,27(0,18) [0,24;0,31]	0,21(0,15) [0,18;0,24]
MFP	0,26(0,17) [0,23;0,29]*	0,16(0,11) [0,14;0,18]*	0,13(0,12) [0,10;0,15]*
VP	0,32(0,24) [0,27;0,37]*	0,19(0,15) [0,16;0,22]*	0,11(0,17) [0,07;0,14]*

Os valores são: média (desvio-padrão) [IC a 95%]

a seleção do BIC. Dessa tabela, vemos que, apesar do MFP_B, em teoria, ser o modelo correto, as médias das estimativas do erro para os dois modelos não apresentam diferenças significativas para cada n_{TRE} considerado. Com relação ao aumento de n_{TRE}, esses dois modelos tiveram um comportamento semelhante ao do modelo ponderado, apresentaram decréscimo nas médias apenas quando n_{TRE} passou de 200 para 500.

Considerando as análises apresentadas, podemos concluir que, para este problema, a média da estimativas do erro de classificação para o método MFP foi sempre estatisticamente menor. Outro aspecto observado é que, o emprego do modelo ponderado, não comprometeu o desempenho do MFP, tendo em vista os resultados com o modelo "correto", o MFP_B. Com relação aos resultados dos métodos paramétricos, isso era esperado, em virtude das distribuições das classes serem bastante distintas de uma normal e, em particular para o ADL, da forma da fronteira de decisão, estritamente não linear. Os métodos

Tabela 5.3: Problema 2 – Estimativas do Erro de Classificação (%)

Método	n_{TRE}		
	200	500	1000
MFP_A	0,26(0,19) [0,22;0,30]*	0,14(0,12) [0,11;0,16]*	0,12(0,11) [0,09;0,14]*
MFP_B	0,34(0,23) [0,29;0,38]*	0,17(0,16) [0,14;0,20]*	0,12(0,12) [0,10;0,15]*

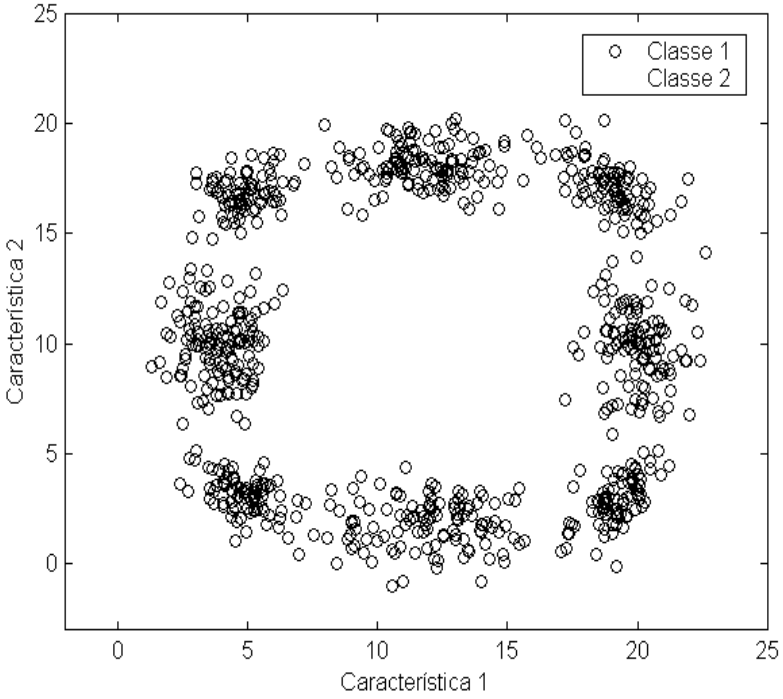
Os valores são: média (desvio-padrão) [IC a 95%]

não-paramétricos, por outro lado, foram capazes de modelar o problema, em particular, com o aumento do tamanho do conjunto de treinamento, apesar da complexidade das distribuições consideradas.

5.1.3 Problema 3

Neste problema, a distribuição de uma das classes era completamente circundada pela distribuição da outra classe. A distribuição da classe mais interior, a Classe 2, foi simulada através de uma normal com probabilidade a priori $P_2 = 0,20$, enquanto para a outra classe, a Classe 1, através de uma mistura de normais com $P_1 = 0,80$. Na Figura 5.3, apresentamos um exemplo com 1000 pontos para essas distribuições, sendo 800 pontos da Classe 1 e 200 da Classe 2. Pode ser visto dessa figura, que as classes são relativamente bem separadas e que a fronteira de decisão é estritamente não linear, contornando a Classe 2. Pela forma das distribuições condicionais, a estimação da densidade para a Classe 2 não apresenta dificuldades, no entanto, para a Classe 1, essa estimação é bem mais complexa.

Figura 5.3: Distribuição das Classes no Problema 3



As observações para a Classe 1 foram geradas empregando-se uma mistura de oito normais, com os valores para os parâmetros dados por

$$\alpha_{11} = 0,10 \quad \boldsymbol{\mu}_{11} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} \quad \boldsymbol{\Sigma}_{11} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$$\alpha_{12} = 0,15 \quad \boldsymbol{\mu}_{12} = \begin{pmatrix} 4 \\ 10 \end{pmatrix} \quad \boldsymbol{\Sigma}_{12} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\alpha_{13} = 0,10 \quad \boldsymbol{\mu}_{13} = \begin{pmatrix} 5 \\ 17 \end{pmatrix} \quad \boldsymbol{\Sigma}_{13} = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$$

$$\alpha_{14} = 0,15 \quad \boldsymbol{\mu}_{14} = \begin{pmatrix} 12 \\ 18 \end{pmatrix} \quad \boldsymbol{\Sigma}_{14} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\alpha_{15} = 0,10 \quad \boldsymbol{\mu}_{15} = \begin{pmatrix} 19 \\ 17 \end{pmatrix} \quad \boldsymbol{\Sigma}_{15} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$$\alpha_{16} = 0,15 \quad \boldsymbol{\mu}_{16} = \begin{pmatrix} 20 \\ 10 \end{pmatrix} \quad \boldsymbol{\Sigma}_{16} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\alpha_{17} = 0,10 \quad \boldsymbol{\mu}_{17} = \begin{pmatrix} 19 \\ 3 \end{pmatrix} \quad \boldsymbol{\Sigma}_{17} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\alpha_{18} = 0,15 \quad \boldsymbol{\mu}_{18} = \begin{pmatrix} 12 \\ 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_{18} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Para a Classe 2, as observações foram geradas segundo uma normal cujos valores dos parâmetros foram

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 12 \\ 10 \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}.$$

O AIC e BIC selecionaram o mesmo modelo em cada uma das classes, $k_1^{(A)} = k_1^{(B)} = 8$ e $k_2^{(A)} = k_2^{(B)} = 1$, coincindindo com os modelos empregados para gerar as observações. Dessa forma, foi estimado o modelo com $k_1 = 8$ componentes para a Classe 1 e com $k_1 = 1$ para a Classe 2. Na seleção dos parâmetros para o EFN e kVP, os valores que apresentaram melhores resultados foram , respectivamente, 0,15 e 5.

Os resultados para as estimativas do erro de classificacão, para cada método e segundo os tamanhos de n_{TRE}, são apresentados na Tabela 5.4. Como nos problemas já descritos,

Tabela 5.4: Problema 3 – Estimativas do Erro de Classificação (%)

Método	n_{TRE}		
	200	500	1000
ADL	20,00(0,00)	20,00(0,00)	20,00(0,00)
ADQ	5,89(0,88) [5,71;6,06]	5,49(0,51) [5,39;5,59]	5,32(0,60) [5,20;5,44]
EFN	2,74(0,97) [2,55;2,93]	1,09(0,51) [0,99;1,20]	0,41(0,60) [0,29;0,53]
kVP	0,45(0,23) [0,40;0,49]*	0,33(0,17) [0,30;0,36]	0,28(0,16) [0,24;0,31]
MFP	0,35(0,21) [0,31;0,40]*	0,22(0,15) [0,19;0,25]*	0,20(0,12) [0,17;0,22]*
VP	0,49(0,28) [0,43;0,54]	0,26(0,16) [0,22;0,29]*	0,13(0,21) [0,08;0,17]*

Os valores são: média (desvio-padrão) [IC a 95%]

são apresentadas a média, o desvio-padrão e o IC para média das estimativas do erro. Também, como nas descrições anteriores, para cada n_{TRE} considerado, os resultados assinalados com " * " não apresentam diferença significatica com relação a menor das médias (sublinhada).

Analisando a Tabela 5.4, vemos que os métodos paramétricos apresentaram as maiores médias. O método ADL, em particular, classificou todas as observações dos conjuntos de teste na Classe 1, independentemente do tamanho do conjunto de treinamento. Com os métodos não-paramétricos, verificou-se decréscimos significativos em suas médias com o aumento de n_{TRE}, exceto para o kVP ao aumentar n_{TRE} de 500 para 1000. Para o MFP, não houve diferença significativa para suas médias com n_{TRE}=500 e n_{TRE}=1000.

Considerando os resultados para cada n_{TRE} em separado, vemos que a média do erro de

classificação do MFP foi significativamente menor para todos os tamanhos de conjunto de treinamento. É possível que esses resultados sejam decorrentes de que o MFP adotou o modelo que, em teoria, era o correto. É importante observar, também, que as componentes da mistura da distribuição na Classe 1, estão relativamente bem separadas, o que facilita o processo de estimação do modelo. Com relação aos resultados dos métodos paramétricos, temos uma situação em que esses métodos poderiam modelar apenas uma das classes, sendo as diferenças entre os resultados desses métodos decorrentes da forma da fronteira de decisão que geram, linear com o ADL e quadrática com o ADQ. De forma diferente, os métodos não-paramétricos modelaram o problema de forma equivalente ao modelo "correto" do MFP.

5.1.4 Problema 4

A estrutura de classes simuladas para este problema, consistia de uma classe com distribuição descontínua composta por duas normais, denominada Classe 1, e a outra com uma distribuição normal, a Classe 2. A localização das classes foi montada de maneira que a distribuição da Classe 1 tivesse suas componentes separadas pela distribuição da Classe 2. As probabilidades a priori foram $P_1 = 0,60$ e $P_2 = 0,40$. Na Figura 5.4 é apresentado um exemplo dessas distribuições com 607 observações da Classe 1 e 393 da Classe 2. Vemos que para esse problema, com um erro de classificação pequeno, uma fronteira de decisão composta de duas retas pode separar as classes. Pelas formas das distribuições, a estimação das densidades condicionais não apresentam grandes dificuldades, pois em uma classe é simplesmente uma normal e na outra uma mistura com duas componentes bem separadas.

Para gerar as observações da Classe 1 foi empegada uma mistura de duas normais com parâmetros

$$\alpha_{11} = 0,50 \quad \boldsymbol{\mu}_{11} = \begin{pmatrix} 4 \\ 10 \end{pmatrix} \quad \boldsymbol{\Sigma}_{11} = \begin{pmatrix} 1 & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & 3 \end{pmatrix}$$

$$\alpha_{12} = 0,50 \quad \boldsymbol{\mu}_{12} = \begin{pmatrix} 16 \\ 10 \end{pmatrix} \quad \boldsymbol{\Sigma}_{12} = \begin{pmatrix} 1 & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & 3 \end{pmatrix}.$$

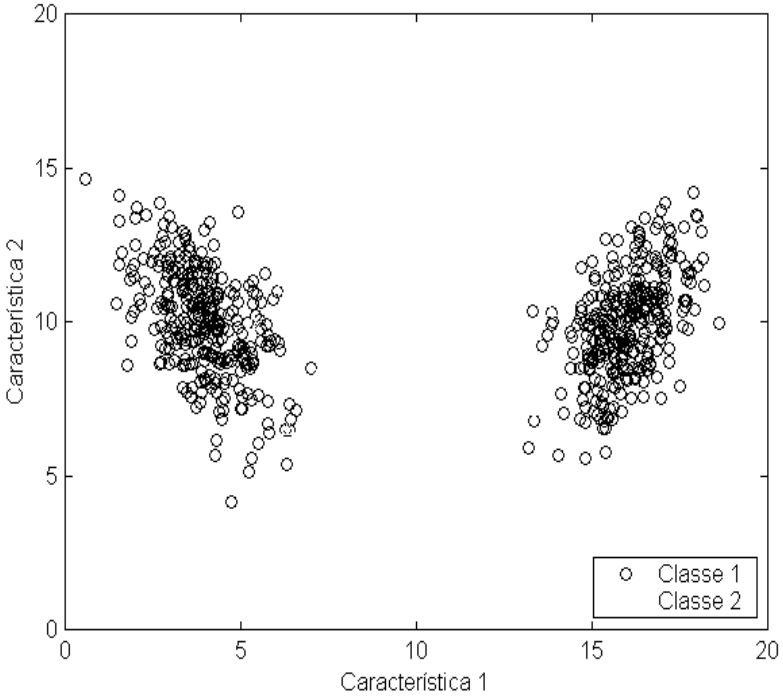
Para Classe 2, empregou-se uma normal com parâmetros

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 10 \\ 10 \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

Os critérios AIC e BIC selecionaram as mesmas dimensões para o modelo em cada classe, $k_1^{(A)} = k_1^{(B)} = 2$ e $k_2^{(A)} = k_2^{(B)} = 1$. Temos, portanto, que os critérios decidiram por modelos equivalentes aos empregados para gerar as observações. Para os métodos EFN e kVP, os valores para seus parâmetros que apresentaram os melhores resultados foram, respectivamente, 0,15 e 5.

Na Tabela 5.5 são apresentados os resultados referentes as estimativas do erro de classificação para os métodos, segundo os valores de n_{TRE} considerados neste problema. Em geral, os resultados sugerem que, com o aumento do tamanho do conjunto de treinamento, ocorre um decréscimo nas médias da estimativa do erro de classificação. Para o MFP, os decréscimos nas médias não foram significativos com o aumento de n_{TRE}. Com os métodos paramétricos, que apresentaram as maiores médias, os decréscimos não são significativos, exceto com o ADL quando n_{TRE} foi aumentado de 200 para 500. Com relação aos métodos não-paramétricos, o EFN e o kVP tiveram decréscimos significativos em suas médias em todas as mudanças de n_{TRE}, já o VP apenas no caso de 500 para 1000.

Figura 5.4: Distribuição das Classes no Problema 4



Analisando os resultados para os valores de n_{TRE}, vemos que o MFP teve a menor média quando n_{TRE}=200. Para n_{TRE}=500, as médias do MFP e do VP foram significativamente menores. Com n_{TRE}=1000, os métodos EFN, kVP, MFP e VP apresentaram as menores médias sem diferenças significativas entre elas.

Considerando as análises feitas, os resultados do MFP eram esperados, dado que as distribuições eram adequadas à modelagem desse método e, em teoria, sem dificuldades para a estimação dos parâmetros. Deve ser ressaltado, também, que os métodos não-paramétricos foram capazes de lidar com a distribuição descontínua da Classe 1, em particular, com o aumento do conjunto de treinamento. Com os métodos paramétricos,

Tabela 5.5: Problema 4 – Estimativas do Erro de Classificação (%)

Método	n_{TRE}		
	200	500	1000
ADL	40,68(1,44)[40,40;40,97]	40,13(0,67)[39,99;40,26]	40,02(0,11)[40,00;40,04]
ADQ	6,01(0,97) [5,82;6,21]	5,66(0,65) [5,53;5,79]	5,50(0,55) [5,39;5,61]
EFN	3,14(1,01) [2,93;3,34]	1,42(0,55) [1,31;1,53]	0,69(0,63) [0,56;0,81]*
kVP	1,33(0,45) [1,24;1,42]	0,95(0,32) [0,89;1,01]	0,72(0,30) [0,66;0,78]*
MFP	0,78(0,26) [0,73;0,83]	0,68(0,29) [0,63;0,74]*	0,72(0,26) [0,66;0,77]*
VP	1,36(0,52) [1,26;1,46]	0,79(0,42) [0,71;0,88]*	0,62(0,56) [0,51;0,72]*

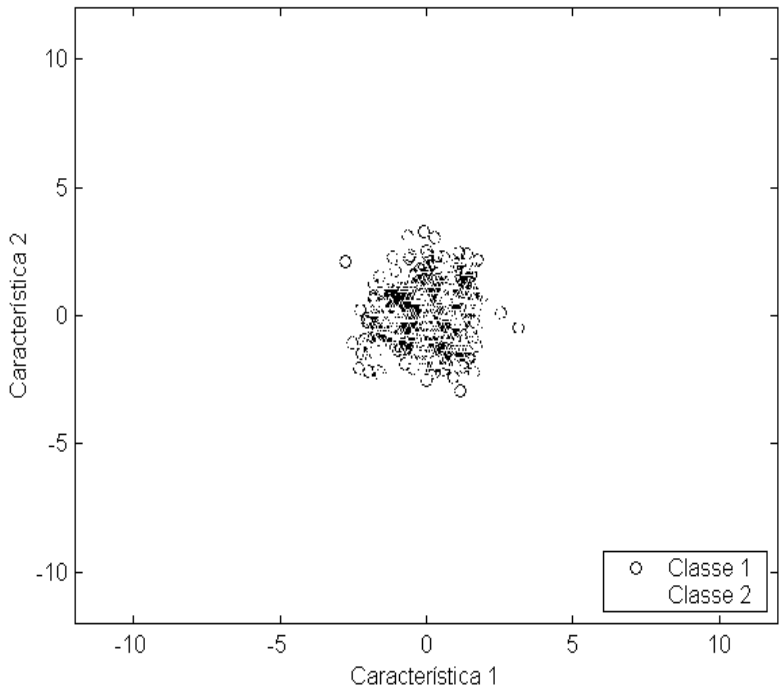
Os valores são: média (desvio-padrão) [IC a 95%]

era possível modelar adequadamente apenas uma das classes e, em particular para o ADL, a forma da fronteira de decisão era uma dificuldade a mais.

5.1.5 Problema 5

Neste problema, o objetivo era simular uma estrutura de classes em que as classes fossem diferenciadas apenas em termos de suas matrizes de dispersão. Foram, então, simuladas duas classes com distribuição normal, com vetores de médias iguais e matrizes de covariâncias diferentes, onde a matriz para uma das classes era oito vezes a da outra. Essa estrutura, retrata algumas situações reais como, por exemplo, dados relativos a assinaturas verdadeiras e falsificadas, onde conjectura-se que os vetores de médias são aproximadamente iguais, porém, a classe das falsas apresentam dispersão muito superior a das

Figura 5.5: Distribuição das Classes no Problema 5



verdadeiras.

As classes foram simuladas com probabilidades a priori $P_1 = 0,40$ para a Classe 1 $P_2 = 0,60$ para a Classe 2. Um exemplo das distribuições dessas classes, com $d = 2$, é apresentado na Figura 5.5, onde temos 407 observações da Classe 1 e 593 da Classe 2. Dessa figura, vemos que este problema é bastante complexo em com relação a separação das classes.

Para este problema, a dimensão das observações variava, tendo sido consideradas as dimensões d= 2, 3, 6 e 10. Os tamanhos dos conjuntos de treinamento considerados foram

n_{TRE}=500 e 1000. Para gerar as observações, nas duas classes, empregou-se como modelo uma normal cujos valores dos vetores de médias e das matrizes de covariâncias foram, respectivamente,

$$\boldsymbol{\mu}_{1(d\times 1)} = \boldsymbol{\mu}_{2(d\times 1)} = (0, 0, 0, ..., 0)^T,$$

$$\boldsymbol{\Sigma}_1 = \mathbf{I}_{(d)} \quad \text{e} \quad \boldsymbol{\Sigma}_2 = 8\mathbf{I}_{(d)},$$

onde $\mathbf{I}_{(d)}$ é a matriz identidade de dimensão d.

Na seleção dos modelos para o MFP, os critérios AIC e BIC selecionaram o número de componentes diferentes para o modelo na Classe 1, sendo $k_1^{(A)} = 2$ e $k_1^{(B)} = 1$. Para a Classe 2, os critérios selecionaram o mesmo número de componentes, $k_2^{(A)} = k_2^{(B)} = 1$. Com relação aos métodos EFN e kVP, os valores selecionados para seus parâmetros foram, respectivamente, 0,1 e 9.

Na Tabela 5.6 são apresentados os resultados referentes as estimativas do erro de classificação para cada um dos métodos. Para cada uma das combinações de d e n_{TRE}, são apresentadas a média, o desvio-padrão e IC para média, a 95%, das estimavas do erro.

Inicialmente, analisamos os resultados na Tabela 5.6 considerando-se o aumento de n_{TRE} para cada valor de d. Dos resultados na tabela, vemos que, com relação aos métodos paramétricos, o ADL apresentou decréscimos significativos nas médias para todo d e o ADQ apenas no caso de d=10. Para os métodos não-paramétricos, no entanto, a aumento de n_{TRE} decresceu as médias para todos os valores de d. Com relação ao MFP, os resultados indicam um decréscimo nas médias somente para d=2 e d=10.

Tabela 5.6: Problema 5 – Estimativas do Erro de Classificação (%)

Método	d	n_{TRE}	
		500	1000
ADL	2	40,38(0,69) [40,22;40,55]	40,10(0,28) [40,04;40,16]
	3	40,91(1,52) [40,61;41,22]	40,22(0,56) [40,11;40,33]
	6	43,12(2,69) [42,59;43,66]	40,63(1,00) [40,43;40,83]
	10	45,98(3,05) [45,38;46,59]	41.61(1,46) [41,32;41,90]
ADQ	2	18,80(1,17) [18,57;19,03]	18,60(1,33) [18,34;18,87]
	3	12,74(1,22) [12,50;12,99]	12,66(1,09) [12,44;12,88]
	6	4,66(0,68) [4,53;4,80]	4,48(0,67) [4,34;4,61]
	10	1,57(0,36) [1,50;1,65]	1,28(0,31) [1,22;1,34]
EFN	2	17,70(2,99) [17,10;18,29]	13,41(4,00)[12,61;14,20]
	3	16,24(5,13) [15,22;17,26]	7,11(8,54) [5,42;8,81]
	6	19,52(6,42) [18,24;20,79]	10,26(15,77) [7,13;13,39]
	10	23,88(8,32) [22,22;25,53]	7,14(14,83) [4,20;10,08]
kVP	2	19,23(1,67) [18,89;19,56]	17,74(1,91) [17,36;18,12]
	3	14,05(1,37) [13,77;14,32]	12,56(1,31) [12,30;12,82]
	6	14,09(1,59) [13,77;14,40]	10,19(1,44) [9,91;10,48]
	10	15,93(2,23) [15,49;16,37]	12,33(2,02) [11,93;12,73]
MFP	2	19,99(1,26) [19,74;20,24]	12,93(1,44)[12,64;13,22]
	3	13,34(1,29) [13,08;13,59]	13,36(1,16) [13,12;13,59]
	6	4,72(0,72) [4,58;4,86]	4,62(0,68) [4,48;4,75]
	10	1,51(0,36) [1,43;1,58]	1,23(0,35) [1,16;1,30]

Os valores são: média (desvio-padrão) [IC a 95%]

Tabela 5.6: Continuação

Método	d	n_{TRE}	
		500	1000
VP	2	15,22(5,03) [14,22;16,22]	4,66(10,02)[2,67;6,65]
	3	11,70(4,43) [10,82;12,58]	4,24(8,07) [2,64;5,84]
	6	7,16(2,42) [6,68;7,64]	3,16(4,89) [2,19;4,13]
	10	10,31(3,50) [9,61;11,00]	2,43(5,07) [1,42;3,43]

Os valores são: média (desvio-padrão) [IC a 95%]

Consideramos, agora, os resultados da Tabela 5.6 dentro de cada valor de n_{TRE}. Para os métodos paramétricos, vemos que o aumento da dimensão das observações produziu acréscimos nas médias para o ADL, porém, levou a decréscimos significativos nas médias do ADQ. Com relação aos métodos não-paramétricos, não se verificou um comportamento bem definido com relação ao aumento de d, às vezes aumentando e outras vezes decrescendo as médias. O método MFP, por outro lado, apresentou um comportamento bem definido, aprensentado decréscimos significativos para as médias com aumento da dimensão dos dados.

Comparando os métodos, vemos da Tabela 5.6 que o método VP apresentou as menores médias no caso de d=2, para os dois casos de n_{TRE}. Para d=3, com n_{TRE}=500 a média do VP e do ADQ foram equivalentes e, para n_{TRE}=1000, o método VP teve média inferior aos demais. Com a dimensões maiores, as médias do ADQ e do MFP foram significativamente inferiores às dos outros métodos.

Na Tabela 5.7, são apresentados as médias das estimativas do erro de classificação para MFP com os modelos selecionados pelo AIC (MFP_A) e pelo BIC (MFP_B) separadamente.

Tabela 5.7: Problema 5 – Método MFP com Seleção do AIC e BIC

Método	d	n_{TRE}	
		500	1000
MFP_A	2	18,87(1,19) [18,63;19,11]	18,56(1,27)[18,31;18,81]
	3	12,77(0,94) [12,58;12,95]	12,41(1,28) [12,15;12,66]
	6	4,65(0,78) [4,50;4,81]	4,32(0,85) [4,15;4,49]
	10	1,44(0,42) [1,36;1,52]	1,06(0,35) [0,99;1,13]
MFP_B	2	18,82(1,11) [18,60;19,04]	18,80(1,28)[18,54;19,05]
	3	13,90(8,76) [12,16;15,64]	12,67(1,14) [12,44;12,89]
	6	4,66(0,69) [4,52;4,79]	4,58(0,71) [4,44;4,73]
	10	1,53(0,42) [1,44;1,61]	1,44(0,45) [1,35;1,53]

Os valores são: média (desvio-padrão) [IC a 95%]

Considerando a variação de n_{TRE} para cada d, vemos um decréscimo significativo na médias apenas para o MFP_A quando d=6 e d=10. Analisando as resultados com relação a variação de d em cada n_{TRE}, os dois modelos tiveram decréscimos em sua médias com o aumento da dimensão das observações, isso para os dois valores de n_{TRE} considerados.

Das análises apresentadas, temos que o método ADL apresentou um desempenho muito inferior aos dos outros métodos, no entanto, isso era esperado, tendo em vista que esse método é completamente inadequado à estrutura de classes do problema considerado aqui. Os reultados mostraram, também, uma certa dificuldade dos métodos não-paramétricos em lidar com o aumento da dimensão das observações. No que diz respeito ao MFP, vimos que, para os três modelos, os resultados indicam um comportamento muito semelhante

e, por outro lado, esses resultados são equivalentes aos do ADQ. Esse último resultado comentado, é provavelmente justificado pelo fato de que o modelo no MFP_B é o mesmo empregado no ADQ.

5.1.6 Problema 6

Para este problema, foram simuladas duas classes relativamente bem separadas, de maneira que, para $d = 2$, a fronteira de decisão fosse aproximadamente linear. Um exemplo das distribuições das classes é apresentada na Figura 5.6. Dessa figura, vemos que as distribuições geradas não são convexas, apresentam uma boa separação e, admitindo-se um pequeno erro de classificação, podemos adotar uma fronteira de decisão linear. A partir dessa configuração bidimensional, foi, então, aumentada a dimensão das observações por acrescentar variáveis independentes e redundantes para efeito de discriminação, obtendo-se vetores de observações com $d = 3$, 5 e 10.

Para gerar as observações com $d = 2$, para cada classe, empregou-se uma mistura de cinco normais. Na Classe 1, os valores para os parâmetros foram

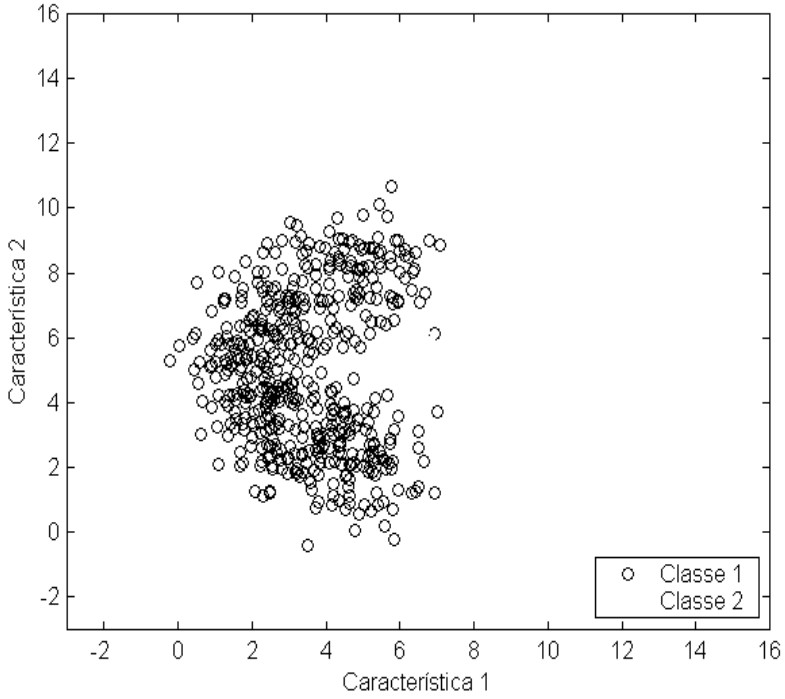
$$\alpha_{11} = \alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{15} = 0,20\,,$$

$$\boldsymbol{\mu}_{11} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \quad \boldsymbol{\mu}_{12} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \quad \boldsymbol{\mu}_{13} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}, \quad \boldsymbol{\mu}_{14} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \quad \boldsymbol{\mu}_{15} = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$$

e $\boldsymbol{\Sigma}_{1j} = \mathbf{I}_{(2)}$, $j = 1, 2, ..., 5$. Para a os parâmetros da Classe 2, os valores foram

$$\alpha_{21} = \alpha_{22} = \alpha_{23} = \alpha_{24} = \alpha_{25} = 0,20\,,$$

Figura 5.6: Distribuição das Classes no Problema 6 com $d = 2$

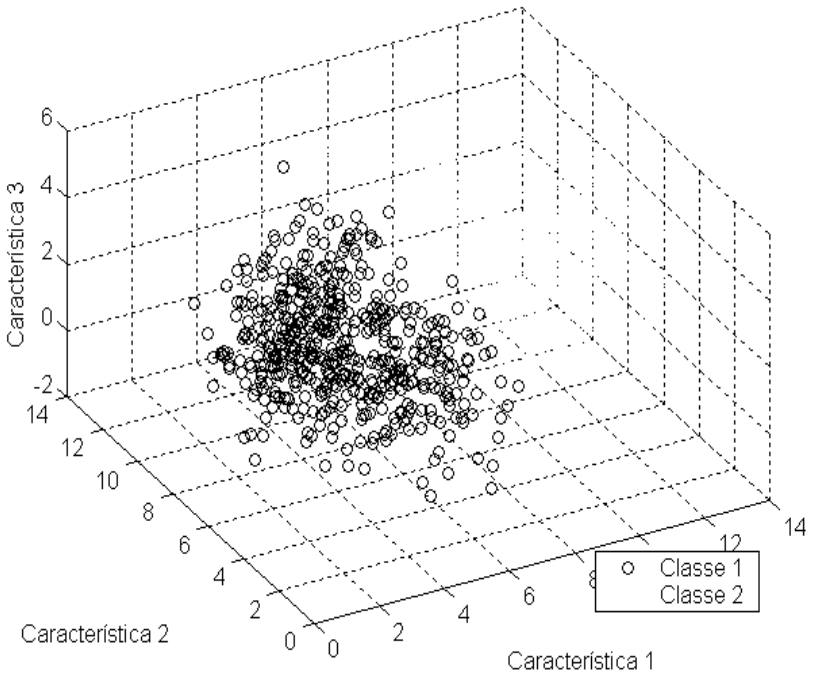


$$\boldsymbol{\mu}_{21} = \begin{pmatrix} 9 \\ 5 \end{pmatrix}, \quad \boldsymbol{\mu}_{22} = \begin{pmatrix} 11 \\ 6 \end{pmatrix}, \quad \boldsymbol{\mu}_{23} = \begin{pmatrix} 12 \\ 8 \end{pmatrix}, \quad \boldsymbol{\mu}_{24} = \begin{pmatrix} 11 \\ 10 \end{pmatrix}, \quad \boldsymbol{\mu}_{25} = \begin{pmatrix} 9 \\ 11 \end{pmatrix}$$

e, também, $\boldsymbol{\Sigma}_{2j} = \mathbf{I}_{(2)}$, $j = 1, 2, ..., 5$.

O procedimento para aumentar a dimensão com váriáveis redundantes foi acrescentar aos vetores de observações bidimensionais, variáveis independentes geradas por uma $N(3, 1)$, isso sendo feito para as observações das duas classes. Dessa forma, as variáveis acrescentadas não continham informação discriminativa. Esse procedimento foi repetido para obter-se observações com dimensões 3, 5 e 10, como mencionado.

Figura 5.7: Distribuição das Classes no Problema 6 com $d = 3$



Para termos uma idéia das distribuições das classes com $d > 2$, a Figura 5.7 apresenta um exemplo das observações com $d = 3$. Pode ser visualizado dessa figura que, no plano (Característica 1;Característica 3), as classes ainda apresentam algum grau de separação, no entanto, considerando o plano (Característica 2;Característica 3), as classes estão completamente superpostas.

Para a seleção do número de componentes dos modelos para o MFP, foi levada em consideração as diferentes dimensões das observações. Para $d = 2$, 3 e 5, o critério AIC selecionou $k_1^{(A)} = k_2^{(A)} = 3$ e o BIC $k_1^{(B)} = k_2^{(B)} = 2$. No caso de d=10, as escolhas foram $k_1^{(A)} = k_2^{(A)} = 2$ e $k_1^{(B)} = k_2^{(B)} = 1$. Por esses resultados, para todas as dimensões

consideradas para as observações, foi estimado o modelo ponderado.

Os valores empregados para os parâmetros do EFN e kVP foram, repectivamente, 0,8883 e 8. No experimento para seleção dos parâmetros, os valores dos parâmetros desses métodos não foram afetados de forma significativa pelo aumento da dimensão das observações.

A Tabela 5.8 apresenta os resultados referentes as estimativas do erro de classificação. São apresentadas a média, o desvio-padrão e o IC para a médias das estimativas, segundo o tamanho dos conjuntos de treinamento e a dimensão dos dados, para cada um dos métodos.

Da Tabela 5.8, vemos que as médias das estimativa da taxa de erro foram pequenas para todos os métodos, isso se verificando para todas as dimensões consideradas. Para os métodos paramétricos, nos dois casos de n_{TRE}, os resultados indicam que esse métodos não foram afetados significativamente pela inclusão das variáveis redundantes. Com relação aos métodos não paramétricos, o kVP apresentou um aumento significativa na média ao considerar d =5 e 10, para os dois valores de n_{TRE}, o EFN no caso d=5 e n_{TRE}=500 e o VP com d=5 e n_{TRE}=1000. O método MFP, para todos os valores de n_{TRE}, teve um aumento significativo somente no caso de d=10.

Com relação tamanho do conjunto de treinamento, à exceção do EFN e do kVP, os métodos não foram afetados com o aumento do conjunto de treinamento. O método EFN teve um aumento na média no caso de d=2 e o kVP apresentou um decréscimo com d=10.

A Tabela 5.9 apresenta os resultados para os métodos MFP_A e MFP_B. Dessa tabela, vemos que os métodos não foram afetados significativamente com o aumento no valor de

Tabela 5.8: Problema 6 – Estimativas do Erro de Classificação

Método	d	n_{TRE}	
		500	1000
ADL	2	0,91(0,38) [0,84;0,99]	0,89(0,29) [0,84;0,95]
	3	0,91(0,38) [0,83;0,98]	0,92(0,33) [0,85;0,98]
	5	0,91(0,42) [0,82;0,99]	0,94(0,30) [0,88;1,00]
	10	0,94(0,42) [0,86;1,03]	0,92(0,28) [0,86;0,97]
ADQ	2	0,88(0,38) [0,80;0,96]	0,90(0,29) [0,84;0,96]
	3	0,91(0,40) [0,83;0,99]	0,92(0,31) [0,86;0,98]
	5	0,93(0,40) [0,85;1,01]	0,94(0,33) [0,88;1,01]
	10	0,95(0,42) [0,86;1,03]	0,92(0,31) [0,86;0,98]
EFN	2	0,47(0,36) [0,39;0,54]	0,89(0,29) [0,84;0,95]
	3	0,14(0,35) [0,07;0,21]	0,27(0,34) [0,20;0,34]
	5	0,33(0,43) [0,25;0,42]	0,41(0,37) [0,33;0,48]
	10	0,43(0,73) [0,28;0,57]	0,36(0,70) [0,22;0,50]
kVP	2	0,39(0,24) [0,34;0,44]	0,36(0,21) [0,32;0,40]
	3	0,45(0,30) [0,39;0,51]	0,41(0,21) [0,36;0,45]
	5	0,64(0,38) [0,57;0,72]	0,53(0,23) [0,48;0,57]
	10	0,96(0,47) [0,86;1,05]	0,76(0,30) [0,70;0,82]
MFP	2	0,38(0,27) [0,32;0,43]	0,35(0,18) [0,32;0,39]
	3	0,33(0,25) [0,28;0,38]	0,35(0,19) [0,31;0,39]
	5	0,39(0,32) [0,32;0,45]	0,39(0,19) [0,35;0,43]
	10	0,55(0,37) [0,48;0,63]	0,50(0,21) [0,46;0,55]

Os valores são: média (desvio-padrão) [IC a 95%]

Tabela 5.8: Continuação

Método	d	n_{TRE}	
		500	1000
VP	2	0,17(0,35) [0,10;0,24]	0,23(0,31) [0,16;0,29]
	3	0,14(0,34) [0,07;0,21]	0,14(0,29) [0,08;0,20]
	5	0,30(0,54) [0,19;0,40]	0,31(0,45) [0,22;0,40]
	10	0,58(0,96) [0,39;0,77]	0,42(0,78) [0,26;0,58]

Os valores são: média (desvio-padrão) [IC a 95%]

n_{TRE}. Com relação ao aumento da dimensão, o MFP_B apresentou um aumento significativo no caso de d=10 para os dois valores de n_{TRE}.

Para este problema, em teoria, a situação mais difícil seria n_{TRE}=500 com d=10 e a mais confortável n_{TRE}=1000 com d=2. Na primeira situação, os métodos EFN, MFP e VP, apresentaram as menores médias sem diferenças significativas entre elas. Na segunda situação, o VP teve uma média inferior à dos outros métodos.

A expectativa era que a inclusão das variáveis redundantes afetasse todos os métodos. De maneira geral, no entanto, essas inclusões não afetaram de forma mais acentuada o desempenho dos métodos. Vemos que as médias das taxas de erros, apesar de aumentarem com o aumento da dimensão, foram sempre muito pequenas. Possivelmente, a forma adotada para incluir essa perturbação às variáveis não foi adequada para alterar, de forma significativa, o desempenho dos métodos.

Tabela 5.9: Problema 6 – Método MFP com Seleção do AIC e BIC

Método	d	n_{TRE}	
		500	1000
MFP_A	2	0,33(0,28) [0,28;0,39]	0,34(0,21) [0,29;0,39]
	3	0,36(0,260 [0,31;0,41]	0,36(0,18) [0,32;0,39]
	5	0,30(0,26) [0,25;0,35]	0,34(0,19) [0,31;0,38]
	10	0,40(0,31) [0,34;0,46]	0,41(0,19) [0,37;0,45]
MFP_B	2	0,42(0,27) [0,37;0,48]	0,43(0,19) [0,39;0,47]
	3	0,35(0,24) [0,31;0,40]	0,42(0,19) [0,38;0,46]
	5	0,38(0,27) [0,33;0,44]	0,44(0,20) [0,40;0,48]
	10	0,89(0,40) [0,81;0,97]	0,95(0,30) [0,89;1,01]

Os valores são: média (desvio-padrão) [IC a 95%]

5.1.7 Considerações Sobre os Problemas Simulados

Nestes experimentos computacionais, não havia a intenção de estabelecer a superioridade de qualquer um dos métodos considerados. Determinações desse tipo, inclusive, não fazem sentido aqui, tendo em vista que os problemas simulados eram mais apropriados à modelagem por mistura de densidades normais. A intenção era gerar estruturas, com um certo grau de complexidade, e observar o comportamento do método proposto, o MFP, com respeito a sua capacidade de classificar corretamente as observações. Nesse sentido, os outros métodos serviram de referência para avaliar esse comportamento.

Com respeito a modelagem, a nossa preocupação não foi recuperar exatamente os mo-

delos simulados mas, sim, obter aproximações para esses modelos capazes de estimar as probabilidades a posteriori de forma satisfatória, avaliadas pelas estimativas do erro de classificação. Essa perspectiva, está baseada nas observações de Ripley (1996, Seção 6.4), onde o autor argumenta que, para os problemas de RPS, é necessário somente uma boa aproximação da mistura, no sentido de uma alta verossimilhança, em vez de uma estimação precisa dos parâmetros do modelo. Por esse argumento, inclusive, se justifica o emprego do algoritmo EM para a estimação dos parâmetros, desde que não há garantias de atingirmos o máximo global no espaço dos parâmetros.

Dos resultados observados, tendo em vista nossos objetivos, temos que o MFP apresentou um desempenho satisfatório. Nos casos onde foi necessário, nos Problemas 2, 5 e 6, vimos que o emprego do modelo ponderado apresentou resultados estatisticamente equivalentes aos dos modelos "corretos". No caso das distribuições das classes, ou de uma das classe, não serem misturas (Problemas 2, 4, e 5), o método se adaptou ao problema, selecionando um modelo que demonstrou ser capaz de descrever os dados. Para o Problema 6, onde as componentes da distribuição simulada estavam muito próximas, o que dificulta a identificação do modelo, o MFP modelou o problema de forma satisfatória, a considerar seus resultados com relação aos dos métodos não-paramétricos. Para esse último problema, como já comentado, a forma adotada para perturbar os dados não permitiu uma avaliação mais precisa dos objetivos com esse problema.

Ainda com relação ao número de componentes para os modelos, condizente com a teoria, em alguns caso o critério AIC indicou um número de componentes maior que o BIC. É importante ressaltar, no entanto, que as taxas de erros com o modelo superestimado foram equivalentes às dos modelos "corretos". Esses resultados indicam uma concordância

com a literatura, em que é argumentado que é necessário apenas uma aproximação das distribuições condicionais nos problemas de RPS.

Embora os resultados com emprego do modelo ponderado tenham sido satisfatórios, temos que considerar o custo computacional. Para esse caso, torna-se necessário estimar os parâmetros para os dois modelos e, dependendo da dismensão das observações e o número de componentes necessários na mistura, isso pode demandar um esforço computacional muito grande. No Problema 2, por exemplo, onde foi necessário estimar um modelo de mistura com 6 componentes e outro com 5, mesmo as observações sendo bidimensionais, o algoritmo EM foi lento e consumindo muita memória computacional (memória RAM).

Outra questão, diz respeito a dimensão das observações com relação a estimação das matrizes de covariâncias das componentes. Como não foi imposta nenhuma restrição sobre a forma dessas matrizes, verificou-se a necessidade de conjuntos de treinamento realmente grandes para efetuar suas estimações. No Problema 6, por exemplo, onde foi estimado um modelo com 3 componentes com a dimensão igual a 10, em muitas repetições ocorreram singularidades para as matrizes de covariâncias, mesmo quando tinhamos n_{TRE}=1000. Estavamos lidando com observações geradas computacionalmente, nós conjecturamos que isso pode agravar-se em problemas com dados reais.

Uma altenativa à questão do parágrafo acima, seria impor restrições sobre as matrizes de covariâncias, considerá-las todas iguais ou diagonais. Foram feitos alguns pequenos experimentos exploratórios para essas alternativas e, em geral, verificou-se a necessidade de mais componentes para os modelos, porém, as estimativas do erro de classificação se mantiveram nos mesmos níveis. Observou-se, também, que, com as matrizes diagonais, parece ser necessário mais componentes que com as matrizes iguais. Seriam necessários experi-

mentos mais planejados, no entanto, isso pode indicar que, em problemas reais, é possível empregar modelos homocedásticos, ao custo de um maior número de componentes, sem prejudicar o erro de classificação do método.

5.2 Aplicação com Dados Reais

Nesta seção descrevemos a aplicação do método MFP em um problema de classificação de assinaturas para um indivíduo. O problema é classificar uma dada observação de um vetor de características que corresponde a uma assinatura, ou seja, a observação deve ser classificada como sendo da *classe das falsas* ou da *classe das verdadeiras*. Para aplicações nesse contexto, são considerados dois tipos de erros: o de *Falsa Aceitação* (FA), que consiste em classificar uma assinatura falsa como verdadeira, e o de *Falsa Rejeição* (FR), em que se classifica uma assinatura verdadeira como falsa. A aplicação feita aqui, utiliza um conjunto de dados apresentado em Lee, Berger e Aviczer (1996).

O conjunto de dados utilizados refere-se à assinatura de um indivíduo. O conjunto é composto de 1825 observações, sendo 1000 correspondentes à assinaturas genuínas e 825 à falsificadas. Esses dados foram adquiridos através de processo de *digitalização*, onde as saídas desse processo são dados relativos a uma função temporal bidimensional, digamos $(X(t), Y(t))$, indicando a posição relativa da caneta sobre a superfície da mesa *digitalizadora*. Através de um processo de extração de características, os dados primitivos adquiridos, $\{X(t_l), Y(t_l)\}$, foram transformados em vetores de característias composto de 42 variáveis preditoras, havendo variáveis discretas e contínuas, que correspondem aos aspectos estáticos e dinâmicos da assinatura. Com mencionado, não será discutida a

questão das variáveis selecionadas, assumimos que essas variáveis são suficientemente discriminativas das classes. Os detalhes da obtenção das assinaturas, o processo de extração de características empregado e a descrição das variáveis preditoras, são apresentados em Lee *et al.* (1996).

Para esta aplicação, tinhamos o vetor de características com $d = 42$, o conjunto de treinamento para a classe das verdadeiras (1) com $n_1 = 1000$ e, para classe das falsificadas (2), com $n_2 = 825$. Seguindo a suposição assumida de que o cojunto de treinamento contém a informação sobre as probabilidades a priori das classes (veja Subseção 1.4.3), para essas probabilidades adotamos como estimativa as proporções das classes dentro do total de observações, ou seja, $P_1 = 0,548$ e $P_2 = 0,452$.

Na análise dos dados, primeiramente fizemos um estudo visando a uma redução de dimensão, para isso, seguimos as sugestões dadas em Nelson, Turin e Hastie (1994). Esses autores argumentam que, além da capacidade de discriminação, as características empregadas no problema devem ser insensíveis às variações típicas dentro das assinaturas verdadeiras visando manter o erro de FR relativamente pequeno. Com essa finalidade, os autores propõem uma ordenação das variáveis preditoras no vetor de características, em ordem crescente, segundo seu coeficiente de variação (cv) e, para obter a redução de dimensão, selecionar as d' ($d' < d$) primeiras variáveis que satisfaçam algum critério estabelecido.

Para a redução da dimensão, portanto, foi formado o vetor

$$\Big(cv_{(1)}, cv_{(2)}, cv_{(3)}, ..., cv_{(d)}\Big),$$

onde $cv_{(r)}$ é o r-ésimo menor coeficiente de variação entre os $cv's$ das variáveis preditoras.

Os vetores de características, então, têm suas variáveis preditoras ordenadas segundo a posição dos seu coeficiente de variação nesse vetor. Dessa forma, a primeira variável no vetor de características será aquela com menor cv e, a última, aquela com o maior cv. O coeficiente de variação amostral para a l-ésima variável preditora é definido por $cv_l = \frac{s_l}{\overline{x}_l}$, onde s_l e $\overline{x}_l$ são, respectivamente, o desvio-padrão amostral e a média amostral da l-ésima variável.

Efetuada a ordenação do vetor de características, a redução de dimensão foi efetuada analisando as estimativas dos erros de classificação dos métodos ADL, ADQ e VP, sobre um conjunto de teste. Esses métodos foram escolhidos por serem aqueles que não dependem de ajuste de parâmetros. O procedimento foi o seguinte: (i) foram retiradas aleatoriamente 200 observações, sendo 110 da classe da verdadeiras e 90 das falsificadas para serem as observações de teste; (ii) as 1625 observações restantes foram empregadas para estimar os métodos ADL, ADQ e VP; e (iii) esses métodos foram aplicados para classificar as 200 observações com diferentes dimensões para os dados e determinadas as estimativas do seus erros de classificação. Na Tabela 5.10, são apresentados alguns dos resultados para as estimativas dos erros no conjunto de teste para os métodos considerados, segundo a dimensão das observações. Nessa tabela, temos as taxas dos erros de FR e FA, mensuradas pelas proporções de classificações erradas segundo o conceito desses erros, e o erro total (ET), determinado pelos erros FR e FA poderados pelas probabilidades a priori estimadas.

Dos resultados na Tabela 5.10, consideramos razoável empregar vetores de dimensão $d = 20$. Definida a dimensão dos dados, prosseguimos para selecionar o número de componentes dos modelos para o MFP. Para esse fim, empregamos o conjunto das 1625

Tabela 5.10: Taxas de Erro (%) no Subconjunto de Teste

Método	Erro	Dimensão							
		42	30	20	15	12	10	7	5
ADL	ET	2,01	4,02	5,52	5,02	8,04	7,53	8,54	16,07
	FR	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	FA	4,44	8,89	12.22	11,11	17,78	16,67	18,89	35,56
ADQ	ET	0,00	0,00	0,00	0,00	0,00	0,00	1,00	1,00
	FR	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	FA	0,00	0,00	0,00	0,00	0,00	0,00	2,22	2,22
VP	ET	2,51	1,51	1,00	1,00	1,00	0,50	0,00	1,00
	FR	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,91
	FA	5,56	3,33	2,22	2,22	2,22	1,11	0,00	1,11

Nota: ET=Erro Total, FR=Falsa Rejeição, FA=Falsa Aceitação

observações para determinar os valores dos critérios AIC e BIC, com o número de componentes no modelo em cada uma das classe variando de 1 a 10. Esse procedimento foi repetido quinze vezes, isso representando, quinze inicializações distintas para o algoritmo EM. Nos experimentos iniciais, ocorreram problemas de singularidade para as matrizes de dispersão, em virtude disso, consideramos modelos homocedásticos para cada classe.

Para ilustração, na Tabela 5.11 apresentamos os resultados para a escolha da dimensão dos modelos, os valores na tabela são o número de vezes que o número de componentes foi selecionado pelos critérios em cada classe. Na Figura 5.8, apresentamos um exemplo típico do comportamento dos critérios AIC e BIC de acordo com o número de componentes para

Tabela 5.11: Número de Componentes para os Modelos

Classe	Critério	Número de Componentes									
		1	2	3	4	5	6	7	8	9	10
1	AIC	11	-	-	-	-	2	1	1	-	-
	BIC	12	-	-	-	-	1	1	1	-	-
2	AIC	-	-	-	-	-	3	11	1	-	-
	BIC	-	-	-	-	-	4	10	-	1	-

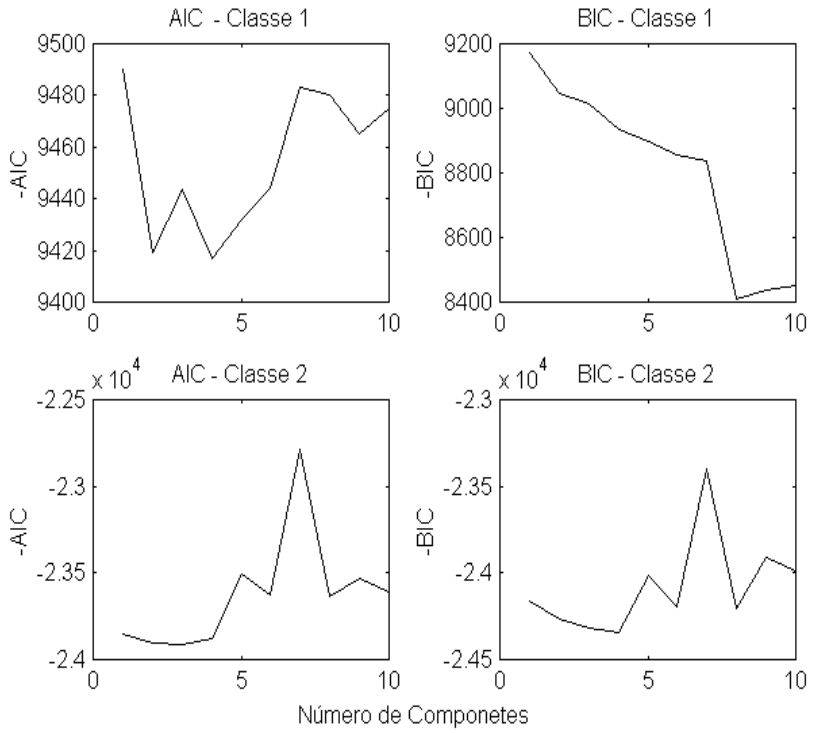
o modelo, em cada uma das classes (para melhor visualização, no gráfico estão os valores de -AIC e -BIC).

De acordo com os resultados na Tabela 5.11, para o modelo na Classe 1 temos $k_1^{(A)} = k_1^{(B)} = 1$ e, para o modelo na Classe 2, $k_2^{(A)} = k_2^{(B)} = 7$, dessa forma, não sendo necessário empregar o modelo ponderado. Definidas as dimensões dos modelos, os parâmetros foram estimados via o algoritmo EM para cada uma das classes, sendo, então, os modelos estimados empregado no método MFP para classificar as 200 observações de teste. Os resultados da classificação do conjunto de teste para o MFP, foram (em %)

$$\text{FR} = 0,00, \quad \text{FA} = 0,00 \quad \text{e} \quad \text{ET} = 0,00.$$

Os resultados observados para os erros de FR e FA, são para um particular conjunto de teste. Com a finalidade de estimar esses erros com maior precisão, realizamos um experimento onde repetimos 100 vezes o procedimento de classificar conjuntos de teste com 200 observações. Em cada repetição, foram selecionadas, aleatoriamente 200 observaões, 110 da classe das verdadeiras e 90 das falsificadas, as quais foram classificadas pelo MFP

Figura 5.8: Critérios AIC e BIC



e pelos outros métodos que estamos considerando, sendo os parâmetros dos métodos estimados com as 1625 observações restantes.

Para o experimento, os valores para os parâmetros dos métodos EFN e kNN, foram selecionados com base no mesmo conjunto de teste empregado para determinar o número de componentes para o MFP. O valores foram, respectivamente, 0,6669 e 3, escolhidos empregando o mesmo procedimento adotado nos problemas com dados simulados.

Na Tabela 5.12, temos os resultados para os erros de FR e de FA observados no experimento, são apresentadas a média, o desvio-padrão (DP) e o intervalo de confiança para

Tabela 5.12: Taxas de Erro (%) nas 100 Repetições

Método	Erro	Taxas (%)		
		Média	DP	IC (95%)
ADL	FR	0,00	0,00	-
	FA	13,48	3.16	[12,86;14,11]
ADQ	FR	0,00	0,00	-
	FA	1,02	1,14	[0,79;1,24]
EFN	FR	0,00	0,00	-
	FA	1,21	0,48	[1,12;1,31]
kNN	FR	0,07	0,24	[0,02;0,12]
	FA	8,67	3,02	[8,07;9,27]
MFP	FR	0,00	0,00	-
	FA	1,03	1,14	[0,80;1,26]
VP	FR	0,10	0,31	[0,03;0,16]
	FA	3,93	1,95	[3,54;4,32]

Nota: FR=Falsa Rejeição, FA=Falsa Aceitação

média (IC), a 95%, das 100 repetições. Dessa tabela, vemos que, à exceção do kNN e VP, os métodos classificaram corretamente as assinaturas verdadeiras em todas as repetições, ou seja, a média para o erro de FR foi igual a 0,00. Por outro lado, todos o métodos apresentaram erros ao classificar as assinaturas falsas, isto é, todos com a média do erro de FA maior que 0,00. Pode ser visto, também, considerando os dois tipos de erros, os métodos ADQ , EFN e o MFP, apresentaram os melhores resultados, somente com erros de FA, cujos valores das médias não têm diferenças significativas.

Os resultados na Tabela 5.12, indicam uma mesma tendência para os erros, as médias para o erro de FR são pequenas, enquanto para o de FA, são sempre maiores com todos os métodos. Isso pode estar sugerindo, que os métodos modelaram de forma adequada a distribuiçào na classe das verdadeiras, encontrando uma maior dificuldade com a distribuição na classe das falsificadas. É possível que isso seja decorrente de uma maior dispersão para as variáveis preditoras dentro da classe das assinaturas falsificadas, o que seria uma comprovação da conjectura a esse respeito mencionada anteriormente na seção anterior.

Os resultados apresentados pelo ADQ e MFP, parecem indicar que a distribuição normal seria compatível com a distribuição condicional da classe das assinaturas verdadeiras. Isso também sugere, que o procedimento de seleção do número de componentes no MFP, foi eficiente neste problema. Com relação a distribuição na classe das assinaturas falsificadas, a tendência demonstrada por todos os métodos para o erro de FA, pode estar indicando uma limitação própria desse conjunto de dados.

Embora não tenha sido empregado neste trabalho, nos problemas de classificação de assinaturas, é adotado um limiar para os erros FA e FR de modo que se obtenha uma relação otimizada entre esses erros. Esse limiar depende dos objetivos da classificação, em que poderia ser fixado um valor máximo para um desses erros e os modelos seriam selecionados procurando minimizar o outro erro, porém respeitando a restrição imposta. Veja as discussões sobre essas questões em Nelson *et al.* (1994) e em Lee *et al.* (1996).

Capítulo 6

Conclusões e Sugestões

Neste capítulo, discutimos os principais resultados obtidos para a proposta apresentada neste trabalho. São feitas sugestões de para trabalhos futuros, em continuidade a pesquisa aqui desenvolvida.

6.1 Discussões e Conclusões

Na modelagem estatistica para os problemas de Reconhecimento de Padrões Supervisionado, considerando a aprendizagem Informativa, torna-se necessário a estimação da distribuições condicionais para as classes consideradas no problema. Para isso, podemos empregar modelos paramétricos ou não-paramétricos. Neste trabalho, no entanto, foi proposto o emprego de misturas finitas de densidades, em particular, *misturas finitas de*

densidades normais, para aproximar essas densidades condicionais.

No desenvolvimento do trabalho, inicialmente, fizemos uma revisão da literatura sobre os métodos estatísticos paramétricos e não-paramétricos, visando caracterizá-los por suas qualidades e deficiências, considerando os aspectos teóricos e práticos. Foi feita uma revisão da teoria das misturas finitas de densidades, abordando as questões de estimação dos parâmetros em um modelo com o número de componentes fixo, enfatizando a estimação de máxima verossimilhança, e a determinação de um número de componentes para o modelo baseada nas observações amostrais. A determinação das estimativas de máxima verossimilhança via o algoritmo EM, também foi abordada, onde dicutimos algumas questões teóricas e outras relativas a implementação desse algoritmo.

A motivação da proposta estava baseada, num primeiro instante, em que qualquer distribuição contínua pode ser aproximada, com precisão arbitrária, por um de modelo de mistura de densidades normais. Para esse fim, em teoria, é necessário um número muito grande de componentes para modelo. Num segundo instante, foram levados em conta os aspectos práticos dos problemas em RPS, em que pesquisadores da área argumentam ser necessário apenas uma "boa" aproximação para as densidades condicionais para obter-se regras de classificação com um desempenho satisfatório. Essa "boa" aproximação, pode implicar em um número menor de componentes, desde que haja uma escolha apropriada das componentes para representar com maior precisão as regiões do suporte da distribuição verdadeira importantes para o objetivo da modelagem. Com essas considerações em mente, a proposta foi empregar *misturas finitas de normais com a determinação do número de componentes direcionada pelas próprias observações da classe no conjunto de treinamento*. Como um meio para determinar o número de componentes adequado aos

dados, foram empregados o *Critério de Informação de Akaike* (AIC) e o *Critério de Informação Bayesiano* (BIC) e, havendo determinações diferentes, adotamos como modelo uma *ponderação dos modelos indicados por esses dois critérios*. Pelo nosso conhecimento, essa abordagem não havia sido mencionada na literatura.

Com a proposta de modelagem das densidades condicionais, foi contruída uma regra de classificação, sendo uma versão empírica da regra de Bayes, que denominamos *metodo de Mistura Finitas Ponderadas* (MFP). Foi mostrado que, sob algumas condições de regularidades, essa regra é *consistente para o risco de Bayes*, inclusive, quando empregada a ponderação dos modelos selecionados pelo AIC e BIC. Entre as condições de regularidade, está a da distribuição verdadeira ser uma mistura, embora isso possa parecer restritivo, a classe das distribuições que podem ser geradas por misturas é extremamente abrangente, suficientemente grande para incluir as distribuições paramétricas ordinárias.

Realizamos experimentos computacionais visando avaliar alguns apectos do método proposto, tendo como referência, os principais métodos estatísticos para RPS. Dos resultados desses experimentos, verificou-se que o MFP modelou distribuições bastante complexas, onde os métodos paramétricos não eram adequados, com resultados estatisticamente equivalentes aos dos métodos não-paramétricos. Quando utilizada a ponderação do modelo, os resultados indicaram um desempenho estatisticamente equivalente ao do modelo que, em teoria, era o correto.

O método foi aplicado a um problema com dados reais, referentes a assinaturas de um indivíduo. O problema era classificar uma dada observação de um vetor de características relativo a uma assinatura como sendo de uma assinatura verdadeira ou de uma assinatura falsificada. Neste problema, o resultados do MFP foram estatisticamente equivalentes aos

dos outros métodos estatísticos com os melhores desempenhos. Em particular, verificou-se que os resultados indicaram que o procedimento de seleção do número de componentes se mostrou eficiente para o problema. Temos, portanto, que o método proposto modelou as distribuições das classes de forma satisfatória, tendo em vista as pequenas taxas de erros observadas.

O procedimento de classificação proposto aqui, pode ser efetuado com o emprego de uma rede neural de funções de base radial (RNFBR), explorando a capacidade de modelar superfícies discriminantes complexas das redes neurais artificiais. O critério para a determinação da dimensão dos modelos, pode ser empregado para selecionar o número de funções de base na rede. Existem, no entanto, algumas diferenças entre as abordagens pelo MFP e por RNFBR. No MFP temos um certo controle sobre as componentes no modelo, no sentido de que, sendo possível atribuir alguma significado físico a essas componentes, podemos identificar e analisar essas componentes separadamente. Esse situação se aplica aos casos em que sabemos que as classes são compostas por subclasses. Outra diferença, é o fato de que, uma vez estimados os modelos para o MFP, será necessário apenas armazenar as estimativas dos parâmetros para posteriores utilizações da regra de classificação, ou seja, é feita uma "sumarização" das observações.

As dificuldades com o MFP, basicamente, são relativas a "lentidão" de convergência do algoritmo EM e as dificuldades com relação a estimação das matrizes de covariâncias das componentes, principalmente, não havendo restrições nessas matrizes. Essas questões, limitam a emprego do método em problemas com dados de alta dimensão e conjunto de treinamento relativamente pequenos e, também, em aplicações *on line*.

Nós acreditamos que existam razões teóricas e empíricas suficientes para que a proposta

apresentada seja eficiente para os problemas de RPS. O que foi feito, nós consideramos apenas como um passo inicial de pesquisa, mais investigações são necessárias visando superar as dificiências observadas e buscar melhoramentos para o procedimento. Na seção a seguir, apresentamos algumas sugestões para a continuidade da pesquisa nessa direção.

6.2 Sugestões para a Continuidade da Pesquisa

Nas avaliações feitas com relação ao método proposto, exploramos a aplicação do aprendizado e não o aprendizado em si mesmo. No entanto, um dos passos necessários em busca de melhorar o desempenho do método, seria investigar sua capacidade de modelagem propriamente dita, avaliando os aspectos do procedimento como estimador de densidades pertinentes aos objetivos de classificação. Nesse sentido, poderiam ser empregados os modelos para gerar observações apresentados em Marron e Wand (1992), investigando meios de gerar observações de alta dimensão correlacionadas com esses modelos e empregando-os para avaliar o erro de estimação da densidade, com o erro médio quadrático, por exemplo, comparando os modelos determinados pelo AIC, pelo BIC e o ponderado.

Outro aspecto fundamental, seria implementar o procedimento com versões do EM que procuram aumentar a taxa de convergência desse algoritmo. Propostas desse tipo são apresentadas em Jamshidian e Jennrich (1997) e em Liu *et al.* (1998).

Podem ser consideradas alternativas à ponderação que empregamos aqui. Na literatura, há propostas de combinação de classificadores que resultam em um classificador tão bom

quanto os classificadores combinados, em particular, sendo os classificadores combinados consistentes, o classificador resultante será consistente. Propostas desse tipo são apresentadas, por exemplo, em Mojirsheibani (1999) e Mojirsheibani (2000). Nesse sentido, poderíamos considerar o MFP, com os modelos selecionados pelo AIC e pelo BIC, como classificadores distintos, utilizar essas propostas de combinação e comparar com a ponderação que empregamos.

Apêndice A

Definições e Conceitos

A.1 Algumas Formas de Convergência

Considere que Ω é um conjunto de pontos, $\mathcal{A}$ é uma σ-álgebra de subconjuntos de Ω e ν uma medida definida sobre $\mathcal{A}$, dessa forma, $(\Omega, \mathcal{A}, \nu)$ denota um *espaço de medida.* Diremos que uma função mensurável $f \in L(\Omega, \mathcal{A}, \nu)$ se

$$\int_{\Omega} |f| \, d\nu < \infty$$

Definição A.1.1 *Seja f uma função mensurável sobre $(\Omega, \mathcal{A}, \nu)$ e considere uma sequência de funções $f_1, f_2, f_3, ... \in L(\Omega, \mathcal{A}, \nu)$. Diremos que a sequência $\{f_n\}$ converge para f em* **norma $\mathbf{L_1}(\nu)$** *se, e somente se,*

$$\int_{\Omega} |f_n - f| \, d\nu \to 0 \quad quando \, n \to \infty.$$

Nas definições a seguir, a medida ν é uma medida de probabilidade e a denotaremos por P, assim, $(\Omega, \mathcal{A}, P)$ denota um *espaço de probabilidade.*

Definição A.1.2 *Seja* $\{X_n\}$ *uma sequência de variáveis aleatórias e* X *uma variável aleatória, todas definidas em* $(\Omega, \mathcal{A}, P)$*. Essa sequência converge para* X **em probabilidade** *se,* $\forall\, \epsilon > 0$*,*

$$\lim_{n\to\infty} Pr(|X_n - X| < \epsilon) = 1$$

Definição A.1.3 *Considere uma sequência de variáveis aleatórias* $\{X_n\}$ *e uma variável aleatória* X*, todas definidas em* $(\Omega, \mathcal{A}, P)$*. Essa sequência converge para* X **com probabilidade 1***, ou* **em quase toda parte** *(***q.t.p.***), se*

$$Pr(\lim_{n\to\infty} X_n = X) = 1$$

Definição A.1.4 *Sejam* $X_1, X_2, X_3, \ldots.$ *e* X *variáveis aleatórias, não necessariamente no mesmo espaço de probabilidade, e* $F_1, F_2, F_3, \ldots.$ *e* F*, respectivamente, suas funções de distribuição. Dizemos que a sequência* $\{X_n\}$ *converge* **em distribuição** *para* X *se*

$$\lim_{n\to\infty} F_n(t) = F(t)$$

para cada t *ponto de continuidade de* F*.*

A.2 Propriedades de Estimadores

Nas definições que seguem, considere uma família de espaços de probabilidade $(\Omega, \mathcal{A}, P_\theta)$, indexada pelo parâmetro $\theta \in \Theta$, e que $X^1, X^2, X^3, \ldots, X^n$ é uma *amostra aleatória* de

alguma distribuição específica (porém desconhecida) P_θ dessa família. Um *estimador* é uma estatística $\widehat{\theta}_n = \widehat{\theta}_n(X^1, X^2, X^3, \ldots, X^n)$, sendo, portanto, uma variável aleatoria, empregada para estimar o valor do parâmetro desconhecido θ.

Definição A.2.1 *Seja $\widehat{\theta}_n$ um estimador para um parâmetro desconhecido θ. O estimador $\widehat{\theta}_n$ é dito ser* **não-tendencioso** *para θ, se $E_{\mathcal{D}}(\widehat{\theta}_n) = \theta$, qualquer que seja $\theta \in \Theta$, com o valor esperado determinado com a distribuição $\mathcal{D}$ de $\widehat{\theta}_n$.*

Definição A.2.2 *Seja $\{\widehat{\theta}_n\}_{n\geq 1}$ uma sequência de estimadores para um parâmetro desconhecido θ. A sequência $\{\widehat{\theta}_n\}_{n\geq 1}$ é dita ser* **consistente** *se ela converge em probabilidade para θ. Se a convergência é com probabilidade 1, então, é dito que $\{\widehat{\theta}_n\}_{n\geq 1}$ é* **fortemente consistente**.

Definição A.2.3 *Considere agora que em P_θ temos $\theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)$. A* **matriz de Informação de Fisher**, *$I(\theta)$, é uma matriz $m \times m$ definida por*

$$I(\theta) = \int_{\Re^n} [\nabla_\theta \ln f(\mathbf{x};\theta)][\nabla_\theta \ln f(\mathbf{x};\theta)]^T f(\mathbf{x};\theta)\, d\mu,$$

sendo, portanto, cada elemento da forma

$$I(\theta)_{ij} = E[\frac{\partial}{\partial\theta_i} \ln f(\mathbf{x};\theta) \frac{\partial}{\partial\theta_j} \ln f(\mathbf{x};\theta)].$$

Definição A.2.4 *Uma sequência de estimadores $\{\widehat{\theta}_n\}_{n\geq 1}$ para um parâmetro desconhecido θ é dita ser* **assintoticamente eficiente** *se $\sqrt{n}(\widehat{\theta}_n - \theta)$ converge em distribuição para uma $N(0, [I(\theta)]^{-1})$, onde $I(\theta)$ é a matriz de informação de Fisher.*

A.3 Estatística Suficiente e Família Exponencial

Nesta seção, consideramos o conceito de *estatística suficiente* e o de uma classe de distribuições denominada de *família exponencial*. Como na seção anterior, temos uma família de espaços de probabilidade $(\Omega, \mathcal{A}, P_\theta)$, uma amostra aleatória $X^1, X^2, X^3, \ldots, X^n$ de uma distribuição P_θ nessa família, com θ desconhecido.

Definição A.3.1 *Uma estatística $T = T(X^1, X^2, X^3, \ldots, X^n)$ é dita ser* **suficiente** *para a família $\{P_\theta; \theta \in \Omega\}$ se temos que a distribuição condicional*

$$f_{X^1, X^2, X^3, \ldots, X^n | T=t}$$

não depende de θ.

A definição A.3.1 diz que a distribuição condicional da amostra, dado o valor da estatística, é independente de θ. Em outras palavras, uma estatística suficiente contém toda a informação sobre θ, no sentido de que, se conhecemos o valor da estatística, as observações na amostra nada têm a informar sobre o parâmetro.

Definição A.3.2 *Uma família de distribuições $\{P_\theta; \theta \in \Omega\}$, onde $\theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)$, é dita pertencer a* **família exponencial de m-parâmetros** *se P_θ tem densidades da forma*

$$p_\theta(x) = \frac{1}{a(\theta)} b(x) \exp\{c(\theta)^T d(x)\}.$$

As distribuições na classe da família exponencial, têm propriedades bem definidas com relação a estimação dos parâmetros. Temos, por exemplo, que as distribuições nessa

família são continuamente diferenciáveis com relação aos seus parâmetros, a matriz de informação de Fisher é definida positiva e as equações de máxima verossimilhança têm, no máximo, uma raíz no interior do espaço dos parâmetros que, quando existe, maximiza a função de verossimilhança, ou seja, é o EMV.

Considere uma família paramétrica de densidades $\{g(x;\theta) : \theta \in \widetilde{\Theta} \subseteq \Re^m\}$ (com respeito a uma medida ν sobre $\Re^d$) que pertence a família exponencial, então, podemos escrever

$$g(x;\theta) = \frac{1}{a(\theta)} b(x) \exp\{c(\theta)^T t(x)\}, \tag{A.1}$$

onde $b : \Re^d \longrightarrow \Re$, $t : \Re^d \longrightarrow \Re^m$ e $a(\theta)$ é uma constante de normalização dada por

$$a(\theta) = \int_{\Re^d} b(x) \exp\{c(\theta)^T t(x)\}\, d\nu.$$

Pode ser provado que as estatísticas em t, $t = (t_1, t_2, t_3, \ldots, t_m)$, são conjuntamente suficientes para $g(\cdot;\theta)$ e, quando existe o EMV, esse estimador é função dessas estatísticas.

A representação da família exponencial como dada em (A.1) é denominada *parametrização natural*. Existe outra forma de representação, denominda *parametrização de valor médio*, que é dada em termos de um parâmetro

$$\phi = E(t(X)|\theta) = \int_{\Re^d} t(x) g(x;\theta)\, d\nu,$$

sob a condição que $\widetilde{\Theta}$ seja aberto e convexo. A parametrização de valor médio é discutida em Redner e Walker (1984) e, uma observação importante sobre esta parametrização, é que a atribuição $\theta \to \phi = E(t(X)|\theta)$ é um mapeamento biunívoco de $\widetilde{\Theta}$ para um conjunto aberto Θ .

Sob as condições estabelecidas na parametrização de valor médio, os membros da família exponencial podem ser representados como

$$h(x;\phi) = g(x;\theta(\phi)) = \frac{1}{a(\theta(\phi))} b(x) \exp\{\theta(\phi)^T t(x)\} \quad x \in \Re^d, \tag{A.2}$$

para $\phi \in \Theta$, onde $\theta(\phi)$ satisfaz $\phi = E(t(X)|\theta(\phi))$.

As distribuições da família exponencial são muito importantes tanto pelas suas propriedades, como pelo fato de que a maioria das distribuições empregadas na prática pertencem a essa família. As distribuições binomial, Poisson, normal e gama são alguns exemplos de distribuições da família exponencial. Para uma descrição mais completa da família exponencial e a discussão de suas propriedades quanto a estimação de máxima verossimilhança sugerimos Sundberg (1974) e Barndorff–Nielsen e Cox (1994).

Bibliografia

Aitkin, M. e Rubin, D. B. (1985). Estimation and Hypotheses Testing in Finite Mixture Models. *Journal of the Royal Statistical Society Series B*, **47**(1), 67–75.

Ash, R. B. (1972). *Real Analysis and Probability*. Academic Press, INC, New York.

Banfield, J. D. e Raftery, A. E. (1993). Model–Based Gaussian and Non–Gaussian Clustering. *Biometrics*, **49**, 803–821.

Barndorff–Nielsen, O. E. e Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, New York.

Basford, K. E. e McLachlan, G. J. (1985). Likelihood Estimation with Normal Mixture Models. *Applied Statistic*, **34**(3), 282–289.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* Second Edition. Springer–Verlag, New York.

Biernacki, C., Celeux, G. e Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

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

Boyles, R. A. (1983). On the Convergence of the EM Algorithm. *Journal of the Royal Statistical Society Series B*, **45**(1), 47–50.

Bozdogan, H. (1987). Model Selection and Akaike's Information Criterion (AIC): General Theory and Its Analytical Extensions. *Psychometric*, **52**(3), 345–370.

Breiman, L., Friedman, J. H., Olshen, R. A. e Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, California.

Celeux, G. e Soromenho, G. (1996). An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. *Journal of Classification*, **13**, 195–175.

Cooley, C. A. e MacEachern, S. N. (1998). Classification via Kernel Product Estimators. *Biometrika*, **85**(4), 823–833.

Cover, T. M. e Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, **IT-13**(1), 21–27.

Dempester, A. P., Laird, N. M. e Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**(1), 1–38.

Devroye, L. e Györfi, L. (1985). *Nonparametric Density Estimation. The L_1 View*. John Wiley, New York.

Devroye, L., Györfi, L. e Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Duda, R. O. e Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley, New York.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**, 179–188.

Fraley, C. e Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers Via Model–Based Cluster Analysis. *Technical Report No. 329, Department of Statistics.*University of Washington.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Second Edition. Academic Press, New York.

Gamerman, D. (1996). *Simulação Estocástica Via Cadeias de Markov*. ABE, 12º Simpósio Nacional de Probabilidade e Estatística, Caxambu,MG.

Gilks, W. R., Richardson, S. e Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice* . Chapman & Hall, New York.

Graybill, F. A. (1969). *Matrices with Applications in Statistics*. Wadsworth Intern. Group, California.

Guyon, I., Makhoul, J., Schwartz, R. e Vapnick, V. (1998). What Size Test Set Gives Good Error Rate Estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(1), 52–64.

Hand, D. J. (1981). *Discrimination and Classification*. John Wiley, New York.

Hand, D. J. (1982). *Kernel Discriminant Analysis*. Research Studies Press, New York.

Hastie, T. e Tibshirani, R. (1996). Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society Series B*, **58**(1), 155–176.

Hathaway, R. J. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics*, **13**(2), 795–800.

Holmström, L., Koinstinen, P., Laaksonen, J. e Oja, E. (1997). Neural and Statistical Classifiers–Taxonomy and Two Case Studies. *IEEE Transactions on Neural Networks*, **8**(1), 5–17.

Jain, A. K. e Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

Jain, A. K., Duin, R. P. W. e Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(1), 4–37.

Jamshidian, M. e Jennrich, R. I. (1997). Acceleration of the EM Algorithm by Using Quasi-Newton Methods. *Journal of the Royal Statistical Society Series B*, **59**(3), 569–587.

Kass, R. E. e Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Kurita, T., Otsu, N. e Abdelmalek, N. (1992). Maximum Likelihood Thresholding Based on Population Mixture Models. *Pattern Recognition*, **25**(10), 1231–1240.

Lavine, M. e West, M. (1992). A Bayesian Method for Classification and Discrimination. *The Canadian Journal of Statistics*, **20**(4), 451–461.

Lee, L. L., Berger, T. e Aviczer, E. (1996). Reliable On-Line Human Signature Verification Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**(6), 643–647.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Second Edition. John Wiley, New York.

Lehmann, E. L. (1999). *Elements of Large–Sample Theory*. Springer, New York.

Leroux, B. G. (1992). Consistent Estimation of a Mixing Distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Lindsay, B. G. (1983). The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics*, **11**(1), 86–94.

Liu, C., Rubin, D. B. e Wu, Y. N. (1998). Parameter Expansion to Accelerate EM: The PX-EM Algorithm. *Biometrika*, **85**(4), 755–770.

Mardia, K. V., Kent, J. T. e Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.

Marron, J. S. e Wand, M. P. (1992). Exact Means Integrated Squared Error. *The Annals of Statistics*, **20**(2), 712–736.

McLachlan, G. J. e Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

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

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley, New York.

McLachlan, G. J. e Peel, D. (1997). On a Resampling Approach to Choosing the Number of Components in Normal Mixture Models. Em Billard, L. e Fisher, N. I. (Eds.). *Computing Science and Statistics,* Vol. 28, pag. 260-266. Fairfax Station, Virginia.

Meng, X. L. e van Dyk, D. (1997). The EM Algorithm–An old folk song sung to a fast new tune. *Journal of the Royal Statistical Society Series B*, **59**(3), 511–567.

Mojirsheibani, M. (1999). Combining Classifiers via Discretization. *Journal of the American Statistical Association*, **94**(446), 600–609.

Mojirsheibani, M. (2000). A Kernel-based combined classification rule. *Statistics & Probability Letters*, **48**, 411–429.

Moon, T. K. (1996). The Expectation–Maximization Algorithm. *IEEE Signal Processing Magazine*, **November**, 47–96.

Nelson, W., Turin, W. e Hastie, T. (1994). Statistical Methods for On-Line Signature Verification. *Intern. Journal of Pattern Recognition and Artificial Intelligence*, **8**(3), 749–770.

Oliver, L. H., Pulsen, R. S., Toussaint, G. T. e Louis, C. (1979). Classification of Atypical Cells in the Automatic Cytoscrening for Cervical Cancer. *Pattern Recognition*, **11**, 205–212.

Peters, B. C. e Walker, H. F. (1978). An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM Journal on Applied Mathematics*, **35**(2), 362–378.

Pollard, D. (1981). Strong Consistency of k-Means Clustering. *The Annals of Statistics*, **19**(1), 135–140.

Polymenis, A. e Titterington, D. M. (1998). On the Determination of the Number of Components in a Mixture. *Statistics & Probability Letters*, **38**, 295–298.

Popat, K. e Picard, R. W. (1997a). Cluster-Based Probability Model and its Applications to Image and Texture Processing. *Media Laboratory Perceptual Computing Section Technical Report No. 351.* Massachusetts Institute of Technology (MIT).

Popat, K. e Picard, R. W. (1997b). Cluster-Based Probability Model and Its Applications to Image and Texture Processing. *IEEE Transactions on Image Processing*, **6**(2), 268–284.

Priebe, C. E. (1993). Adaptive Mixtures Density Estimation. *Pattern Recognition*, **26**(5), 771–785.

Priebe, C. E. (1994). Adaptive Mixtures. *Journal of the American Statistical Association*, **89**(427), 796–806.

Redner, R. (1981). Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions. *The Annals of Statistics*, **9**(1), 225–228.

Redner, R. A. e Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, **26**(2), 195–239.

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

Roeder, K. (1990). Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. *Journal of the American Statistical Association*, **85**(411), 617–624.

Roeder, K. (1992). Semiparametric Estimation of Normal Mixture Densities. *The Annals of Statistics*, **20**(2), 929–943.

Roeder, K. e Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixture of Normals. *Journal of the American Statistical Association*, **92**(439), 894–902.

Rubinstein, Y. D. e Hastie, T. (1997). Discriminative vs Informative Learning. *http://www-stat.stanford.edu/∼hastie/Papers*.

Rudin, W. (1976). *Principles of Mathematical Analysis.* Third Edition. McGraw–Hill, New York.

SAS (1988). *SAS INSTITUTE INC. SAS/IML. User's Guide, Release 6.03 Edition*. SAS INSTITUTE INC, Cary, NC.

SAS (1989). *SAS INSTITUTE INC. SAS/STAT. User's Guide, Version 6, Fourth Edition*. SAS INSTITUTE INC, Cary, NC.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**(2), 461–464.

Sclove, S. L. (1983). Application of the Conditional Population-Mixture Model to Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**(4), 428–433.

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

Seber, G. A. F. (1984). *Multivariate Observations*. John Wiley, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Solka, J. L., Wegman, E. J., Priebe, C. E., Poston, W. L. e Rogers, G. W. (1998). Mixture Structure Analysis using the Akaike Information Criterion and the Bootstrap. *Statistics and Computing*, **8**, 177–188.

Soromenho, G. (1994). Comparing Approaches for Testing the Number of Componentes in Finite Mixture Model. *Computational Statistics*, **9**, 65–78.

Sundberg, R. (1974). Maximum Likelihood Theory for Incomplete Data from an Exponential Family. *Scandinavian Journal of Statistics*, **1**(2), 49–58.

Teicher, H. (1963). Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, **34**, 1265–1269.

Titterington, D. M., Smith, A. F. M. e Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

Tråvén, H. G. C. (1991). A Neural Network Approach to Statistical Pattern Classification by 'Semiparametric' Estimation of Probability Density Functions. *IEEE Transactions on Neural Networks*, **2**(3), 366–377.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, **11**(1), 95–103.

Yakowitz, S. J. e Spragins, J. D. (1968). On the Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, **39**(1), 209–214.

Zhu, Q. e Cai, Y. (1998). A Subclass Model for Nonlinear Pattern Classification. *Pattern Recognition Letters*, **19**(2), 19–29.