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

APLICAÇÕES ESTATÍSTICAS EM RECONHECIMENTO DE PADRÕES COM ÊNFASE EM VERIFICAÇÃO DE ASSINATURAS

Manoel Raimundo de Sena Junior

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

Este exemplar corresponde à redação final da tese defendida por MANOEL RAIMUNDO DE SENA JUNIOR e aprovada pela Comissão Julgadora em 31/3/97
Orientador

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.

Março - 1997

Campinas - SP.

UNICAMP
BIBLIOTECA CENTRAL

UNIDADE BC
N.º CHAMADA :
T/UNICAMP
Se55a
V. Ex.
TOMBO BC/ 30382
PROC. 281/97
C [ ] D [X]
PREÇO R$ 11,00
DATA 22/05/97
Nº CPD

CM-00099349-0

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

Se55a

Sena Júnior, Manoel Raimundo de
Aplicações estatísticas em reconhecimento de padrões com ênfase em verificação de assinatura / Manoel Raimundo de Sena Júnior.--Campinas, SP: [s.n.], 1997.

Orientador: Lee Luan Ling.
Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Reconhecimento de padrões. 2. Estatística - Classificação. 3. Análise multivariada. I. Lee, Luan Ling. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Comptação. III. Título.

À Claudia Regina,
Leonam e Arani,
pelo apoio
incondicional.

Agradecimentos

- Ao Professor Doutor Lee Luan Ling, pela sua orientação eficiente e segura, com quem pude aprender e cultivar uma grande amizade, que, juntamente com seu conhecimento, foram fatores decisivos na realização deste trabalho.

- Aos Professores Doutores Dalton Francisco de Andrade, Clarice Azevedo de Luna Freire, Jaime Portughesis e Reginaldo Palazzo Júnior pelas valiosas sugestões.

- Aos Professores Doutores do DE-UFSCar: Carlos Alberto Ribeiro Dinis e Luis Aparecido Milan, pelo incentivo e apoio ao me apontar novos e seguros caminhos.

- À minha esposa Prof*a*. Dr*a*. Claudia Regina Oliveira de Paiva Lima, pela inesgotável paciência, compreensão e carinho nestes anos todos, além de um decisivo apoio profissional.

- Aos meus familiares em nome de minha mãe Diva, pelo amor constante.

- Aos colegas do curso de Doutorado, pelo companheirismo e ajuda mútua.

- Aos amigos e colegas do Departamento de Estatística da UFPb, que possibilitaram a realização deste trabalho e a credibilidade em mim depositada, em especial ao Prof. Dr. João Agnaldo do Nascimento.

- Aos amigos que me acompanharam e me incetivaram desde o princípio em São Paulo: Antônio, Filidor, Manuel, Miguel, Oscar, Thelma, Wânia e notadamente a família de Romeu Ignacchitti.

- À UFPb pela licença concedida.

- Ao apoio financeiro da CAPES/PICD.

Resumo

Neste trabalho apresentamos um estudo estatístico sobre seleção de variáveis e classificação de objetos em relação a dois padrões.

Descrevemos alguns métodos de classificação conhecidos na literatura, construídos com base nas observações dos dois padrões. Dentre eles destacamos a classificação via Análise Discriminante e Modelos de Regressão Logística. Mais resumidamente descrevemos o Classificador de Bayes e um outro classificador, baseado nas observações de apenas um padrão fixado, que utiliza uma aproximação da distância de um objeto ao centro da distribuição normal, pela distribuição Qui-quadrado. Introduzimos um novo classificador, baseado apenas nas observações de um padrão fixado, que utiliza uma aproximação da distância de Mahalanobis de um objeto ao centro da distribuição normal, pela distribuição T^2 de Hotelling. Comparamos seu desempenho, através de um estudo de simulação, com os demais classificadores em termos de taxa de classificação errônea.

Descrevemos também alguns métodos de seleção de características significantes com base nos respectivos classificadores, além de outros métodos de seleção conhecidos na literatura, construídos com base nas observações apenas de um padrão fixado. Introduzimos um novo método de seleção com base no novo classificador. Um estudo de simulação para comparar o desempenho dos métodos de seleção apresentados foi realizado, onde o processo de classificação utiliza apenas as características significantes. O desempenho de cada método também foi avaliado em termos de classificação errônea.

Uma atenção especial foi dada ao problema de verificação de assinaturas, através de exemplos envolvendo assinaturas dinâmicas e estáticas.

Abstract

In this work, we describe and compile statistical classification methods about feature selection and object recognition for the case of two pattern classification.

First of all, we describe some well-known statistical classification methods when observed samples of both patterns are avaliable. We provide some special attention to the description of the Discriminant Analysis and the Logistic Regression Models. A brief description of the Bayes classification method is provided. We also analyse the case when only one pattern is observable. In this case, the distance measure between the input feature vector and reference vector is assumed to have Qui-squared distribution. We also introduce a novel statistical classification method for the one-pattern observable case, by considering the distance measure as being the T^2 Hotelling distribution. Performance comparison in terms of misclassification error is done via simulation.

Next, we describe some, well-known, selection methods in the literature and suggest some feature selection strategies based on the classification methods previously described. Simulation investigation is performed to evaluate the classification methods using only significant selected features in terms of misclassification error.

Finally, we apply the proposed feature selection and classification methods to signature verification problems, involving both static and dynamic human signatures.

Conteúdo

1 Introdução **1**

1.1 Histórico 1

1.2 Objetivo do Trabalho 3

1.2 Resumo do Conteúdo 5

2 Classificação de Padrões **7**

2.1 Introdução 7

2.2 Análise Discriminante (AD) 9

2.2.1 Introdução 9

2.2.2 Limites de separação de classes 10

2.2.3 Classificador de máxima verossimilhança sob normalidade 12

2.2.4 Função discriminante linear de Fisher 17

2.2.5 Classificador a partir de dados amostrais 18

2.3 Modelo de Regressão Logística (MRL) 23

2.3.1 Introdução 23

2.3.2 Modelo de classificação com resposta binária 25

2.3.3 Estimadores de máxima verossimilhança para o MRL 29

2.3.4 Generalidade da classificação usando MRL 33

2.4 Modelo de Classificação Usando apenas uma das Classes 35

2.4.1 Introdução 35

2.4.2 Estimação de parâmetros sob normalidade 36

2.4.3 Estimação amostral de parâmetros 39

2.5 Aplicação com Dados Simulados 46

2.5.1 Introdução 46

2.5.2 O Classificador de Bayes 46

2.5.3 Classificador usando aproximação pela distribuição Qui-quadrado .. 48

2.5.4 Simulação dos dados 49

2.5.2 Resultados da simulação 51

2.6 Comentários e Conclusões 56
3 Seleção de Variáveis Explicativas 57
3.1 Introdução 57
3.2 Seleção de Variáveis Através de Análise Discriminante 59
3.2.1 Introdução 59
3.2.2 A distribuição da distância Studentizada D^2 60
3.2.3 Escolha de variáveis discriminantes 64
3.3 Seleção de Variáveis Através de Modelos de Regressão Logística 69
3.3.1 Introdução 69
3.3.2 Teste para ajuste do Modelo de Regressão Logística 70
3.3.3 Procedimentos de seleção sistemática 71
3.4 Seleção de Variáveis Usando Observações Apenas de Uma das Classes 73
3.4.1 Introdução 73
3.4.2 Distribuição da distância D^2 restrita a uma das classes 74
3.4.3 Seleção de variáveis discriminates 76
3.5 Outros Métodos de Seleção de Variáveis 80
3.5.1 Introdução 80
3.5.2 Seleção de variáveis através do Coeficiente de Variação (CV) 81
3.5.3 Seleção de variáveis através de Análise de Componentes Principais (CP) 82
3.5.4 Seleção de variáveis através do Coeficiente de Correlação Múltipla (CR) 86
3.6 Testes Experimentais com Dados Simulados 90
3.6.1 Introdução 90
3.6.2 Resultado da seleção e classificação com dados simulados 91
3.7 Cometários e Conclusões 97
4 Aplicação em Verificação de Assinaturas 99
4.1 Introdução 99

4.2 Aplicação de Técnicas de Seleção de Variáveis e Classificação de Assinaturas: Dinâmicas e Estáticas 101
4.2.1 Introdução 101
4.2.2 Descrição dos resultados 102
4.2.3 Análise dos resultados 106
4.3 Estudo Empírico Para Escolha dos Limites de Separação de Classes 108
4.3.1 Introdução 108
4.3.2 Apresentação dos resultados empíricos 109
4.3.3 Análise dos resultados empíricos 118
4.4 Comentários e Conclusões 118
5 Conclusões Finais e Sugestões para Futuros Trabalhos 120
5.1 Introdução 120
5.2 Contribuição do Trabalho 121
5.3 Sugestões para Futuros Trabalhos 122
Apêndice A 123
Apêndice B 131
Bibliografia 138

Chapter 1

Introdução

1.1 Histórico

Desde a invenção dos computadores digitais, existe um esforço entre os pesquisadores no sentido de expandir sua aplicação nas diversas áreas do conhecimento. As técnicas de Reconhecimento de Padrões (RP) são, em geral, um importante componente dos sistemas de inteligência artificial, que associadas com outras áreas têm crescido em motivação no desafio de programar essas máquinas com tarefas cada vez mais ousadas.

Basicamente os sistemas de RP têm como objetivo decidir, com base em informações fornecidas, qual o padrão mais aproximado de um determinado objeto, supostamente desconhecido. De forma simples, RP é a ciência que se preocupa com a descrição ou classificação das medidas (Schalkoff [1992]). Historicamente havia duas grandes áreas de aproximação ou estimação dessa classificação: Estatística (ou Teoria da Decisão) e Sintática (ou Análise Estrutural). Mas com os avanços tecnológicos, pesquisadores têm dado destaque a uma terceira área, conhecida como Redes Neurais Artificiais, que através da implementação de algoritmos chamados de caixa preta, tem surgido como alternativa quando os dois métodos citados acima apresentam-se ineficientes. Entretanto, cada uma delas apresenta limitações,

sendo importante considerar o interrelacionamento entre elas na solução de um problema de RP.

A modelagem de um sistema estatístico de RP de dois padrões deve se inciar pela tentativa de formular matematicamente um modelo de classificação, com base na Teoria da Probabilidade. O conhecimento passado sobre as características do problema, conduzem a uma solução natural através da Teoria da Decisão, ou mais precisamente através da Teoria de Teste de Hipóteses, onde o objetivo é decidir, com certa margem de erro, em aceitar ou rejeitar uma hipótese de interesse. Portanto, é intuitivamente plausível utilizar técnicas estatísticas na elaboração desses modelos. Quando não for possível estabelecer tais condições para a construção dos modelos, uma segunda tentativa conduz a solução do problema a uma das outras técnicas (Sintática ou Redes Neurais) citadas acima.

A estrutura básica de um sistema de RP (ver Figura 1), consiste de:

1 - Um sensor ou receptor de sinal (uma câmera fotográfica ou uma caneta ótica, por exemplo), de modo que a informação recebida possa ser codificada;

2 - Um pré-processador ou um tradutor de sinal, onde os códigos recebidos possam ser armazenados em um banco de dados;

3 - Um mecanismo de seleção de características, sendo que aquelas mais importantes venham a ser futuramente utilizadas para fins de reconhecimento;

4 - Um algoritmo de classificação, que dependa da aproximação utilizada. Adicionalmente, um conjunto contendo informações passadas sobre os possíveis objetos, chamado conjunto de treinamento, são utilizadas para descrever o padrão que irá representar cada objeto na comparação com outros, supostamente desconhecidos.

Particularmente, os dois últimos blocos da estrutura da Figura 1, interessam de perto aos procedimentos estatísticos. A seleção de variáveis tem como objetivo reduzir o espaço original diminuindo a complexidade do sistema de RP sem perder a eficiência de desempenho do mesmo. De fato, a seleção de características sempre é desejável, visto que isso facilita a execução do procedimento de classificação, e consequentemente, será necessário menos espaço de memória computacional. Em geral, um bom desempenho do procedimento de classificação está intimamente relacionado a um bom procedimento de seleção.

Figura : 1 – Diagrama Funcional de um Sistema Básico de Reconhecimento de Padrões

O desenvolvimento deste trabalho pode ser dividido em duas partes: A primeira, trata de forma geral o problema da classificação sob o ponto de vista estatístico, apresentando um novo procedimento para ser comparado aos existentes na literatura, enquanto que a segunda parte avalia os procedimentos de seleção de variáveis, também sob o ponto de vista estatístico, de modo a manter as taxas de classificação errôneas baixas.

1.2 Objetivo do Trabalho

O objetivo deste trabalho é estudar o desempenho dos procedimentos que utilizam métodos estatísticos na solução dos problemas típicos de RP. Mais precisamente vamos avaliar a aplicação de técnicas estatísticas em relação aos dois últimos passos da estrutura apresentada na Figura 1. E, particularmente, enfatizar os casos com apenas dois padrões básicos, embora uma generalização ao caso de mais de dois padrões, possa ser feita sem grandes dificuldades. Nesse sentido, é comum assumir que os sistemas utilizam um conjunto de treinamento para a construção dos modelos de classificação, composto de observações de

ambos os padrões. Press e Wilson [1978] sugerem as técnicas de Análise Discriminante como procedimento de classificação estatística, se a distribuição condicional das características dos objetos for multinormal com matrizes de covariância iguais, ou os Modelos de Regressão Logística, se alguma das características for qualitativa e/ou na presença de não normalidade nas variáveis. Usualmente, não utilizamos informação *a priori* do objeto pertencer ou não a um determinado padrão, quando empregamos as técnicas de decisão clássica. Entretanto, estas informações *a priori* são incorporadas se adotarmos as técnicas da decisão Bayesiana.

Na realidade, nem sempre é possível observar os dois padrões simultaneamente para que possamos compor o conjunto de treinamento. Em outras palavras, as observações de um dos padrões podem ser inacessíveis, como por exemplo, as observações de uma assinatura falsa. No caso de assinaturas, onde dispomos apenas de um dos padrões para treinamento, as pesquisas tem indicado uma solução por simulação dessas falsificações. Este trabalho propõe um novo classificador, baseado apenas nas observações de um dos padrões. Para verificar o comportamento deste novo método, propositadamente comparamos seu desempenho com aqueles que utilizam tradicionalmente as observações dos dois padrões, além de um classificador construído com base apenas em um dos padrões, que assume variáveis com distribuição Qui-quadrado (Nelson et al [1994]). As comparações são feitas em termos taxas de classificação errôneas através do critério de Neyman-Pearson, isto é, erros em se rejeitar um objeto verdadeiro (chamados de erros do Tipo I) e erros em se aceitar um objeto falso (chamados erros do Tipo II) .

Em seguida, os classificadores são usados na seleção de variáveis (ou características) discriminantes, onde procura-se escolher aquelas mais significativas na construção do modelo de classificação. A principal motivação para tal seleção está na redução do custo de implementação prática do classificador e menor disponibilidade de memória computacional (Heydorn [1970]). Este trabalho também apresenta uma nova técnica de seleção no caso de dispormos apenas das observações de um dos padrões. De forma similar, uma comparação

com as técnicas estatísticas usuais de seleção é feita em termos de erros de classificação. Em outras palavras, as técnicas de seleção de variáveis são incorporadas aos classificadores e seus desempenhos são avaliados conjuntamente através das taxas de erros. Algumas técnicas de seleção de variáveis, conhecidas na literatura, que utilizam apenas um dos padrões, são apresentadas e incorporadas aos classificadores que utilizam apenas um dos padrões, e seus desempenhos também são avaliados.

As técnicas de *RP* têm sido amplamente utilizadas para solucionar problemas de Inteligência Artificial. De fato, características biométricas, tais como manuscritos, impressões digitais, íris, assinaturas, etc., representam excelentes referenciais que podem ser utilizados para identificar a origem do padrão observado, e classificá-lo como pertencente a um determinado indivíduo. A assinatura pessoal é uma das formas mais comum de se autenticar a identidade de um indivíduo. Plamondon e Lorette [1989] publicaram uma seleção de 180 referências a cerca do assunto. Grande parte desses trabalhos utiliza técnicas automáticas de classificação, as chamadas técnicas *on-line*, onde características dinâmicas de uma assinatura, tais como velocidade, aceleração e tempo de execução da assinatura, são utilizadas para caracterizar o processo de escrita. Outra parte desses trabalhos utiliza apenas a imagem fixa das assinaturas, as chamadas técnicas *off-line*, que também tem recebido atenção, pois é a forma mais utilizada nas identificações bancárias. Em geral, espera-se que os sistemas que utilizam as técnicas *on-line* obtenham desempenho muito superior do que os sistemas que utilizam as técnicas *off-line* (Lee [1991]). Neste trabalho, exemplos com os dois tipos de assinaturas (*on-line* e *off-line*) são apresentados como aplicação.

1.3 Resumo do Conteúdo

No Capítulo 2, são discutidas algumas técnicas estatísticas tradicionais de classificação de padrões e apresentado um novo procedimento de classificação com base na distribuição T^2 de Hotelling. O capítulo é encerrado com um estudo de simulação para avaliar as taxas

de erros e comparar o desempenho de cada método apresentado no capítulo.

No Capítulo 3, são discutidas as técnicas usuais de seleção de variáveis baseadas nos procedimentos de classificação de objetos apresentados no Capítulo 2. Uma nova técnica de seleção com base na distância amostral de Mahalanobis entre um ponto e a origem da distribuição é introduzida e comparada com algumas outras técnicas conhecidas na literatura. Um estudo simulado comparativo para avaliação das técnicas discutidas também finaliza esse capítulo.

No Capítulo 4, apresentamos uma aplicação das técnicas de seleção e classificação combinadas na solução de um problema de verificação de assinaturas. Um estudo empírico é feito para avaliar cada método estudado.

Finalmente, no Capítulo 5, são apresentadas as discussões finais a respeito das técnicas apresentadas, juntamente com as contribuições deste trabalho e nossas indicações para futuras pesquisas.

Capítulo 2

Classificação de padrões

2.1 Introdução

Dado um conjunto de características, o problema da classificação de padrões consiste basicamente em encontrar-se uma partição ótima no espaço das características, de modo a se obter sub-regiões, uma para cada classe ou padrão envolvido. O ideal é que essa partição não induza erros ao se tomar as decisões. Quando isso não for possível, esperamos que pelo menos esses erros sejam mínimos. Em geral, cada erro está associado a um custo, que dependendo da situação pode ser maior ou menor. Nesse caso, uma solução seria procurar essas sub-regiões de maneira a minimizar o custo médio dos erros. Essa versão do problema de classificação pode ser abordado como um problema de decisão estatística envolvendo teste de hipóteses, e essa formulação tem tido cada vez mais aplicação na área de reconhecimento de padrões.

Neste Capítulo, são descritas três técnicas usuais de classificação estatística envolvendo apenas duas classes. A proposta é particularizar o problema da classificação geral, quando uma das classes é confrontada com as demais. Na visão estatística de teste de hipóteses, seria equivalente a testarmos apenas duas hipóteses prováveis: verdadeira ou falsa. A aplicação dessa abordagem é comum em várias áreas. Em Medicina substituindo as hipóteses por

doente ou curado. Em Direito Penal, seria o caso de testarmos: inocente ou culpado. E, de forma geral, em Engenharia estaríamos interessados no funcionamento ou não de um determinado equipamento.

Dividimos a apresentação das técnicas em três formas diferentes de abordagem do problema de classificação. Lembrando que a proposta em comum entre elas é classificar um novo objeto, supostamente desconhecido, em uma das duas classes ou padrões. A primeira técnica é conhecida na literatura como Análise Discriminante (AD). Sua formulação básica é supor o conhecimento da distribuição do vetor de características, ou vetor de variáveis explicativas, $\underline{x}$, condicionadas às respectivas classes, π_i, isto é, supõe-se conhecida $p(\underline{x}/\pi_i)$ e a partir daí estimam-se alguns parâmetros envolvidos nessas distribuições, de modo que seja possível estabelecer sub-regiões no espaço das características com base nesses estimadores. A segunda técnica que discutimos é a construção de um modelo de regressão, cuja resposta do modelo, ou variável dependente, é uma variável aleatória com distribuição dicotômica. O modelo também é conhecido como Modelo de Regressão Logística (MRL), devido ao relacionamento do parâmetro da distribuição da variável resposta, y, e o vetor de variáveis que a explica, $\underline{x}$, ser suposto ter uma distribuição logística. Isto é, considerando θ como o parâmetro que associa a probabilidade de sucesso da variável resposta, diremos que $\theta = 1 - F_U(0/\underline{x})$, onde $F_U(0/\underline{x})$ é a função de distribuição acumulada de U, no ponto zero, condicionada ao vetor de variáveis explicativas, $\underline{x}$, onde U é uma variável aleatória não observável, que segue uma distribuição logística com parâmetros de locação $(\alpha + \underline{x}'\underline{\beta})$ e escala unitária. Finalmente, uma terceira técnica aborda o problema na mesma linha da primeira, mas supondo que uma das classes não pode ser avaliada para treinamento do modelo de classificação.

No final deste capítulo apresentaremos um estudo de simulação, onde são comparados, em termos de erros de classificação, todas as técnicas descritas a seguir, juntamente com duas outras usuais encontradas na literatura.

2.2 Análise Discriminante (AD)

2.2.1 Introdução

A formulação matemática da análise discriminante (AD) é feita da seguinte forma (Mardia et al [1979]). Considere J classes denotadas por π_1, π_2, ..., π_J e um vetor aleatório, $\underline{x}$, p-dimensional, que está associado a cada uma das classes através de uma função densidade de probabilidade (*Probability Density Function - pdf*), $p(\underline{x}/\pi_j)$, num espaço R de dimensão p. Desta forma, se um indivíduo é da classe π_j ele tem *pdf* $p(\underline{x}/\pi_j)$. O objetivo da AD é decidir entre as J classes aquela que devemos alocar um novo indivíduo com base nas observações de seu vetor $\underline{x}$. Em outras palavras, a regra discriminante consiste em dividir o espaço R em sub-regiões disjuntas R_1, R_2, ..., R_J (com $\bigcup R_j = R$) tal que,

$$\text{associamos } \underline{x} \text{ na classe } \pi_j \text{ se } \underline{x} \in R_j, \tag{2.1}$$

para $j = 1, 2, ..., J$. Obviamente, quanto mais concentradas estiverem as observações da classe π_j em R_j, para cada j, mais precisa será a discriminação.

Assumimos que não dispomos de qualquer informação *a priori* sobre as classes. Isto é, não conhecemos previamente as chances de um indivíduo pertencer às respectivas classes sem antes termos efetuado as medidas do vetor $\underline{x}$. Entretanto, se essa informação pode ser incorporada, um outro enfoque, conhecido com Regra Discriminante de Bayes, pode ser avaliada. Assim, considerando o conhecimento de todas as probabilidades *a priori*, $P(\pi_j)$, $(j = 1, 2, ..., J)$ e as respectivas *pdf* 's condicionais de $\underline{x}$ dada a classe π_j, $p(\underline{x}/\pi_j)$, e utilizando a regra de Bayes pode-se definir um classificador de acordo com a probabilidade *a posteriori* de classe π_j, que é dada por:

$$p(\pi_j/\underline{x}) = \frac{p(\underline{x}/\pi_j)P(\pi_j)}{p(\underline{x})} \qquad j = 1, 2, ..., J,$$

onde

$$p(x) = \sum_{j=1}^{J} p(\underline{x}/\pi_j)P(\pi_j),$$

sendo natural associar $\underline{x}$ à classe π_j que tiver maior probabilidade *a posteriori*, $p(\pi_j/\underline{x})$.

Entretanto, a utilização dessa informação não será nosso objetivo central neste trabalho. Particularmente, o objetivo aqui será estudar a construção de uma regra para classificar um indivíduo em uma dentre duas classes possíveis, $(J = 2)$, supondo conhecidas apenas as distribuições condicionais do vetor $\underline{x}$, dado a classe π_j. A nossa proposta é estudar um classificador equivalente ao de Bayes, supondo que as probabilidades *a priori*, sejam iguais. Lembramos ainda que a generalização $(J > 2)$, pode ser obtida com um pouco de álgebra a partir de (2.1), com as devidas modificações.

2.2.2 Limites de separação entre classes

Seja R um espaço p-dimensional contendo os pontos do vetor $\underline{x}$. Queremos dividir essa região, R, em duas outras, digamos, R_1 e R_2, de modo que possamos estabelecer uma regra, que baseada em algum método ótimo, seja capaz de decidir se o objeto representado por $\underline{x}$ pertence a classe π_1 ou π_2, isto é, queremos encontrar R_1 e R_2 tal que,

$$\begin{cases} \text{se } \underline{x} \in R_1, \text{ associamos } \underline{x} \text{ com a classe } \pi_1, \\ \text{se } \underline{x} \in R_2, \text{ associamos } \underline{x} \text{ com a classe } \pi_2. \end{cases} \tag{2.2}$$

Se as classes não são linearmente separáveis, isto é, se não é possível traçar um hiperplano separando as observações das classes em regiões distintas, qualquer método que leve a essa partição carrega consigo, inevitavelmente, uma margem de erro de classificação. Usando a linguagem de Teste de Hipóteses, a regra pode alocar $\underline{x}$ na classe π_2, quando na realidade $\underline{x}$ pertence a classe π_1, sendo esse erro definido como *erro do Tipo I* , e sua probabilidade denotada aqui por α_1. Mas também pode ocorrer o contrário, a regra pode alocar $\underline{x}$ na classe π_1, quando na verdade ele pertence a classe π_2, sendo esse erro definido como *erro do Tipo II* e α_2 sua probabilidade. Para definir formalmente a probabilidade desses dois tipos

de erros, considere, $p(\underline{x}/\pi_1)$ e $p(\underline{x}/\pi_2)$ as pdf's de $\underline{x}$ condicionadas respectivamentes as duas classes, π_1 e π_2. Assim, a probabilidade do erro total, α, consiste de duas partes,

$$\begin{aligned}\alpha_1 &= \text{Prob}(\underline{x} \text{ ser classicado na classe } \pi_2 \ / \ \underline{x} \text{ pertence a classe } \pi_1)\\ \alpha_2 &= \text{Prob}(\underline{x} \text{ ser classicado na classe } \pi_1 \ / \ \underline{x} \text{ pertence a classe } \pi_2),\end{aligned}$$

isto é,

$$\alpha_1 = \int_{R_2} p(\underline{x} \ / \ \pi_1) \ d\underline{x} \qquad \text{e} \qquad \alpha_2 = \int_{R_1} p(\underline{x} \ / \ \pi_2) \ d\underline{x} \tag{2.3}$$

A menos que exista uma razão formal, não se faz necessário estabelecer α_1 diferente de α_2. Numa situação mais específica poderíamos estabelecer uma relação do tipo: $\alpha_1 = k\alpha_2$, cujo valor de k dependeria do pesquisador. Portanto, é mais comum adotar-se $k = 1$, e salientar que a troca do valor de k, por qualquer outro valor diferente de 1, pode modificar completamente a regra de decisão.

O problema agora é dividir R em R_1 e R_2, tal que α, seja minimizado, sujeito a restrição de que $\alpha_1 = \alpha_2$, onde

$$\begin{aligned}\alpha &= \alpha_1 + \alpha_2 = \int_{R_2} p(\underline{x} \ / \ \pi_1) \ d\underline{x} + \int_{R_1} p(\underline{x} \ / \ \pi_2) \ d\underline{x}\\ &= 1 - \left\{ \int_{R_2} [p(\underline{x} \ / \ \pi_2) \ - p(\underline{x} \ / \ \pi_1)] \ d\underline{x} \right\},\end{aligned}$$

com

$$\int_{R_1} p(\underline{x} \ / \ \pi_2) \ d\underline{x} = 1 - \int_{R_2} p(\underline{x} \ / \ \pi_2) \ d\underline{x}. \tag{2.4}$$

A solução desse problema é baseada no Lema de Neyman-Pearson, que pode ser encontrado na literatura estatística, como por exemplo em Rao[1973]. O critério avalia a razão de

verossimilhanças no limite de separação entre as classes como:

$$\frac{p(\underline{x} \ / \ \pi_1)}{p(\underline{x} \ / \ \pi_2)} = c, \tag{2.5}$$

onde c é uma constante real escolhida de modo que $\alpha_1 = \alpha_2$. Observe ainda que assumimos que $\underline{x}$ tem uma *pdf* contínua, mas o resultado é equivalente ao se utilizar uma função de probabilidade discreta. O mais importante aqui é frisar que a solução é a razão de verossimilhanças para as observações do vetor $\underline{x}$.

A seguir descreveremos a construção do limite de separação entre as classe sob a suposição que as *pdf*'s condicionais às classes π_1 e π_2 são multinormais.

2.2.3 Classificador de máxima verossimilhança sob normalidade

A situação onde as *pdf*'s $p(\underline{x} \ / \pi_i)$, $i = 1, 2$, são conhecidas é uma tentativa de simplificação no caso teórico. Embora seja pouco real na prática, isso pode ser utilizado para testar a precisão dessas distribuições, desde que o tamanho das amostras seja grande. Entretanto, a grande vantagem de supor o conhecimento das *pdf*'s é que inferências podem ser feitas a partir desta suposição. Assim, é possível utilizar a Regra Discriminante de Máxima Verossimilhança, cuja definição segue:

Definição 2.2.1 - A regra Discriminante de Máxima Verossimilhança para alocação de uma observação de $\underline{x}$ em uma das classes π_1 ou π_2, consiste de alocar $\underline{x}$ na classe em que ocorrer a maior verossimilhança para $\underline{x}$.

Observe que essa definição também vale para $J > 2$, e significa que a regra aloca $\underline{x}$ em π_j se

$$L_j(\underline{x}) = \max_i L_i(\underline{x}), \qquad i = 1, 2, ..., J,$$

onde $L_j(\underline{x})$ é a *pdf* sob as observações do vetor $\underline{x}$ condicionadada à classe π_j, isto é,

$$L_j(\underline{x}) = p(\underline{x} \ / \ \pi_j), \qquad j = 1, 2, ..., J.$$

Para uma dada observação de $\underline{x}$, quando $J = 2$ e ambas as verossimilhanças têm o mesmo valor, qualquer uma das classes pode ser escolhida. Quando isto ocorre, significa que temos um ponto de separação ótimo entre as classes, e neste caso, qualquer uma das classes poderia ser escolhida como a mais provável. A seguir apresentamos uma regra de classificação usando o método de máxima verossimilhança com dados que seguem uma distribuição Normal.

Considere que a *pdf* de $\underline{x}$ condicionada às suas respectivas classes têm distribuição Normal, e que possuem vetores de médias diferentes, dependendo da respectiva classe, mas com a mesma matriz de covariância, isto é:

$$\begin{aligned} p(\underline{x} \ / \ \pi_1) &\sim N_p(\underline{\mu}_1, \Sigma), \qquad \text{se } \underline{x} \in \pi_1 \\ p(\underline{x} \ / \ \pi_2) &\sim N_p(\underline{\mu}_2, \Sigma), \qquad \text{se } \underline{x} \in \pi_2. \end{aligned}$$

Aplicando o operador logaritmo natural em (2.5) temos que,

$$\begin{aligned}\log\frac{p(\underline{x}/\pi_1)}{p(\underline{x}/\pi_2)} &= \frac{1}{2}\{(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)-(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)\}\\ &= \underline{\delta}'\Sigma^{-1}\underline{x}-\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\Sigma^{-1}\underline{\delta}\\ &= \underline{a}'\underline{x}-\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\underline{a}, \qquad (2.6)\end{aligned}$$

onde $\underline{a}'$ é o vetor transposto de $\underline{a}$, com $\underline{a}=\Sigma^{-1}\underline{\delta}$ e $\underline{\delta}=(\underline{\mu}_1-\underline{\mu}_2)$, sendo a fronteira de separação obtida através de $\underline{a}'\underline{x}=h$, onde $h=\frac{1}{2}(\underline{\mu}_1+\underline{\mu}_2)'\underline{a}$ é uma constante que não depende das observações de $\underline{x}$. De fato, queremos alocar $\underline{x}$ em π_1 se $\log\frac{p(\underline{x}/\pi_1)}{p(\underline{x}/\pi_2)}>0$, fazendo sentido definir as sub-regiões como:

$$\begin{aligned}R_1 \text{ é a região tal que } \underline{a}'\underline{x} &\geq h\\ R_2 \text{ é a região tal que } \underline{a}'\underline{x} &< h.\end{aligned}$$

A inclusão de h em R_1 é arbitrária, embora um procedimento aleatório possa ser utilizado no caso de $\underline{a}'\underline{x}=h$, incluindo h em uma das regiões dependendo, por exemplo, de um sorteio.

Para definir α_1 e α_2 as *pdf*'s Normais, são substituídas em (2.3), isto é,

$$\alpha_1=\int_{\underline{a}'\underline{x}<h}N_p(\underline{\mu}_1,\Sigma)d\underline{x} \qquad \text{e} \qquad \alpha_2=\int_{\underline{a}'\underline{x}\geq h}N_p(\underline{\mu}_2,\Sigma)d\underline{x}, \qquad (2.7)$$

substituindo agora o vetor $\underline{x}$ em (2.7) pela variável $u=\underline{a}'\underline{x}$, e as outras $(p-1)$ transformações convenientes nas variáveis de $\underline{x}$, obtemos,

$$\alpha_1=\int_{u<h}\phi_1(u)du \qquad \text{e} \qquad \alpha_2=\int_{u\geq h}\phi_2(u)du, \qquad (2.8)$$

onde $\phi_i(u)$ representa as *pdf*'s da variável u com respeito a classe π_i $(i=1,2)$, respectivamente. Observe também que os limites de integração dependem apenas de u. Como uma função linear de variável com distribuição Normal tem também distribuição Normal (ver

James[1981]), segue que u tem distribuição Normal univariada com valor esperado $E(u)$, dependendo da classe a que $\underline{x}$ pertença, isto é,

$$\begin{aligned} E(u) &= \underline{a}'\underline{\mu}_1 \text{ quando } \underline{x} \in \pi_1 \quad \text{ou} \\ E(u) &= \underline{a}'\underline{\mu}_2 \text{ quando } \underline{x} \in \pi_2, \end{aligned}$$

e variância comum em ambas as classes, isto é,

$$Var(u) = \underline{a}'\Sigma\underline{a} = \underline{\delta}'\Sigma^{-1}\underline{\delta} = \Delta_p^2, \qquad \text{para ambos os casos.} \tag{2.9}$$

Assim,

$$\phi_i(u) = \frac{1}{\Delta_p\sqrt{2\pi}} \exp\{-\frac{1}{2\Delta_p^2}(u - \underline{a}'\underline{\mu}_i)^2\}, \qquad i = 1, 2. \tag{2.10}$$

Usando agora a transformação $z = \{u - E(u)\}/\Delta_p$, temos que z tem distribuição Normal padrão univariada, assim:

$$\begin{aligned} \alpha_1 &= \text{Prob}(u < h/\underline{x} \in \text{a classe } \pi_1) = \text{Prob}(z < \frac{h - \underline{a}'\underline{\mu}_1}{\Delta_p}) \qquad \text{e} \\ \alpha_2 &= \text{Prob}(u \geq h/\underline{x} \in \text{a classe } \pi_2) = \text{Prob}(z \geq \frac{h - \underline{a}'\underline{\mu}_2}{\Delta_p}) = \text{Prob}(z \leq -\frac{h - \underline{a}'\underline{\mu}_2}{\Delta_p}), \end{aligned}$$

portanto,

$$\alpha_1 = \Phi\left(\frac{h - \underline{a}'\underline{\mu}_1}{\Delta_p}\right) \qquad \text{e} \qquad \alpha_2 = \Phi\left(\frac{\underline{a}'\underline{\mu}_2 - h}{\Delta_p}\right), \tag{2.11}$$

onde $\Phi(\cdot)$ representa a Função Distribuição Acumulada (FD), de uma variável aleatória Normal padronizada, isto é,

$$\Phi(z_0) = \int_{z<z_0} \frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\} dz. \tag{2.12}$$

Se estabelecermos que $\alpha_1 = \alpha_2$, de (2.11) segue que

$$h = \frac{1}{2}\underline{a}'(\underline{\mu}_1 + \underline{\mu}_2). \tag{2.13}$$

Substituindo (2.13) em (2.11), obtemos

$$\alpha_1 = \alpha_2 = \Phi\left(-\frac{1}{2}\Delta_p\right). \tag{2.14}$$

Em resumo, o procedimento discriminatório entre duas classes com distribuições Normais com diferentes vetores de médias e mesma matriz de covariância, segue os seguintes passos:

i) Calcule $u = \underline{a}'\underline{x} = \underline{\delta}'\Sigma^{-1}\underline{x}$, onde $\underline{\delta} = \left(\underline{\mu}_1 - \underline{\mu}_2\right)$.

ii) Calcule $h = \frac{1}{2}\underline{a}'(\underline{\mu}_1 + \underline{\mu}_2)$.

iii) Se $u \geq h$ associe $\underline{x}$ à classe π_1, caso contrário, associe $\underline{x}$ à classe π_2.

(2.15)

2.2.4 Função discriminante linear de Fisher

Fisher[1936] elaborou uma função discriminante considerando o caso univariado, para duas classes com distribuições Normais, com diferentes médias, μ_1 e μ_2, e mesmas variâncias $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Ele observou que é intuitivamente claro que se $(\mu_1 - \mu_2)/\sigma$ é grande, o risco de má classificação é pequeno, pois as regiões envolvendo as duas classes tem uma pequena área sobreposta. Caso contrário, se $(\mu_1 - \mu_2)/\sigma$ é muito próximo a zero, essa área sobreposta é muito grande crescendo assim o risco de má classificação. Fisher sugeriu que as p características do vetor $\underline{x}$, fossem combinadas linearmente, reduzindo, assim, a dimensão do espaço, e que a medida de distância $(\mu_1 - \mu_2)/\sigma$ para essa combinação linear fosse máxima. A regra de classificação seria então baseada nessa combinação linear ótima. Escrevendo de outra forma, a proposta dele seria reduzir a classificação feita em dimensão p para uma única dimensão, escolhendo uma combinação linear do vetor $\underline{x}$, $\underline{a}'\underline{x}$, tal que

$$\frac{|\mu_1^* - \mu_2^*|}{\Sigma^*} = \frac{\underline{a}'\underline{\delta}}{(\underline{a}'\Sigma\underline{a})^{1/2}}, \tag{2.16}$$

seja maximizado com respeito ao vetor $\underline{a}$, onde μ_i^* é média de $\underline{a}'\underline{x}$ supondo que $\underline{x}$ pertence a classe π_i $(i = 1, 2)$, Σ^* é o desvio padrão de $\underline{a}'\underline{x}$ em ambas as classes, $\delta = \underline{\mu}_1 - \underline{\mu}_2$ e Σ a matriz de covariância comum a ambas as classes . Diferenciando a equação (2.16) em relação a $\underline{a}$ (ver Apêndice A-3) e igualando o resultado a zero, obtemos

$$\underline{\delta} = \left(\frac{\underline{a}'\underline{\delta}}{\underline{a}'\Sigma\underline{a}}\right)\Sigma\underline{a}. \tag{2.17}$$

Observe que uma possível solução para a equação (2.17) é tomar $\underline{a}$ como sendo:

$$\underline{a} = \Sigma^{-1}\underline{\delta}, \tag{2.18}$$

embora qualquer vetor proporcional a $\underline{a}$, também satisfaça a equação (2.17), isto é, $c\underline{a}$ também é solução da equação (2.17), onde c é qualquer constante diferente de zero e $\underline{a}$ definido na equação (2.18). Portanto a combinação linear ótima é $\underline{a}'\underline{x}$ ou qualquer função linear de $\underline{x}$ que seja proporcional a $\underline{a}'\underline{x}$, de modo que a classificação de um indivíduo com p características seria baseada simplesmente na função linear $\underline{a}'\underline{x}$, que é conhecida como função discriminante linear.

Note que a função discriminante $u = \underline{a}'\underline{x}$ definida pela equação (2.18) é a mesma que a obtida, através da razão de verossimilhanças sob a condição de normalidade, em (2.6), embora as justificativas para cada método sejam diferentes. Na Subseção 2.2.3, existe uma suposição explícita de normalidade, além da condição $\alpha_1 = \alpha_2$, enquanto que nesta Subseção, a busca foi baseada na sensibilidade da função discriminante. Essa é também uma justificativa para se utilizar esta mesma função discriminante, mesmo quando a suposição de normalidade não é satisfeita.

Embora os elementos de $\underline{a}$ sejam únicos, seus coeficientes não são, como foi mencionado anteriormente. Assim, como qualquer função da forma $c\underline{a}'\underline{x}$, sendo c constante, também é uma função discriminante, e a classificação é baseada nos passos de (2.15) com $\underline{a}'\underline{x} = \underline{\delta}'\Sigma^{-1}\underline{x}$. Substituindo $\underline{a} = \Sigma^{-1}\underline{\delta}$, em (2.16) a distância entre as classes π_1 e π_2, baseada na função discriminante $\underline{a}'\underline{x}$, se reduz a Δ_p, conhecida como a distância de Mahalanobis, sendo $\Delta_p^2 = \underline{\delta}'\Sigma^{-1}\underline{\delta} = \underline{a}'\underline{\delta}$.

2.2.5 Classificador a partir de dados amostrais

Na prática, dificilmente conhecemos os verdadeiros valores dos parâmetros das distribuições Normais envolvidas, $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ. Em geral, tomamos duas amostras de tamanhos

n_1 e n_2, respectivamente, em cada classe π_1 e π_2, e usamos estas amostras para estimar os valores dos parâmetros, sendo um procedimento natural substituir os parâmetros pelos seus estimadores. Formalmente, organizando as observações em forma de matrizes, consideramos X_1 e X_2 duas matrizes (n_1 x p) e (n_2 x p) de observações, respectivamente, das classes π_1 e π_2, como sendo

$$X_1 = \begin{bmatrix} \underline{x}_1^{(1)\prime} \\ \underline{x}_1^{(1)\prime} \\ \cdots \\ \underline{x}_{n_1}^{(1)\prime} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_{11}^{(1)} \; \mathrm{x}_{12}^{(1)} \; \ldots \; \mathrm{x}_{1\mathbf{p}}^{(1)} \\ \mathrm{x}_{21}^{(1)} \; \mathrm{x}_{22}^{(1)} \; \ldots \; \mathrm{x}_{2\mathbf{p}}^{(1)} \\ \cdots \\ \mathrm{x}_{n_1 1}^{(1)} \; \mathrm{x}_{n_1 2}^{(1)} \; \ldots \; \mathrm{x}_{n_1\mathbf{p}}^{(1)} \end{bmatrix} \text{ e } X_2 = \begin{bmatrix} \underline{x}_1^{(2)\prime} \\ \underline{x}_2^{(2)\prime} \\ \cdots \\ \underline{x}_{n_2}^{(2)\prime} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_{11}^{(2)} \; \mathrm{x}_{12}^{(2)} \; \ldots \; \mathrm{x}_{1\mathbf{p}}^{(2)} \\ \mathrm{x}_{21}^{(2)} \; \mathrm{x}_{22}^{(2)} \; \ldots \; \mathrm{x}_{2\mathbf{p}}^{(2)} \\ \cdots \\ \mathrm{x}_{n_2 1}^{(2)} \; \mathrm{x}_{n_2 2}^{(2)} \; \ldots \; \mathrm{x}_{n_2\mathbf{p}}^{(2)} \end{bmatrix},$$

onde $\underline{x}_i^{(1)}$ segue uma distribuição $N_p(\underline{\mu}_1, \Sigma)$ e $\underline{x}_j^{(2)}$ segue uma distribuição $N_p(\underline{\mu}_2, \Sigma)$, ($i = 1, 2, ..., n_1$), e ($j = 1, 2, ..., n_2$). Definimos,

$$\begin{aligned} \overline{\underline{x}}^{(1)} &= \frac{1}{n_1} X_1' \mathbf{1}_{\mathbf{n_1,1}}, \qquad \overline{\underline{x}}^{(2)} = \frac{1}{n_2} X_2' \mathbf{1}_{\mathbf{n_2,1}} \\ S_1 &= X_1'(I_{n_1} - \frac{1}{n_1}\mathbf{1}_{\mathbf{n_1,n_1}})X_1, \qquad S_2 = X_2'(I_{n_2} - \frac{1}{n_2}\mathbf{1}_{\mathbf{n_2,n_2}})X_2 \\ S &= S_1 + S_2, \qquad m = n_1 + n_2 - 2, \qquad \underline{d} = \overline{\underline{x}}^{(1)} - \overline{\underline{x}}^{(2)} \quad \text{e} \end{aligned} \tag{2.19}$$

$$C^2 = \frac{n_1 n_2}{n_1 + n_2}$$

onde $\mathbf{1_{a,b}}$ representa uma matriz ($\mathbf{a}$ x $\mathbf{b}$) cujos elementos são todos iguais a 1 e I_a a matriz identidade de ordem a. Usando o Lema 2.4.1 (ver pag. 40), temos que $C\underline{d}$ é independente de S, onde $C\underline{d}$ tem distribuição $N_p(C\underline{\delta}, \Sigma)$ e S tem distribuição Wishart com parâmetros Σ e m, $\{W_p(\Sigma, m)\}$ (ver Definição 2.4.1, pag.40),onde $\underline{d}$ e $\frac{1}{m}S$, são estimadores não viciados de $\underline{\delta}$ e Σ, respectivamente. Admitindo que um novo vetor, $\underline{z}$, seja objeto de classificação, e substituindo na função discriminante seus estimadores, chegamos a

$$\underline{l}'\underline{z} = \underline{d}' \left(\frac{1}{m}S\right)^{-1} \underline{z} = m\underline{d}' S^{-1}\underline{z}, \tag{2.20}$$

onde $l = mS^{-1}\underline{d}$ é um estimador de $\underline{a}$ obtido pela substituição dos estimadores não viciados de Σ e de δ, definindo assim a função discriminante amostral $\underline{l}'\underline{z}$. Portanto o procedimento de classificação (2.15) sugere os seguintes passos:

i) Calcule $\underline{l}'\underline{z}$ como em (2.20).

ii) Calcule $\frac{1}{2}\underline{l}'(\overline{\underline{x}}^{(1)} + \overline{\underline{x}}^{(2)})$ como em (2.19).

iii) Se $\underline{l}'\underline{z} - \frac{1}{2}\underline{l}'(\overline{\underline{x}}^{(1)} + \overline{\underline{x}}^{(2)}) \geq 0$ associe $\underline{z}$ à classe π_1, caso contrário associe $\underline{z}$ à classe π_2.

(2.21)

Esta regra de classificação é chamada Classificação Estatística de Anderson (Anderson [1958]). A probabilidade de classificação errônea de (2.21) pode ser estimada substituindo Δ_p^2 pelo seu estimador em (2.14), isto é:

$$\widehat{\alpha}_1 = \widehat{\alpha}_2 = \Phi(-\frac{1}{2}D), \tag{2.22}$$

onde

$$D^2 = m\underline{d}' S^{-1}\underline{d}. \tag{2.23}$$

Infelizmente as probabilidades de má classificação, $\widehat{\alpha}_1$e $\widehat{\alpha}_2$, definidas em (2.22) são otimistas, isto é, tendem a subestimar as verdadeiras probabilidades (ver Mardia et al [1979]). Aitchison et al [1977] argumentam que essas probabilidades, baseadas na regra de Bayes, são mais

realísticas. Na Seção 2.5 apresentamos uma versão particular da regra de Bayes de acordo com o problema aqui apresentado.

Uma justificativa para adotarmos a regra (2.21) considerando $\underline{l} = S^{-1}\underline{d}$ é apresentada a seguir:

Uma outra forma de obtermos o valor de $\underline{l}$ é considerar a expressão (2.16), onde seus parâmetros são substituídos pelas suas estimativas, levando a:

$$\frac{\left|\widehat{\mu_1^*} - \widehat{\mu_2^*}\right|}{\widehat{\Sigma^*}} = \frac{|\underline{l}'\underline{d}|}{(m^{-1}\underline{l}'S\underline{l})^{\frac{1}{2}}}. \tag{2.24}$$

O objetivo, portanto, é encontrar uma combinação linear do valor observado, $\underline{l}'\underline{z}$, que maximize (2.24) com respeito a $\underline{l}$, onde $\underline{z}$ é uma nova observação. Elevando (2.24) ao quadrado e multiplicando pela constante C^2 obtemos:

$$C^2\left(\frac{|\underline{l}'\underline{d}|}{(m^{-1}\underline{l}'S\underline{l})^{\frac{1}{2}}}\right)^2 = \frac{C^2(\bar{t}_1 - \bar{t}_2)^2}{m^{-1}\underline{l}'(S_1 + S_2)\underline{l}}, \tag{2.25}$$

onde, C^2 foi definido em (2.19) e

$$\bar{t}_1 = \underline{l}'\overline{\underline{x}}^{(1)}, \bar{t}_2 = \underline{l}'\overline{\underline{x}}^{(2)}.$$

Definindo agora $t = \underline{l}'\underline{z}$, temos que $C^2(\bar{t}_1 - \bar{t}_2)^2 = C^2(\underline{l}'\underline{d})^2$ e $\underline{l}'S_1 + S_2\underline{l} = \underline{l}'S\underline{l}$ que são denominadas, respectivamente, a Soma de Quadrados entre classes ($S.Q.Entre$) e Soma de Quadrados intra classes ($S.Q.Intra$) na análise de variância da variável t, onde $\underline{l}'\underline{z}$ é uma

combinação linear para uma nova observação, $\underline{l}'\underline{x}^{(1)}$ uma combinação linear nas observações da classe π_1 e $\underline{l}'\underline{x}^{(2)}$ uma combinação linear nas observações da classe π_2.

A Tabela 2.1 apresenta o quadro usual da Análise de Variância ($ANOVA$) (ver Draper e Smith [1981]), da variável aleatória t, onde a variação total (Soma de Quadrados Total) é dividida em duas parcelas: Soma de Quadrados entre Classes (variabilidade entre observações de classes distintas) e Soma de Quadrados Intra Classes (variabilidade entre observações dentro da mesma classe).

Tabela 2.1 - Tabela da $ANOVA$ para a variável t

Fonte	G.L.	S.Q.
Entre classes	1	$C^2(\underline{l}'\underline{d})^2$
Intra classes	m	$\underline{l}'S\underline{l}$
Total	$m+1$	$C^2(\underline{l}'\underline{d})^2 + \underline{l}'S\underline{l}$

Observe que a razão entre as Soma de Quadrados (SQ) das fontes Entre classes e Intra classes, divididas respectivamente, pelos seus Graus de Liberdade (GL), é a mesma que a expressão (2.25). Em outras palavras, maximizar (2.24) é equivalente a maximizar a razão F, da análise de variância da variável t, onde

$$F = \frac{\frac{C^2(\underline{l}'\underline{d})^2}{1}}{\frac{\underline{l}'S\underline{l}}{m}}. \tag{2.26}$$

Através do argumento usado em (2.18), onde não existe uma solução única para $\underline{l}$, segue que F é maximizado quando $\underline{l}$ é proporcional a $S^{-1}\underline{d}$. Substituindo $\underline{l} = S^{-1}\underline{d}$ em (2.24) obtemos a distância amostral dada por

$$D_p = (m\underline{d}' S^{-1} \underline{d})^{\frac{1}{2}}, \tag{2.27}$$

ou equivalentemente,

$$D_p^2 = m\underline{d}' S^{-1} \underline{d},$$

como obtida anteriormente em (2.23). Na literatura D_p^2 é conhecida como a distância D^2 de Mahalanobis Studentizada, baseada em p características do vetor $\underline{z}$. Esta distância também será usada no próximo capítulo, no sentido de definir a contribuição de cada característica de $\underline{z}$ na construção da distância entre as classes π_1 e π_2.

2.3 Modelo de Regressão Logística (MRL)

2.3.1 Introdução

Na seção anterior descrevemos a formulação matemática de um classificador linear, com base nas observações de um vetor de características, $\underline{x}$, dando ênfase à distribuição condicional desse vetor dada a classe π_1 ou π_2, respectivamente, $p(\underline{x}/\pi_i)$ $(i = 1, 2)$. Consideramos agora uma técnica alternativa de abordagem do problema de classificação através de modelos de regressão linear. É importante lembrar que este trabalho trata o problema da estimação da variável resposta, via modelos de regressão, de uma forma bastante particular.

Existe uma vasta literatura sobre este tema, na sua formulação geral, como por exemplo, Draper e Smith [1981], McCullagh e Nelder [1989] entre outros, e que para maiores esclarecimento, uma consulta inicial sobre os procedimentos usuais de estimação e interpretação dos modelos, se faz necessário.

Basicamente, os modelos de regressão buscam estudar o relacionamento entre uma variável resposta, y, chamada de variável dependente, e um vetor de variáveis explicativas,

$\underline{x}$, chamada de variáveis independentes. O modelo de regressão linear é definido por:

$$\underline{y} = X'\underline{\beta} + \underline{\varepsilon} \qquad \text{com} \quad E\{\underline{y}/X\} = X'\underline{\beta} \tag{2.28}$$

onde

$\underline{y}$ é um vetor de observações da variável dependente (usualmente contínua),

X é uma matriz de dados observados das variáveis independentes, sendo X' sua transposta,

$\underline{\beta}$ é um vetor de parâmetros desconhecidos e

$\underline{\varepsilon}$ é um vetor de erros na estimação de $\underline{y}$ pelo modelo $X'\underline{\beta}$, independentes das observações.

Assumindo alguma *pdf* para o vetor de erros, $\underline{\varepsilon}$, um dos objetivos dos modelos de regressão é estimar o vetor de parâmetros $\underline{\beta}$, de modo que o modelo (2.28) possa ser usado como preditor da resposta, $\underline{y}$. Nesse sentido, a abordagem do problema de classificação, via modelos de regressão, propõe uma formulação diferente da seção anterior, sobretudo devido ao fato de não se impor explicitamente condições sobre a matrix de dados X, como por exemplo, que as linhas de X tenham distribuição Normal multivariada.

Para o problema de classificação de um indivíduo em apenas duas classes, definimos a i-ésima observação, y_i, como uma variável dicotômica. Sem perda de generalidade, podemos assumir $y_i = 1$, se um determinado evento, condicionado ao vetor de variáveis independentes, $\underline{x}$, ocorre, e $y_i = 0$, se o evento condicional não ocorre. Mas, nesse caso, o modelo (2.28) não se aplica, devido a variável resposta y_i, não ser contínua, sendo recomendado o modelo logístico definido por:

$$E\{y/\underline{x}\} = u = \text{Prob}(y = 1/\underline{x}) \quad \text{e}$$

$$g(u) = \log(\frac{u}{1-u}) = \underline{x}'\underline{\beta} \tag{2.29}$$

onde $g(u)$ é o logito da probabilidade de y assumir o valor 1, dado o vetor de variáveis independentes, $\underline{x}$, e $\underline{\beta}$ é o vetor de parâmetros desconhecidos. O logito de uma probabilidade p definida por: $\log\left(\frac{p}{1-p}\right)$.

Essa técnica de transformação logística foi introduzida inicialmente por Cornfield [1962] e seguido por Truett et al [1967] para dados relacionados com doenças coronarianas. A partir da introdução do procedimento de estimação de $\underline{\beta}$, através do método de máxima verossimilhança, nos trabalhos de Walker e Duncan [1967], Day e Kerridge [1967] e Cox [1970] essa técnica passou a ser amplamente utilizada em outras áreas tais como nos experimentos biológicos relacionados a estudos de dose-resposta, na área de saúde principalmente em Epidemiologia, etc. (Breslow e Day [1980]). Seu uso é mais frequente em modelos com respostas dicotômicas (apenas duas respostas possíveis, $J = 2$), mas o modelo também pode ser aplicado quando a resposta é politômica ($J > 2$), sendo recomendado um estudo mais detalhado sobre a estrutura das respostas, que podem apresentar uma ordenação natural (ordinal) ou nominal se a resposta estiver em forma de classes (ver Aidar [1992]).

2.3.2 Modelo de classificação com resposta binária

Uma forma alternativa de avaliarmos o problema da classificação, envolvendo apenas duas classes, pode ser interpretada do seguinte modo: Suponhamos que uma observação possa somente pertencer a uma de duas classes, digamos π_1 ou π_2. Sem perda de generalidade podemos codificar, $Y_i = 1$ se a i-ésima observação ($i = 1, 2, ..., n$) pertence a π_1 e $Y_i = 0$ caso contrário. Podemos agora escrever:

$$\begin{aligned} E\{Y_i/\underline{x}_i\} &= \text{Prob}(Y_i = 1/\underline{x}_i) = \theta_i, \qquad \text{e} \\ \text{Prob}(Y_i &= 0/\underline{x}_i) = 1 - \theta_i, \end{aligned} \tag{2.30}$$

onde θ_i é a probabilidade da i-ésima observação pertencer a π_1. Por conveniência chamamos $Y_i = 1$ de sucesso e consequentemente $Y_i = 0$ de fracasso, fazendo sentido chamar tais observações de binárias. Considerando que essas respostas são avaliadas em n indivíduos, supostamente independentes, a proposta é desenvolver um método que associe qualquer dependência de θ_i com $\underline{x}_i$, onde $\underline{x}_i$ representa um conjunto de atributos que explicam Y_i. O objetivo é avaliarmos a equação de regressão:

$$\theta_i = \alpha + \underline{x}_i'\underline{\beta}, \qquad i = 1, 2, ..., n, \tag{2.31}$$

onde $\underline{\beta}$ é um vetor de p coeficientes desconhecidos da regressão e α é um parâmetro também desconhecido, que representa o intercepto do hiperplano.

A principal restrição imposta a (2.31) é que θ_i, sendo uma probabilidade, deve pertencer ao intervalo [0,1], isto é:

$$0 \leq \theta_i \leq 1, \tag{2.32}$$

podendo causar alguns problemas na estimação dos parâmetros, especialmente usando o método dos Mínimos Quadrados Ordinários, que utiliza as derivadas parciais para encontrar os valores de α e $\underline{\beta}$ que minimizam

$$\sum_{i=1}^{n} \left(\theta_i - (\alpha + \underline{x}_i'\underline{\beta})\right)^2,$$

e isso torna o modelo (2.31) com domínio (valores possíveis de θ_i) e interpretação limitada (Cox [1970]). Uma forma de resolver esse problema é supor a existência de uma variável latente, U, que condicionada ao vetor $\underline{x}$ tenha uma *pdf* contínua. Isto é, dada a Função Distribuição Acumulada, FD, $F_U(u/\underline{x})$, podemos definir sem perda de generalidade, a resposta $Y = 1$, se $U > 0$. O ponto de corte considerado foi zero, já que U não está sendo observada diretamente e seus valores estão definidos no conjunto dos números reais ($\Re$). Assim podemos escrever:

$$\theta = \text{Prob}(Y = 1/\underline{x}) = 1 - F_U(0/\underline{x}). \tag{2.33}$$

Observe que se uma variável aleatória U tem *pdf* logística com parâmetro de locação a e de escala b então sua função de densidade é dada por:

$$p(u) = \frac{1}{b} \quad \frac{\exp[(u - a)/b]}{\{1 + \exp[(u - a)/b]\}^2}, \qquad u, a \in \Re \qquad \text{e} \qquad b > 0,$$

sendo uma distribuição simétrica de média a e desvio padrão $\pi\ b/\sqrt{3}$.

O MRL adota U tendo uma *pdf* logística, com parâmetros de locação $(\alpha + \underline{x}'\underline{\beta})$ e escala unitária, embora no contexto de regressão com respostas binárias outros tipos de distribuições podem ser usadas, assim como outros parâmetros. No caso logístico a FD de U será:

$$F_U(u/\underline{x}) = \frac{\exp(u - \alpha - \underline{x}'\underline{\beta})}{1 + \exp(u - \alpha - \underline{x}'\underline{\beta})}, \tag{2.34}$$

de modo que

$$F_U(0/\underline{x}) = \frac{1}{1 + \exp(\alpha + \underline{x}'\underline{\beta})}. \tag{2.35}$$

Segue portanto que

$$\theta = \text{Prob}(Y = 1/\underline{x}) = 1 - F_U(0/\underline{x}) = \frac{\exp(\alpha + \underline{x}'\underline{\beta})}{1 + \exp(\alpha + \underline{x}'\underline{\beta})} = \frac{1}{1 + \exp\{-(\alpha + \underline{x}'\underline{\beta})\}}$$

e

$$1 - \theta = \text{Prob}(Y = 0/\underline{x}) = \frac{1}{1 + \exp(\alpha + \underline{x}'\underline{\beta})}. \tag{2.36}$$

Usando a transformação $\lambda = \log(\frac{\theta}{1-\theta}) = \alpha + \underline{x}'\underline{\beta}$, o relacionamento entre $\underline{x}$ e λ passa a ser linear.

Lembrando mais uma vez, outras suposições sobre a FD de U, condicionada ao vetor $\underline{x}$, podem ser adotadas, inclusive uma outra formulação do problema poderia ser estabelecida, pois a proposta aqui foi fixar o ponto crítico de U (o limiar zero) e variar a distribuição de U conforme $\underline{x}$. Outra proposta, mais comum em ensaios biológicos, onde as variáveis explicativas são dosagens, seria fixar a distribuição de U e variar o ponto crítico com $\underline{x}$. No caso unidimensional, U seria a dose que produz a resposta, y, algumas vezes chamada de tolerância. Em geral, define-se $y = 1$ se a dose excede uma tolerância recomendada. Se a dose recomendada for " $\alpha + x\beta$ ", então

$$\text{Prob}(Y = 1/x) = \text{Prob}(U \leq \alpha + x\beta),$$

relaciona a probabilidade de que $y = 1$, diretamente com a distribuição de U . Todavia, a primeira formulação, juntamente com a distribuição logística, é mais comum, devido a esse tipo de distribuição está mais relacionada com respostas binárias (ver Cox[1970]). A seguir apresentaremos o método usual de estimação dos parâmetros do modelo.

2.3.3 Estimadores de máxima verossimilhança para o MRL

Várias formas de estimação dos parâmetros $(\alpha, \underline{\beta})$ do MRL têm sido propostas e também têm sido tema de várias pesquisas, entretanto em situação geral, os estimadores de máxima verossimilhança, quando comparados empiricamente com outros competidores, têm se mostrado mais eficientes. Em situações muito particulares, estimadores utilizando, por exemplo, função discriminante, podem ser preferíveis (ver Press e Wilson [1978]), especialmente quando o vetor de variáveis explicativas, $\underline{x}$, não segue uma distribuição Normal multivariada, que é uma das condições no problema interpretado via AD. A discussão sobre o melhor método de estimação dos parâmetros tem surgido justamente porque o método de máxima verossimilhança gera um sistema de equações através das derivadas parciais da função de verossimilhança, sendo esse sistema não linear em $\underline{\beta}$. Nesse caso, métodos alternativos de iteração, como por exemplo de Newton-Raphson e Mínimos Quadrados Ponderados Iterativos, devem ser aplicados.

O método de estimação por máxima verossimilhança considera inicialmente o modelo

$$\lambda = \log\left(\frac{\theta}{1-\theta}\right) = X'\underline{\beta}, \tag{2.37}$$

onde $\theta = (\theta_1, \theta_2, ..., \theta_n)$ =Prob$(Y = \mathbf{1}/X)$, $X = (\underline{x}_1,\ \underline{x}_2,\ ..., \underline{x}_n)'$ é a matriz $(n$ x $(p+1))$ do modelo (a matriz de dados acrescida de uma coluna onde todos os elementos são iguais a 1, isto é, $\underline{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$, $(i = 1, 2, ..., n)$, $\underline{y} = (y_1, y_2, ..., y_n)$ com y_i tendo distribuição de Bernoulli(θ_i) e $\underline{\beta} = (\alpha, \beta_1, \beta_2, ..., \beta_p)$ é o vetor de parâmetros desconhecidos.

A função de verossimilhança considerando fixas todas as n observações de $\underline{y}$, é definida por:

$$
\begin{aligned}
L(\underline{\beta}/\underline{x}_i', \underline{y}) &= L_i = \prod_{i=1}^{n} \text{Prob}\,(y = \mathrm{y_i}/\underline{x}_i) \\
&= \prod_{i=1}^{n} \left(\frac{\exp(\underline{x}_i'\underline{\beta})}{1+\exp(\underline{x}_i'\underline{\beta})} \right)^{\mathrm{y_i}} \left(\frac{1}{1+\exp(\underline{x}_i'\underline{\beta})} \right)^{1-\mathrm{y_i}} \\
\text{com} \qquad \mathrm{y}_i &= 0 \qquad \text{ou} \qquad \mathrm{y}_i = 1.
\end{aligned} \tag{2.38}
$$

Já que L_i e $\log L_i$ têm seus valores máximos no mesmo ponto de $\underline{\beta}$, a solução é equivalente quando aplicamos logaritmo em L_i, isto é,

$$
\log\left(L(\underline{\beta}/\underline{x}_i', Y)\right) = \log L_i = \sum_{i=1}^{n} \left[\mathrm{y_i}\underline{x}_i'\underline{\beta} - \log\left(1+\exp(\underline{x}_i'\underline{\beta})\right) \right]. \tag{2.39}
$$

Considerando $b'(.)$ e $b\,''(.)$ como as derivadas primeira e segunda de uma função $b(.)$, onde $b(.) = b(\underline{x}_i'\underline{\beta}) = \exp(\underline{x}_i'\underline{\beta})$, obtemos

$$
\begin{aligned}
\partial\frac{\log L_i}{\partial\beta_j} &= \sum_{i=1}^{n} \left[\mathrm{y}_i\mathrm{x}_{ij} - \mathrm{x}_{ij}\partial\frac{b(\underline{x}_i'\underline{\beta})}{\partial\beta_j} \right] \\
&= \sum_{i=1}^{n} \left[\mathrm{x}_{ij}\left(\mathrm{y}_i - b'(\underline{x}_i'\underline{\beta})\right) \right] = 0
\end{aligned}
$$

e também, sabendo que (ver Dobson [1990], pag 143),

$$
E\{\frac{\partial L_i}{\partial\beta_j}\} = E\{\mathrm{y_i} - b'(\underline{x}_i'\underline{\beta})\} = 0
$$

temos que

$$E\{y_i\} = b'(\underline{x}_i'\underline{\beta}) = \frac{\exp(\underline{x}_i'\underline{\beta})}{1+\exp(\underline{x}_i'\underline{\beta})} = \theta_i$$

e ainda (ver novamente Dobson [1990], pag 143),

$$\begin{aligned} E\{\left(\frac{\partial^2 L_i}{\partial\beta_i\partial\beta_j}\right) + \left(\frac{\partial L_i}{\partial\beta_j}\right)^2\} &= 0 \\ &= -b''(\underline{x}_i'\underline{\beta}) + E\{\mathrm{y_i} - b'(\underline{x}_i'\underline{\beta})\}^2 \end{aligned}$$

segue que

$$b''(\underline{x}_i'\underline{\beta}) = E\{\mathrm{y_i} - b'(\underline{x}_i'\underline{\beta})\}^2 = \theta_i(1-\theta_i) = Var(\mathrm{y_i}) \geq 0.$$

Resolvendo o sistema de equações dado pelas $(p+1)$ derivadas parciais de $\log(L_i)$, em relação a $\underline{\beta}$ e igualando o resultado a zero, temos:

$$\partial\frac{\log\left(L(\underline{\beta}/\underline{x}_i', Y)\right)}{\partial\beta_j} = 0, \qquad j = 0, 1, ..., p, \qquad \text{com} \qquad \beta_0 = \alpha, \tag{2.40}$$

resultando no sistema,

$$\begin{aligned} \sum_{i=1}^{n} \mathrm{x_{ij}}\left[\mathrm{y_i} - \frac{1}{1+\exp(-\underline{x}_i'\underline{\beta})}\right] &= 0, \qquad j = 0, 1, ..., p, \\ \sum_{i=1}^{n} \mathrm{x_{ij}}\left[\mathrm{y_i} - \widehat{\theta}_i\right] &= 0, \end{aligned} \tag{2.41}$$

sendo a solução do sistema, $\underline{\widehat{\beta}}$, não linear em $\underline{\beta}$. A segunda derivada de $\log L_i$ é dada por:

$$\begin{aligned} \partial^2 \frac{\log\left(L(\underline{\beta}/\underline{x}_i', Y)\right)}{\partial\beta_i \partial\beta_j} &= -\sum \mathrm{x}_{ij} b''(\underline{x}_i'\underline{\beta})\mathrm{x}_{ij} \\ &= -\sum \mathrm{x}_{ij}^2 var(\mathrm{y}_i) \leq 0. \end{aligned}$$

Portanto, a solução do sistema (2.41) consiste de pontos de máximo para $L(\underline{\beta}/\underline{x}_i, Y)$, definindo, assim o estimador de máxima verossimilhança de $\underline{\beta}$.

Como a segunda derivada, neste caso, foi de fácil determinação, podemos empregar algum método iterativo, como por exemplo o de Newton-Raphson, para resolver o problema da não linearidede, que procura uma solução iterativa até que uma convergência seja estabelecida, isto é, considera-se $\underline{\beta}^{(0)}$ como solução inicial e deriva-se sucessivamente, $\underline{\beta}^{(1)}$, $\underline{\beta}^{(2)}$,...,$\underline{\beta}^{(t)}$ de modo que,

$$\underline{\beta}^{(t)} = \underline{\beta}^{(t-1)} + (\sum \mathrm{x}_{ij} b''(\underline{x}_i'\underline{\beta})\mathrm{x}_{ij})^{-1} X(\underline{y} - \underline{\widehat{\theta}}),$$

onde t indica que o processo está na t-ésima iteração, até que

$$\left|\underline{\widehat{\beta}}^{(t)} - \underline{\widehat{\beta}}^{(t-1)}\right| < \underline{C},$$

onde $\underline{C}$ é uma precisão pré-definida, $\underline{\widehat{\beta}}^{(t)}$ é última solução e $\underline{\widehat{\beta}}^{(t-1)}$ é a solução do passo anterior. Quando a precisão é estabelecida, o processo é interrompido e o estimador é dado por $\underline{\widehat{\beta}} = \underline{\widehat{\beta}}^{(t)}$. Soluções equivalentes são obtidas através de outros métodos interativos, como por exemplo, o de Mínimos Quadrados Ponderados Iterativos, cuja implementação pode ser encontrada em pacotes estatísticos como o GLIM (General Linear Interactive Model), entre outros. Uma discussão mais detalhada sobre esses e outros métodos, pode ser encontrada em McCullagh e Nelder [1989] e Aidar [1992].

2.3.4 Generalidades da classificação usando MRL

Além do pacote estatístico citado acima, outros como S-PLUS, SAS, BMDP, ect., apresentam formas eficientes de estimação de $\underline{\beta}$. O interesse, contudo, está no significado prático desse parâmetro, como é visto em Regressão Linear onde os coeficientes representam o efeito de cada variável explicativa. Aqui, o parâmetro também representa uma medida de efeito da variável explicativa em relação ao logito da probabilidade de ocorrência da variável resposta ($\lambda = \log(\theta/1-\theta)$). Para um melhor entendimento, considere que uma determinada componente de $\underline{x}$, digamos x_j, assume somente os valores 0 ou 1, de forma que possamos avaliar uma razão entre os modelos nos dois casos, conhecida como a razão de *odds* para a variável x_j, dada por,

$$\Psi_j = \frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}$$

onde o índice 1 em θ_1 indica que o modelo inclui a variável, isto é,

$$g(\theta_1) = \log(\theta_1/1-\theta_1) = \left(\underline{x}'\underline{\beta}\right)^* + \beta_j$$

e o índice 0 em θ_0 indica que o modelo não inclui a variável, isto é,

$$g(\theta_0) = \log(\theta_0/1-\theta_0) = \left(\underline{x}'\underline{\beta}\right)^*,$$

e $\left(\underline{x}'\underline{\beta}\right)^*$ não inclui a variável x_j nem o parâmetro β_j. Portanto a razão entre os modelos com e sem a variável x_j pode ser escrita como:

$$\Psi_j = \frac{\exp(\left(\underline{x}'\underline{\beta}\right)^* + \beta_j)}{\exp(\underline{x}'\underline{\beta})^*} = \exp(\beta_j). \tag{2.42}$$

Aplicando o logaritmo temos

$$\beta_j = \log\left(\Psi_j\right).$$

Portanto, β_j representa a mudança no logaritmo da razão entre as duas probabilidades de ocorrência do evento, condicionada à categoria da variável x_j, ou ainda à ponderação em se passar de $x_j = 0$ para $x_j = 1$.

Quando x_j é uma variável contínua, a idéia pode ser estendida em termos de incremento da variável explicativa na mesma unidade de escala da variável, sendo β_j a mudança da razão das probabilidades de ocorrência do evento, condicionada ao incremento desta variável.

Lembramos que o uso mais frequente do MRL é no estudo de variável com resposta binária, mas o modelo também se aplica ao caso da resposta apresentar mais de duas possibilidades. Embora não seja nosso objetivo nesse trabalho, este tipo de abordagem pode servir também como competidor ao tipo de abordagem feita via AD, principalmente quando as variáveis explicativas não seguem uma distribuição Normal. A seguir resumimos os passos da classificação de um objeto, $\underline{z}$, em uma de duas classes possíveis:

1)Encontre as estimativas de $\underline{\beta}, \underline{\widehat{\beta}} = (\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_p)$, através dos dados.

2) Calcule: $\widehat{\theta}(\underline{z}) = \frac{\exp\{\widehat{\alpha}+\underline{z}'\underline{\widehat{\beta}}\}}{1+\exp\{\widehat{\alpha}+\underline{z}'\underline{\widehat{\beta}}\}}$, com $\widehat{\lambda} = \log(\frac{\widehat{\theta}(\underline{z})}{1-\widehat{\theta}(\underline{z})}) = \widehat{\alpha} + \underline{z}'\underline{\widehat{\beta}}$.

2) Se $\widehat{\theta}(\underline{z}) > 0,5$ associe $\underline{z}$ à classe π_1 ou associe $\underline{z}$ à classe π_2 caso contrário.

(2.43)

Observe que o valor crítico 0,5, significa que ambas as classes são supostas igualmente prováveis. Portanto, dependendo da aplicação, esse valor pode ser modificado.

Outro tema que está fortemente ligado à estrutura do MRL é a seleção de variáveis explicativas, pois, o primeiro objetivo do modelo é descrever o comportamento da variável resposta, Y, com o conjunto de variáveis explicativas através dos seus efeitos. Esse problema será estudado com mais detalhes no Capítulo 3. A próxima seção aborda o problema da

classificação de um objeto em uma de duas possíveis classes, quando existe alguma restrição às observações referentes a uma das classes, isto é, não temos acesso às observações quando $Y_i = 0$.

2.4 Modelo de Classificação Usando apenas uma das Classes

2.4.1 Introdução

Os dois métodos de classificação apresentados nas Seções 2.2 e 2.3, consideravam a suposição de que observações de ambas as classes, π_1 e π_2, são disponíveis. Nessa Seção abordamos o mesmo problema da classificação de um objeto, mas com uma restrição: não temos como obter as observações de uma das classes para estimar os parâmetros do modelo de classificação. A proposta é apresentar um modelo baseado apenas nas observações de uma das classes, ou ainda baseado na distribuição do vetor aleatório, $\underline{x}$, condicionada apenas a uma das classes. Na visão estatística, o interesse é testar se um novo objeto é membro ou não dessa classe, digamos π_1, que temos acesso as informações, supondo que $p(\underline{x}/\pi_1)$ seja conhecida.

Considere R o espaço geral p-dimensional que contém os pontos das observações do vetor $\underline{x}$. O objetivo é encontrar uma sub-região de R, digamos R_1, de modo que possamos ou não associar $\underline{x}$ na classe π_1, se $\underline{x}$ pertencer ou não a região R_1, respectivamente. Como na seção 2.2, quando as classes não são linearmente separáveis, qualquer método que se propõe encontrar essa sub-região, carrega consigo as chances de classificação errônea. O método pode deixar de associar um verdadeiro elemento da classe ou vice-versa, associar erroneamente um elemento de outra classe. Estamos assim, diante de um teste de hipóteses estatístico, fazendo sentido definir:

$$\begin{aligned}\alpha_1^* &= \Pr(\text{ Rejeitar } H_0 \text{ / } H_0 \text{ é correta }) = \int_{R-R_1} p(\underline{x}/\pi_1)d\underline{x} \qquad \text{e} \\ \alpha_2^* &= \Pr(\text{ Aceitar } H_0 \text{ / } H_0 \text{ é falsa }) = \int_{R_1} p(\underline{x}/\pi_2)d\underline{x},\end{aligned} \tag{2.44}$$

onde α_1^* denota a probabilidade de erro do Tipo I, α_2^* denota a probabilidade de erro do Tipo II, e H_0 é a hipótese que associa $\underline{x}$ em π_1. O superescrito * em α_1^* e α_2^* é usado aqui para diferenciar a definição de α_1 e α_2 dada em (2.3). Sem perda de generalidade supomos que π_1 seja a classe a qual temos acesso às observações. Lembramos, ainda, que α_2^* não pode ser avaliada, em termos teóricos, devido a restrição às observações da classe π_2, sendo necessário um estudo de simulação para sua estimação. Neste caso, o método para determinar R_1 deve considerar apenas α_1^*.

2.4.2 Estimação de parâmetros sob normalidade

A suposição com respeito à distribuição do vetor $\underline{x}$ condicionado a classe π_1 ser Normal, adotada na Seção 2.2, é também adotada aqui com a finalidade de simplificar o caso teórico. Apresentamos, a seguir, uma regra de classificação baseada na distância de uma nova observação ao centro de uma distribuição Normal.

Considere que a *pdf* de $\underline{x}$ condicionada a classe π_1 seja Normal com vetor de média $\underline{\mu}$ e matriz de covariância Σ, ambos conhecidos, isto é,

$$p(\underline{x}/\pi_1) \sim N_p(\underline{\mu}, \Sigma). \tag{2.45}$$

seja $\underline{z}$ uma nova observação que queremos verificar se pertence a essa classe. O objetivo aqui é definir uma medida de distância, $d(\underline{z})$, entre $\underline{z}$, um indivíduo supostamente desconhecido e o centro da distribuição $N_p(\underline{\mu}, \Sigma)$, com base apenas nas observações da classe π_1, de modo

que possamos decidir por associar $\underline{z}$ com π_1 se

$$d(\underline{z}) \leq C, \tag{2.46}$$

e decidir por não associar se caso contrário, onde C é escolhida, depois de fixado α_1^*,satisfazendo

$$\Pr(d(\underline{x}/\pi_1) > C) = \alpha_1^*, \tag{2.47}$$

onde $d(\underline{x}/\pi_1)$, representa a distância de um elemento com *pdf* $N_p(\underline{\mu}, \Sigma)$ ao centro dessa distribuição. O significado de (2.47) é que apenas $\alpha_1^*(100)\%$ das observações de um vetor aleatório com a distribuição $N_p(\underline{\mu}, \Sigma)$ tem distância ao centro da distribuição superior a C, ou equivalentemente, que apenas $\alpha_1^*(100)\%$ deles não pertencem a R_1. Sendo, portanto, uma maneira de controlarmos a probabilidade de falsa rejeição.

Assumindo que a distribuição de $\underline{x}$ condicionada à classe π_1 é $N_p(\underline{\mu}, \Sigma)$,e substituindo sua *pdf* na equação (2.44), podemos escrever α_1^* com:

$$\alpha_1^* = \int_{d(\underline{x})>C} \left[(2\pi)^n \, |\Sigma|\right]^{-\frac{1}{2}} \exp\{-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}\right)\} d\underline{x}, \tag{2.48}$$

onde $|\Sigma|$ é o determinante da matriz Σ. Uma forma convencional de definir $d(\underline{x}/\pi_1)$ é relacionar a rejeição da hipótese de associação com uma pequena probabilidade α_1^*, isto é, condicionar as observações a uma região pouco provável, e isso é feito fazendo com que

$$\exp\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{x}-\underline{\mu}\right)\},$$

seja pequeno, ou de forma equivalente fazendo com que

$$\left(\underline{x}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{x}-\underline{\mu}\right), \tag{2.49}$$

seja grande. Na literatura (2.49) é conhecida como o quadrado da distância de Mahalanobis ao centro da distribuição de um vetor Normal, tendo densidade elíptica constante, sendo também conhecida como densidade de contorno constante. Pode ser mostrado que (2.49) tem uma distribuição Qui-quadrado, com p graus de liberdade (χ^2_p) (ver Mardia et al [1979]). A expressão de uma variável aleatória com *pdf* Qui-quadrado está definida no Apêndice A-4. Portanto, podemos definir nossa medida de distância por (2.49) e depois de fixado α_1^*, encontrar C tal que (2.47) se cumpra. Em outras palavras, se uma nova observação $\underline{z}$, vem de fato da classe π_1, espera-se que $d(z)=\left(\underline{z}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{z}-\underline{\mu}\right)$, seja inferior a C, com probabilidade α_1^*.

Os passos do procedimento de classificação de um vetor aleatório Normal, com seus parâmetros conhecidos, podem ser resumidos em:

1) Fixe α_1^* e encontre C tal que $\Pr(\chi^2_p > C) = \alpha_1^*$, onde $\Pr(\chi^2_p > C) = \alpha_1^*$ representa que a probabilidade de uma variável aleatória com distribuição Qui-quadrado, com p graus de liberdade, exceder uma constante C é igual a α_1^*.

2) Calcule $d(\underline{z})=\left(\underline{z}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{z}-\underline{\mu}\right)$.

UNICAMP
BIBLIOTECA CENTRAL

3) Associe $\underline{z}$ à classe π_1 se $d(\underline{z}) \leq C$, ou não, se $d(\underline{z}) > C$.

(2.50)

Lembramos que existem outras formas de resolver esse problema, mas certamente exigiria um procedimento mais complicado.

2.4.3 Estimação amostral de parâmetros

O procedimento de classificação de um novo objeto, adotado em (2.50), assume que os parâmetros da distribuição Normal, $\underline{\mu}$ e Σ, são conhecidos, mas na prática eles devem ser estimados abrindo uma discussão sobre a melhor forma de escolher a medida de distância entre um indivíduo e o centro da distribuição. Um procedimento adotado por Nelson et al [1994] foi substituir os parâmetros, $\underline{\mu}$ e Σ, em (2.49), pelos seus respectivos estimadores não viciados, e usar a aproximação da distância amostral de um vetor aleatório com distribuição $N(\underline{\mu}, \Sigma)$ pela distribuição Qui-quadrado, sendo a distância amostral definida por:

$$\widehat{d}(\underline{x}) = \left(\underline{x} - \underline{\widehat{\mu}}\right)' \widehat{\Sigma}^{-1} \left(\underline{x} - \underline{\widehat{\mu}}\right), \tag{2.51}$$

com

$$\underline{\widehat{\mu}} = \overline{\underline{x}} = \frac{1}{n} X' \mathbf{1_{n,1}} \qquad \text{e} \qquad \widehat{\Sigma} = \frac{1}{n-1} X'(I - \frac{1}{n}\mathbf{1_{n,n}})X = \frac{1}{n-1} S,$$

onde X é uma matriz (n x p) contendo n observações do vetor de p características, $\mathbf{1_{a,b}}$ é uma matriz (**a** x **b**) com todos os elementos iguais a 1 e I é a matriz identidade de ordem n.

O problema é que a expressão (2.51) não segue uma distribuição Qui-quadrado como em (2.49), devido a $\underline{\widehat{\mu}}$ e $\widehat{\Sigma}$ serem aleatórios. De fato, como os estimadores são funções amostrais, podem assumir diferentes valores a cada processo amostral, tendo cada estimador uma *pdf*

definida, como pode ser visto no Lema 2.4.1 a seguir. A distribuição exata de (2.51) é justamente o que propõe o Teorema 2.4.1 e isso tem sido uma das principais motivações desse trabalho, que tem como objetivo justificar analiticamente a aplicação dessa medida de distância na prática. Antes, porém, vamos introduzir alguns lemas e definições que servirão de suporte na sua apresentação.

Definição 2.4.1(Mardia et al [1979]) - Seja $M = X'X$ uma matriz (p x p), sendo X uma matriz (n x p), onde as linhas de X representam n observações amostrais do vetor aleatório, $\underline{x}$, com distribuição $N_p(0, \Sigma)$. Nesse caso M é dito ter distribuição Wishart com parâmetros de escala Σ e n graus de liberdade e escrevemos

$$M \sim W_p(\Sigma, n).$$

A expressão da *pdf* de uma matriz aleatória M com distribuição de Wishart com parâmetro de escala e n graus de liberdade pode ser encontrada no Apêndice A-4.

A definição acima é na verdade uma versão multivariada da distribuição Qui-quadrado. Basta observar o caso particular quando $p = 1$, $W_1(\sigma^2, n)$, com $M = \underline{x}'\underline{x}$, onde $\underline{x}$ representa n obsevações de uma variável aleatória com distribuição $N_1(0, \sigma^2)$. Nesse caso $M \sim \sigma^2\chi^2_n$. Outros resultados importantes que serão usados na demonstração do Teorema 2.4.1 seguem.

Lema 2.4.1 - Se $\overline{\underline{x}} = \frac{1}{n}X'\mathbf{1_{n,1}}$ e $S = X'(I - \frac{1}{n}\mathbf{1_{n,n}})X$, onde X é uma matriz ($n$ x p) cujas linhas são n observações de um vetor aleatório com distribuição $N_p(\underline{\mu}, \Sigma)$, então $\overline{\underline{x}}$ e S são independentes com as seguintes distribuições

$$\underline{\bar{x}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma) \qquad \text{e} \qquad S \sim W_p(\Sigma, n-1)\ .$$

Prova:(ver Mardia et al [1979], pag.68)•

A importância do lema acima está na generalização do caso unidimensional, onde a média e a variância amostral são independentes e seguem distribuições de probabilidades conhecidas. Neste caso, suas distribuições são versões multivariadas das respectivas distribuições univariadas.

Lema 2.4.2 - Se M tem distribuição $W_p(\Sigma, n)$, então

$$\Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}} \sim W_p(I, n).$$

Prova:(ver Mardia et al [1979], pag. 67)•

Na verdade o Lema 2.4.2 é uma caso particular de tranformações da classe de matrizes da família Wishart, onde produtos do tipo $B'MB$, continuam pertencendo à família Wishart, e B deve ter dimensão compatível com M. A prova desse fato pode ser feita facilmente partindo da Definição 2.4.2. A principal importância, entretanto, é que o lema define uma versão padronizada de uma matriz aleatória com distribuição de Wishart.

Introduzimos a seguir expressões envolvendo variáveis com distribuições multivariadas Normais e matrizes com distribuição Wishart, tais como $\underline{d}' M^{-1} \underline{d}$, onde $\underline{d}$ tem distribuição Normal multivariada, M tem distribuição Wishart e $\underline{d}$ e M são independentes. Segue a definição formal dessa forma quadrática que foi um trabalho iniciado por Hotelling (1931).

Definição 2.4.2(Mardia et al [1979]) - Se λ pode ser escrito como $\lambda = n\underline{d}' M^{-1}\underline{d}$, onde $\underline{d}$ e M são independentes e distribuídos, respectivamente, segundo uma $N_p(\underline{0}, I)$ e $W_p(I, n)$, diremos que λ tem uma distribuição T^2 de Hotelling com parâmetros p e n, e escrevemos

$$\lambda \sim T^2(p, n).$$

A expressão da *pdf* de uma variável com distribuição T^2 de Hotelling pode ser encontrada no Apêndice A-4. Uma importante aplicação da definição acima segue quando conhecemos os verdadeiros valores dos parâmetros da distribuição Normal, e estamos interessados numa forma alternativa de avaliar a distância (2.49). Isso é o que diz o lema a seguir:

Lema 2.4.3 - Se $\underline{x}$ e M são independentemente distribuídos, respectivamente, por $N_p(\underline{\mu}, \Sigma)$ e $W_p(\Sigma, n)$, então

$$\lambda = n(\underline{x} - \underline{\mu})' M^{-1}(\underline{x} - \underline{\mu})$$

segue uma distribuição T^2 de Hotelling com parâmetros p e n.

Prova: Considere $M = X'X$ e X uma matriz de dados como foi definido anteriormente no Lema 2.4.1. Por suposição $\underline{x} \sim N_p(\underline{\mu}, \Sigma)$, escrevendo $\underline{d}^* = \Sigma^{-\frac{1}{2}}(\underline{x} - \underline{\mu})$ e como uma função linear de variáveis Normais tem também distribuição Normal (ver James [1981]), segue que $\underline{d}^*$ $\sim N_p(\underline{0}, I)$. Observe agora que por hipótese $M \sim W_p(\Sigma, n)$, e pelo Lema 2.4.2, $M^* = \Sigma^{-\frac{1}{2}} M$ $\Sigma^{-\frac{1}{2}} \sim W_p(I, n)$. Assim, pela Definição 2.4.2,

$$\lambda = n\underline{d}^{*\prime} M^{*-1}\underline{d}^{*} = n(\underline{x}-\underline{\mu})' M^{-1}(\underline{x}-\underline{\mu}) \sim T^2(p,n)\bullet$$

O caso particular da Definição 2.4.2 com $p = 1$, $T^2(1,n)$, é na verdade o quadrado de uma variável aleatória t-Student com n graus de liberdade. A importância dessa particularidade é que a partir daí podemos relacionar a distribuição T^2 com outras. Mais precisamente, o quadrado de uma variável aleatória com distribuição t-Student tem uma distribuição $F_{1,n}$, a conhecida distribuição F de Snedecor com 1 grau de liberdade no denominador e n graus de liberdade no numerador. Em outras palavras, a distribuição $T^2(1,n)$ coincide com a $F_{1,n}$. A generalização desse fato está no seguinte lema:

Lema 2.4.4 - $T^2(p,n) = \frac{np}{n-p+1} F_{p,n-p+1}$.

Prova:(ver Mardia et al [1979], pag.74)•

Agora temos condições de avaliar o significado da forma quadrática,

$$(n-1)(\underline{x}-\overline{\underline{x}})' S^{-1}(\underline{x}-\overline{\underline{x}}),$$

onde os parâmetros da expressão (2.49) (μ e Σ) são substituídos pelos seus respectivos estimadores não viciados e assim possamos testar a hipótese de um novo objeto pertencer ou não a classe das distribuições Normais, com base em dados amostrais.

Teorema 2.4.1 - Sejam $\underline{\overline{x}}$ e S como definidos anteriormente no Lema 2.4.1, se $\underline{x}$ tem distribuição $N_p(\underline{\mu}, \Sigma)$, então

$$(n-1)\left(\frac{n}{n+1}\right)(\underline{x}-\underline{\overline{x}})'S^{-1}(\underline{x}-\underline{\overline{x}}) \sim T^2(p, n-1)$$

e consequentemente,

$$\left(\frac{n(n-p)}{(n+1)p}\right)(\underline{x}-\underline{\overline{x}})'S^{-1}(\underline{x}-\underline{\overline{x}}) \sim F_{p,n-p}.$$

Prova: Pelo Lema 2.4.1, $\underline{\overline{x}}$ e S são independentes e distribuídos por $N_p(\underline{\mu}, \frac{1}{n}\Sigma)$ e $W_p(\Sigma, n-1)$, respectivamente. Por hipótese $\underline{x}$, uma nova observação, tem distribuição $N_p(\underline{\mu}, \Sigma)$ e definindo

$$\underline{d} = \left(\frac{n+1}{n}\Sigma\right)^{-\frac{1}{2}}(\underline{x}-\underline{\overline{x}}),$$

segue que $\underline{d}$ tem distribuição Normal, pois é função linear de vetores com distribuições Normais multivariadas(James [1981]), com

$$E(\underline{d}) = 0 \qquad \text{e} \qquad Cov(\underline{d}) = I.$$

Definindo agora

$$M = \Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}},$$

temos pelo Lema 2.4.2 que, $M \sim W_p(I, n-1)$. Portanto $\underline{d}$ e M satisfazem à Definição 2.4.2, isto é, $\underline{d}$ segue uma distribuição $N_p(\underline{0}, I)$ e M segue uma distribuição Wishart padronizada, independente de $\underline{d}$, pois S é independente de $\overline{\underline{x}}$ (Lema 2.4.1) e $\underline{x}$ é uma nova observação, portanto independente de $\overline{\underline{x}}$ e S . Assim podemos concluir que:

$$\lambda = (n-1)\,\underline{d}'M^{-1}\underline{d} = n\left(\frac{n-1}{n+1}\right)(\underline{x}-\overline{\underline{x}})'S^{-1}(\underline{x}-\overline{\underline{x}}) \sim T^2(p, n-1),$$

provando a primeira parte. A segunda parte segue da aplicação direta do Lema 2.4.4, isto é,

$$T^2(p, n-1) = \frac{(n-1)p}{n-p}F_{p,n-p}\bullet$$

Observe que o objetivo do teorema acima é formalizar uma estatística teste ou uma medida de distância com uma distribuição conhecida e tabelada na versão univariada. Em outras palavras, através do teorema acima podemos formalizar a redução da regra de classificação p-dimensional para apenas uma dimensão. Nessa base, podemos testar se um novo objeto, $\underline{z}$, pertence ou não a classe das distribuições Normais, seguindo os passos desenvolvidos em (2.50) com os estimadores necessários. A seguir apresentamos o resumo dos passos para classificar um novo objeto na classe das distribuições Normais:

1) Fixe α_1^* e encontre C tal que $\Pr(F_{p,n-p} > C) = \alpha_1^*$.

2) Calcule $d(\underline{z}) = \frac{n(n-p)}{(n+1)p}(\underline{z}-\overline{\underline{x}})'S^{-1}(\underline{z}-\overline{\underline{x}})$.

3) Associe ou não $\underline{z}$ à classe π_1 se $d(\underline{z}) \leq C$ ou $d(\underline{z}) > C$, respectivamente.

(2.52)

Na próxima seção apresentaremos um estudo simulado onde esse método de classificação, além dos demais vistos anteriomente, são comparados através de suas taxas de erros na classificação de objetos supostamente desconhecidos. E no Capítulo 3, esse método será utilizado como base na seleção de variáveis que mais contribuem na construção da medida de distância do passo 2, acima.

2.5 Aplicação com Dados Simulados

2.5.1 Introdução

O objetivo desta Seção é apresentar um estudo simulado dos métodos de calssificação apresentados nas três seções anteriores em relação a um problema típico de classificação envolvendo apenas duas classes. Formalmente, queremos avaliar o desempenho de cada método, através de suas taxas de erros de classificação, e compará-los com seus principais competidores: O classificador de Bayes, quando temos acesso às observações de ambas as classes, e o classificador que utiliza a aproximação pela distribuição Qui-quadrado, quando temos acesso às observações de apenas uma das classes.

São apresentados a seguir os resumos de cada um dos métodos competidores: Classificador de Bayes e o Classificador usando a aproximação pela distribuição Qui-quadrado, uma simulação para dois conjuntos de amostras e os seu respectivos resultados através das taxas de erros.

2.5.2 O Classificador de Bayes

O classificador de Bayes supõe uma *pdf a priori* para os parâmetros da distribuição condicional de $\underline{x}$ dada a classe π_i. Por exemplo, no caso de duas classes com distribuição Normais multivariadas, o método introduz uma *pdf a priori* para o conjunto de parâmetros $\theta = (\underline{\mu}_1, \underline{\mu}_2, \Sigma), p(\theta)$, (ver Aitchison et al [1977]). A proposta então é avaliar a verossimilhança

de uma observação, $\underline{x}$, dada a matriz de dados X, supondo que $\underline{x}$ pertence à i-ésima classe π_i, $i = 1, 2$. A verossimilhança é dada pela média das *pdf's* de $\underline{x}$ dado θ, $p(\underline{x}/\theta)$, com respeito à *pdf a posteriori* de θ dado X, $p(\theta/X)$, isto é:

$$L_i(\underline{x}/X) = \int p_i(\underline{x}/\theta)\ p(\theta/X)\ d\theta, \qquad i = 1, 2,$$

onde $p(\theta/X)$ é proporcional ao produto $p(X/\theta)\ p(\theta)$. Decidimos associar $\underline{x}$ em π_1 se $L_1(\underline{x}/X) \geq L_2(\underline{x}/X)$.

O caso particular da classificação de um indivíduo com p características em uma das duas classes com distribuição Normal multivariada e com matrizes de covariância iguais, é apresentado em Mardia et al [1979], pag. 317. Considerando $p(\theta)$ não informativa, isto é, $\theta = (\underline{\mu}_1, \underline{\mu}_2, \Sigma)$ de acordo com o princípio da invariância de Jeffreys (ver Jeffreys [1961]), eles concluiram que a i-ésima verossimilhança L_i é proporcional ($\propto$) a uma quantidade amostral, isto é:

$$L_i(\underline{x}/X) \propto \left\{ 1 + (n-1-p)^{-1} (\underline{x} - \overline{\underline{x}}_i)' \left[\frac{n\left(1 + \frac{1}{n_i}\right)}{(n-p-1)} S \right]^{-1} (\underline{x} - \overline{\underline{x}}_i) \right\}^{-\frac{n-1}{2}}, \qquad (2.53)$$
$$i = 1, 2,$$

onde $n = n_1 + n_2$, $nS = n_1 S_1 + n_2 S_2$, $S_i = \frac{1}{n_i} X_i' (I_{n_i} - \mathbf{1}_{\mathbf{n_i},\mathbf{n_i}}) X_i$ e X_i a matriz (n_i x p) de dados com n_i observações do vetor $\underline{x}$ pertencente à classe π_i ($i = 1, 2$). Um resumo do procedimento de classificação de um indivíduo através do Classificador de Bayes segue:

1 - Para uma nova observação, $\underline{z}$, calcule $L_i(\underline{z}/X)$ de acordo com (2.53).

2 - Associe $\underline{z}$ à classe π_1 se $L_1(\underline{z}/X) \geq L_2(\underline{z}/X)$ e à classe π_2 se $L_1(\underline{z}/X) < L_2(\underline{z}/X)$.

(2.54)

2.5.3 Classificador usando aproximação pela distribuição Qui-quadrado

O outro método que destacamos aqui, foi apresentado em Nelson et al [1994] num problema de verificação de assinaturas. Supondo que apenas uma das classes poderia ser observada. Eles consideraram que o vetor de observações $\underline{x}$, com p características, tinha distribuição Normal multivariada com média $\underline{\mu}$ e matriz de covariância DVD, isto é:

$$\underline{x} \sim N_p(\underline{\mu}, DVD),$$

onde $\underline{\mu} = E\{\underline{x}\}$ e DVD é a matriz de covariância, sendo D uma matriz (p x p) diagonal cujos elementos da diagonal representam os desvios padrões das coordenadas (ou características) de $\underline{x}$ e V a parte da matriz de covariância que representa a correlação entre as características. Essa forma de definir a matriz de covariância $\Sigma = DVD$ é utilizada pelos autores para futuras suposições a respeito da matriz Σ, como por exemplo independência entre as características ($V = I$).

Eles propuseram a distância amostral:

$$\widehat{d}(\underline{x}) = (\underline{x} - \widehat{\underline{\mu}})'(\widehat{D}\widehat{V}\widehat{D})^{-1}(\underline{x} - \widehat{\underline{\mu}}), \tag{2.55}$$

onde

$$\widehat{\underline{\mu}} = \frac{1}{n}X'1_{n,1}, \qquad \widehat{D} = \frac{1}{(n-1)^{\frac{1}{2}}} diag\left(\left[\sum_i (\mathrm{x}_{i1} - \overline{\mathrm{x}}_1)^2\right]^{\frac{1}{2}}, ..., \left[\sum_i (\mathrm{x}_{ip} - \overline{\mathrm{x}}_p)^2\right]^{\frac{1}{2}}\right)$$

Eles aproximaram $\widehat{d}(\underline{x})$ pela distribuição Qui-quadrado e estabeleceram o seguinte procedimento de classificação:

1 - Fixe α e encontre C tal que $\Pr(\chi_p^2 \geq C) = \alpha$. ($C$ um valor tabelado pela distribuição Qui-qudrado com p graus de liberdade).

2 - Para uma nova observação, $\underline{z}$, calcule $\widehat{d}(\underline{z})$ através da equação (2.55).

3 - Associe $\underline{z}$ em π_1 (a classe que temos acesso) se $\widehat{d}(\underline{z}) \leq C$ e não associe se $\widehat{d}(\underline{z}) > C$.

(2.56)

2.5.4 Simulação dos dados

Através do aplicativo *MATLAB* 4.0, foram gerados dois conjuntos de dados, ambos considerando duas classes com distribuição Normal multivariada: o primeiro com vetores de médias e matrizes de covariância diferentes e o segundo com vetores de médias iguais e matrizes de covariância proporcionais. Dentro de cada conjunto também foram gerados dois grupos denotados por dados de treinamento e dados para testes: os dados de treinamento são usados especificamente para a construção dos modelos, enquanto os dados para testes são usados para a validação dos modelos.

No primeiro conjunto (com vetores de médias e matrizes de covariância defirentes) as matrizes de dados de treinamento, X_1 e X_2, foram geradas supondo conhecidos os parâmetros da distribuição, $\underline{\mu}_i$ e Σ_i $(i = 1, 2)$, sendo X_1 uma matriz (100 x 42) de dados $\left(X_1 = (\underline{x}_1^{(1)}, \underline{x}_2^{(1)}, ..., \underline{x}_{100}^{(1)})'\right)$, referentes a 100 observações do vetor $\underline{x}^{(1)}$ com 42 características pertencentes à classe π_1, e X_2 uma matriz (50 x 42) de dados $\left(X_2 = (\underline{x}_1^{(2)}, \underline{x}_2^{(2)}, ..., \underline{x}_{50}^{(2)})'\right)$, referentes a 50 observações do vetor $\underline{x}^{(2)}$ com 42 características pertencentes à classe π_2, através de:

$$\begin{aligned} \underline{x}_i^{(1)} &= \underline{r}_i \Sigma_1^{\frac{1}{2}} + \underline{\mu}_1, \qquad i = 1, 2, .., 100, \qquad \text{e} \\ \underline{x}_j^{(2)} &= \underline{r}_j \Sigma_2^{\frac{1}{2}} + \underline{\mu}_2, \qquad j = 1, 2, .., 50, \end{aligned} \tag{2.57}$$

onde as observações de $\underline{r}$ são amostras geradas a partir da distribuição $N_p(\underline{0}, I)$. Portanto $\underline{x}^{(1)}$ e $\underline{x}^{(2)}$ têm distribuição $N_p(\underline{\mu}_1, \Sigma_1)$ e $N_p(\underline{\mu}_2, \Sigma_2)$, respectivamente. A proposta aqui foi simular uma situação parecida com a realidade, onde o acesso às observações de uma das classes é um pouco mais difícil que a outra. As matrizes de dados para testes seguiram exatamente o mesmo procedimento de geração dos dados de treinamento. A diferença, no entanto, foi a geração de 1.000 observações para cada classe. Os vetores de médias $\underline{\mu}_1$ e $\underline{\mu}_2$ e as matrizes de covariância Σ_1 e Σ_2, supostamente conhecidas, foram obtidas através de uma amostra de 200 observações reais de um vetor com 42 características da assinatura de um indivíduo, verdadeiras e falsificadas, sendo 100 verdadeiras e 100 falsificadas, coletadas na Universidade de Cornell (EUA) e apresentados em Lee e Berger [1996]. A descrição dos vetores de médias, assim como, dos desvios padrões utilizados nesta simulação é apresentada no Apêndice B, além da definição de cada uma das 42 características.

No segundo conjunto (com vetores de médias iguais e matrizes de covariância proporcionais), os dados foram gerados com base apenas nas observações da classe π_1. Supondo conhecidos μ_1 e Σ_1, e seguindo o mesmo procedimento do primeiro conjunto, foram geradas 150 observações para treinamento e 2.000 para teste, através de

$$\underline{x}_i^{(1)} = \underline{r}_i \Sigma_1^{\frac{1}{2}} + \underline{\mu}_1, \qquad i = 1, 2, .., 1100,$$

e

$$\underline{x}_j^{(2)} = \underline{r}_j \Sigma_1^{*\frac{1}{2}} + \underline{\mu}_1, \qquad j = 1, 2, .., 1050,$$

onde 100 observações de $\underline{x}^{(1)}$ foram usadas para treinamento e 1000 para teste, enquanto 50 observações de $\underline{x}^{(2)}$ foram usadas para treinamento e 1000 para teste. A matriz de covariância

Σ_1^* foi obtida a partir da decomposição da matriz de covariânca Σ_1 em auto-valores e auto-vetores. Em particular, o maior auto-valor de Σ_1 foi acrescido em 10% e então calculado Σ_1^* através de uma operação inversa. Formalmente, primeiro encontra-se Γ e Δ tal que

$$\Gamma'\Sigma_1\Gamma = \Delta,$$

onde Γ é uma matriz ortogonal (com $\Gamma'\Gamma = \Gamma\Gamma' = I$, onde I é uma matriz identidade de ordem p), que representa a matriz de auto-vetores de Σ_1, e Δ uma matriz diagonal, cujos elementos da diagonal, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, representam os auto-valores associados aos auto-vetores Γ. Então, calcula-se Σ_1^* através de

$$\Sigma_1^* = \Gamma\Delta^*\Gamma'$$

onde $\Delta^* = diag(1.1\lambda_1, \lambda_2, ..., \lambda_p)$. O objetivo aqui foi estudar o caso onde as duas classes apresentam características semelhantes, tendo em média os mesmos valores, mas com matrizes de covariância proporcionais. De fato esse é um caso típico quando uma classe representa uma tentativa de imitação da outra, bastante comum em estudos de verificação de assinaturas.

2.5.5 Resultados da simulação

O objetivo nesta simulação foi apresentar uma situação onde as classes sejam linearmente separáveis, e em seguida apresentar uma situação onde as classes não sejam linearmente separáveis, para que possamos avaliar o desempenho dos métodos de classificação em ambas

situações. Os valores dos parâmetros foram obtidos a partir de um caso real e utilizados para transformação de amostra de vetores com distribuição Normal padronizada em amostra de vetores com os respectivos valores. Observe que μ_1 e Σ_1 representam a média e covariância amostral de assinaturas genuínas, enquanto que μ_2 e Σ_2 representam a média e covariância amostral de assinaturas falsas de um mesmo indivíduo coletadas em laboratório. Entretanto seus valores foram supostos como os verdadeiros valores dos respectivos parâmetros e utilizados para gerar nossos dois conjuntos amostrais.

As Tabela 2.2 e Tabela 2.3, a seguir, apresentam as taxas de erros de classificação de acordo com os métodos descritos anteriormente, isto é, os erros dos Tipos I e II estimados, onde as quantidades que aparecem nos corpos das tabelas representam o número (em percentual) de vezes em que o respectivo método classificou erroneamente os elementos, sabendo-se que eles realmente pertencem às respectivas classes. Na Tabela 2.2 as taxas referem-se aos dados simulados no primeiro conjunto, onde as classes têm distribuições Normais multivariadas distintas, isto é, $\underline{x}_i^{(1)}$ tem distribuição $N_p(\underline{\mu}_1, \Sigma_1)$ e $\underline{x}_j^{(2)}$ tem distribuição $N_p(\underline{\mu}_2, \Sigma_2)$. Enquanto que na Tabela 2.3 as classes têm distribuições Normais multivariadas semelhantes, com vetores de médias iguais e matrizes de covariâncias proporcionais, isto é, $\underline{x}_i^{(1)}$ tem distribuição $N_p(\underline{\mu}_1, \Sigma_1)$ e $\underline{x}_j^{(2)}$ tem distribuição $N_p(\underline{\mu}_1, \Sigma_1^*)$. Os métodos ultilizados na classificação seguem os resumos descritos de acordo com: AD - Análise Discriminante - resumo (2.21), pag.20, RL - Modelos de Regressão Logística - resumo (2.43), pag.34 T^2 - Aproximação pela distribuição T^2 de Hotelling - resumo (2.52), pag.45, CB - Classificador de Bayes - resumo (2.54), pag.47 e χ^2 - Aproximação pela distribuição Qui-quadrado - resumo (2.56), pag. 49.

Realizamos também a classificação em relação aos dados de treinamento como forma de verificação. Os níveis de significâncias ou as probabilidades teóricas de rejeitar uma observação genuína, adotados nos métodos T^2 e χ^2 foram de 5%, com (42,58) graus de liberdade, respectivamente, para distribuição F de Snedecor e 42 para a distribuição Qui-

quadrado.

Um valor apresentado nas tabelas indica quanto (em percentual) o respectivo método classificou erroneamente uma observação. Por exemplo, na Tabela 2.2 o classificador de Bayes (CB) estimou em zero o erro do Tipo I em relação às amostras de treinamento e teste (não rejeitou qualquer observação como pertencente a classe π_1, quando de fato elas vinham desta classe). Assim como estimou em zero o erro do Tipo II, em relação aos dois tipos de amostras: treinamento e teste (rejeitou todas as observações em relação à classe π_1, quando estas vinham da classe π_2). Já na Tabela 2.3, o mesmo classificador, CB, estimou em 2% o erro do Tipo I em relação à amostra de treinamento e 6,7% em relação à amostra de teste (rejeitou 2 observações dentro de 100, e 67 dentro de 1000, respectivamente, em relação às amostras de treinamento e teste, quando estas vinham de fato da classe π_1). Enquanto que estimou em 20% o erro do Tipo II em relação à amostra de treinamento, e 59,2% em relação à amostra de teste (aceitou 20 observações dentre de 100 e 592 dentre de 1000, como pertencentes a classe π_1, quando de fato elas vinham da classe π_2).

Observe que não apresentamos resultados quando o método de classificação era T^2 ou χ^2, em relação à amostra de treinamento, pois estas observações não foram utilizadas na construção dos modelos de classificação para estes dois tipos de classificadores.

Tabela 2.2 - Taxas de erro de classificação, para o primeiro conjunto, estimadas (em %) para simulação com dados $N_p(\underline{\mu}_1, \Sigma_1)$ e $N_p(\underline{\mu}_2, \Sigma_2)$ (α_1 igual a 5%)

Amostras	CB	AD	RL	T^2	χ^2
Treinamento					
π_1	0,0	0,0	0,0	0,0	2,0
π_2	0,0	0,0	0,0	-	-
Teste					
π_1	0,0	0,0	0,3	5,0	75,5
π_2	0,0	0,0	0,0	0,0	0,0

Os resultados da Tabela 2.2 mostram o bom desempenho de todos os métodos, exceto o χ^2, para os dados de teste. Entre eles podemos destacar que os métodos utilizando Análise Discriminate (AD) e Classificador de Bayes (CB) levam pequena vantagem contra os demais, pois não houve erro de classificação estimado. Embora estejamos tratando com dados em classes linearmente separáveis, um erro estimado de 0,3% para os dados de teste através do RL confirmam os resultados de Press e Wilson [1978], que indicam um melhor desempenho do método que utiliza AD em relação ao método que utiliza RL quando os dados seguem uma distribuição Normal.

Quanto aos métodos usando apenas uma das amostras para treinamento do modelo de classificação, o método usando aproximação T^2 leva grande vantagem em relação aquele que usa aproximação χ^2, pois o erro do Tipo I previsto para as amostras de teste no método

T^2, foi confirmado em 5,0%, enquanto que esse erro usando uma aproximação pela χ^2 ficou muito acima (75,5%).

Tabela 2.3 - Taxas de erro de classificação, para o segundo conjunto, estimadas (em %) para simulação com dados $N_p(\underline{\mu}_1, \Sigma_1)$ e $N_p(\underline{\mu}_1, \Sigma_1^*)$ (α_1 igual a 5%)

Amostras	CB	AD	RL	T^2	χ^2
Treinamento					
π_1	2,0	2,0	2,0	0,0	1,0
π_2	20,0	24,0	22,0	-	-
Teste					
π_1	6,7	4,5	5,6	0,2	72,0
π_2	59,2	63,5	60,1	0,0	0,0

Na Tabela 2.3, destaca-se o excelente desempenho dos métodos que usam apenas as observações de uma das classes para construção dos modelos, T^2 e χ^2, mostrando que, neste caso, em situação que as regiões que contém os pontos das duas classes estão bastante sobrepostas, os métodos usuais de classificação linear têm taxas de erros estimados mais elevados. Observe que, para o erro do Tipo I estimado pela aproximação via χ^2, com amostras de teste, os resultados não são bons (72%), confirmando a necessidade de uma correção nesse tipo de aproximação.

Esses dados simulados serão novamente utilizados nos capítulos seguintes, especialmente no próximo, onde várias técnicas de seleção de variáveis serão comparadas.

2.6 Comentários e Conclusões

Neste capítulo, foram apresentadas algumas técnicas estatísticas usadas na classificação de padrões ou classes. Quando não utilizamos informações *a priori* sob a distribuição do indivíduo condicionada a uma das classes, Press e Wilson [1978] indicam que sob condições de normalidade para o vetor de características, a classificação estatística deve ser feita via AD. Em condicões mais gerais o classificador de Bayes tem a vantagem de poder controlar as taxas de erros dos Tipos I e II na construção do classificador, embora sua implementação exija um planejamento mais específico. Este tipo de controle não pode ser exercido no procedimento via AD, pois a construção do modelo de classificação supõe taxas de erros iguais. Uma alternativa, nesse caso, tem sido a utilização do MRL, pois dependendo da aplicação, um controle sob a aceitação e rejeição pode ser feita em termos da probabilidade da variável resposta assumir o valor 1 ($\Pr(y_i = 1/X)$).

Quando não dispomos das observações de ambas as classes para treinamento, o novo classificador, usando uma aproximação pela distribuição T^2 de Hotelling deve ser preferível à aproximação pela distribuição Qui-quadrado. Inclusive, em situações onde as médias das classes são supostamente iguais, essa aproximação deve ser preferível aos procedimentos que utilizam as observações de ambas as classes na construção dos modelos de classificação.

A principal contribuição deste capítulo foi mostrar a necessidade de uma correção no modelo que utiliza uma aproximação pela distribuição Qui-quadrado, pois a substituição dos verdadeiros valores dos parâmetros pelos estimadores não viciados, impõe na prática, uma situação mais realística.

Capítulo 3

Seleção de Variáveis Explicativas

3.1 Introdução

No capítulo anterior descrevemos algumas técnicas estatísticas para classificar um objeto, supostamente desconhecido, em uma determinada classe ou padrão através de modelos construídos a partir dos dados observados nas respectivas classes. Introduzimos agora uma das etapas de um sistema de reconhecimento de padrões que precede ao problema da classificação: a seleção de variáveis explicativas ou seleção de características discriminantes. O termo Extração de Características tem sido muitas vezes confundido com o termo Seleção de Características, pois ambos definem a necessidade de se reduzir o espaço das características através de algum procedimento de escolha que satisfaça determinadas condições. A diferença básica entre os termos é que a extração procura mapear as características originais em outro espaço com dimensões reduzidas, enquanto que o termo seleção, adotado por este trabalho, está associado ao problema de exclusão de variáveis consideradas redundantes ou menos discriminantes. Na extração, não necessariamente ocorre a redução do número de características originais, pois em geral o problema é definido em termos de tranformações que geram um novo espaço obtido quase sempre através de combinações lineares do espaço original.

Uma das principais motivações para abordar o problema da seleção se deve ao fato de que

o custo associado à construção de um classificador está diretamente relacionado ao número de características utilizadas no modelo de classificação (Mucciardi e Gose [1970]). Nesse sentido, podemos definir o problema da seleção de características, como sendo: determinar um conjunto de características que representem uma observação, enquanto reduza a dimensão do espaço das características. Mas a redução desse espaço pode induzir ao acréscimo nas taxas de erros de classificação, pois algumas características importantes podem vir a ser desprezadas (Lee e Landgrebe [1993]). Portanto, o problema da seleção é determinar um número mínimo de características que representem uma observação sem comprometer o poder de discriminação do classificador.

Este trabalho trata o problema da seleção sob o contexto estatístico, cujo objetivo principal é manter a identidade original das características. Em outras palavras, o objetivo aqui é eliminar as características cujo poder discriminatório possa ser considerado estatisticamente não significativo. Lembrando que, cada processo de classificação, descrito anteriormente, define sua própria medida de distância, sendo portanto, nosso objetivo avaliar a contribuição de cada característica na construção dessas medidas.

São apresentadas inicialmente três técnicas de seleção baseadas nos processos de classificação descritos no Capítulo 2. Na Seção 3.2, a abordagem é feita via Análise Discriminante (AD), onde a contribuição de cada variável é comparada em relação a um distância amostral, chamada de distância Studentizada D^2. Na Seção 3.3, o Modelo de Regressão Logística (MRL) é usado como preditor, e o processo de seleção é descrito para que o ajuste do modelo seja feito apenas com as variáveis cujo efeito seja significativo, dando destaque aos procedimentos sistemáticos (Draper e Smith [1981]). Na Seção 3.4, uma nova proposta de seleção é apresentada seguindo a formulação da AD, mas com a restrição ao uso das observações de uma das classes, onde a contribuição de cada variável ou característica é avaliada em relação a uma distância amostral restrita a apenas uma das classes, denotada aqui por D^2_* para diferenciar da distância amostral que será utilizada na Seção 3.2. Final-

mente, na Seção 3.5, outras técnicas de seleção conhecidas na literatura, são apresentadas de forma resumida, para servir de base para avaliar a nova técnica.

No final do Capítulo apresentamos também um estudo com dados simulados, onde os critérios de seleção são avaliados em termos de erros de classificação. Os dois conjuntos de dados, descritos no capítulo anterior, são novamente utilizados para medir as taxas de erros estimados para cada método.

3.2 Seleção de Variáveis Através de Análise Discriminante

3.2.1 Introdução

A distância de Mahalanobis Δ_p, definida na Subseção 2.2.4, é usada aqui para avaliar se um conjunto com um número menor de variáveis, digamos k, descreve o mesmo desempenho que quando baseada em todas as p variáves ($p \geq k$). A proposta é reescrever essa distância como soma de duas parcelas: a primeira baseada em k variáveis e a segunda baseada nas restantes $(p - k)$. O objetivo é verificar se a contribuição da segunda parcela pode ser considerada estatisticamente não significativa. O critério básico de avaliação será a razão F, definida na Subseção 2.2.5, que indica o quanto uma das parcelas contribui para construção da medida de distância total. Nesse sentido é fundamental um estudo inicial sobre as distribuições das variáveis envolvidas nessa razão. Isso inclui a distribuição da distância Studentizada D^2 e suas relações com as distribuições T^2 de Hotelling e Wishart. Nessa seção, apresentaremos inicialmente a distribuição dessa distância, e em seguida, descrevemos um método de seleção relacionado com ela.

3.2.2 A distribuição da distância Studentizada D^2

Um dos méritos da distribuição T^2 de Hotelling está no fato de que sua distribuição não envolve o parâmetro desconhecido Σ. Aliado a isso, está a aditividade da distribuição Wishart (ver Lema 3.2.1 a seguir). Assim, podemos utilizar a formulação dessas duas distribuições para obter a distribuição proporcional da distância D^2, apresentada na Subseção 2.2.5 como:

$$D_p^2 = m\underline{d}' S^{-1}\underline{d},$$

onde as quantidades amostrais foram definidas naquela seção.

Existem várias formas de se definir uma variável com distribuição T^2 de Hotelling. Uma delas foi apresentada no Lema 2.4.3. Uma outra, que usaremos aqui, considera a suposição de normalidade, e é construída da seguinte forma:

Considere incialmente as matrizes de dados X_1 e X_2 denotadas por:

$$X_1 = \begin{bmatrix} \underline{x}_1^{(1)} \\ \underline{x}_2^{(1)} \\ \cdots \\ \underline{x}_{n_1}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_{11}^{(1)} & \mathrm{x}_{12}^{(1)} & \cdots & \mathrm{x}_{1\mathrm{p}}^{(1)} \\ \mathrm{x}_{21}^{(1)} & \mathrm{x}_{22}^{(1)} & \cdots & \mathrm{x}_{2\mathrm{p}}^{(1)} \\ & \cdots & & \\ \mathrm{x}_{n_1 1}^{(1)} & \mathrm{x}_{n_1 2}^{(1)} & \cdots & \mathrm{x}_{n_1\mathrm{p}}^{(1)} \end{bmatrix} \text{ e } X_2 = \begin{bmatrix} \underline{x}_1^{(2)} \\ \underline{x}_2^{(2)} \\ \cdots \\ \underline{x}_{n_2}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_{11}^{(2)} & \mathrm{x}_{12}^{(2)} & \cdots & \mathrm{x}_{1\mathrm{p}}^{(2)} \\ \mathrm{x}_{21}^{(2)} & \mathrm{x}_{22}^{(2)} & \cdots & \mathrm{x}_{2\mathrm{p}}^{(2)} \\ & \cdots & & \\ \mathrm{x}_{n_2 1}^{(2)} & \mathrm{x}_{n_2 2}^{(2)} & \cdots & \mathrm{x}_{n_2\mathrm{p}}^{(2)} \end{bmatrix}.$$

onde as linha de X_1, $\underline{x}_i^{(1)}$ $(i = 1, 2, ..., n_1)$, e as linhas de X_2, $\underline{x}_j^{(2)}$ $(j = 1, 2, ..., n_2)$, são observações de vetores com distribuição $N_p(\mu_1, \Sigma)$ e $N_p(\mu_2, \Sigma)$, respectivamente.

Considere também a hipótese nula, H_0, que a diferença entre a média das duas classes é igual a um vetor constante, digamos $\underline{\delta}_0$, isto é:

$$H_0 : \underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2 = \underline{\delta}_0,$$

onde μ_1 e μ_2 representam as verdadeiras médias das classes π_1 e π_2, respectivamente. Sejam $\overline{\underline{x}}^{(1)}$, $\overline{\underline{x}}^{(2)}$, $\frac{1}{n_1-1}S_1$, $\frac{1}{n_2-1}S_2$, os vetores de médias e matrizes de covariância amostrais, respectivamente, em relação as duas classes, onde

$$\begin{aligned} \overline{\underline{x}}^{(1)} &= \frac{1}{n_1}X_1'\mathbf{1}_{\mathbf{n_1,1}}, \qquad \overline{\underline{x}}^{(2)} = \frac{1}{n_2}X_2'\mathbf{1}_{\mathbf{n_2,1}} \\ S_1 &= X_1'(I - \frac{1}{n_1}\mathbf{1}_{\mathbf{n_1,n_1}})X_1, \qquad S_2 = X_2'(I - \frac{1}{n_2}\mathbf{1}_{\mathbf{n_2,n_2}})X_2. \end{aligned} \tag{3.1}$$

E finalmente, sejam $\frac{1}{m}S$ e $\underline{d}$ estimadores não viciados para a covariância comum e a diferença entre as médias das classes, respectivamente, com

$$S = S_1 + S_2, \qquad m = n_1 + n_2 - 2, \qquad \underline{d} = \overline{\underline{x}}^{(1)} - \overline{\underline{x}}^{(2)}.$$

Por extenção do Lema 2.4.1 (pag. 40), podemos concluir que:

$\overline{\underline{x}}^{(1)} \sim N_p(\underline{\mu}_1, n_1^{-1}\Sigma)$,

$\overline{\underline{x}}^{(2)} \sim N_p(\underline{\mu}_2, n_2^{-1}\Sigma)$,

$S_1 \sim W_p(\Sigma, n_1 - 1)$ e

$S_2 \sim W_p(\Sigma, n_2 - 1)$.

Como combinções lineares de vetores com distribuições Normais tem também distribuição Normal (ver Kshirsagar [1972]), segue que $\underline{d} = \overline{\underline{x}}^{(1)} - \overline{\underline{x}}^{(2)}$, tem distribuição Normal e definindo $\underline{u}$ por:

$$\underline{u} = C\Sigma^{-\frac{1}{2}}(\underline{d} - \underline{\delta_0})$$

onde $C^2 = \frac{n_1 n_2}{n_1+n_2}$, segue que $\underline{u}$ também tem distribuição Normal. Considerando H_0 verdadeira, isto é, $E\{\underline{d}\} = \underline{\mu}_1 - \underline{\mu}_2 = \underline{\delta} = \underline{\delta}_0$, temos que $E\{\underline{u}\} = \underline{0}$. Além disso a matriz de covariância de $\underline{d}$ é $(\frac{1}{n_1} + \frac{1}{n_2})\Sigma$ e a matriz de covariância de $\underline{u}$ é a matriz identidade I. Assim,

$$\underline{u} \sim N_p(0, I) \qquad \text{e} \qquad \underline{d} \sim N_p(\underline{\delta}_0, (\frac{1}{n_1} + \frac{1}{n_2})\Sigma). \tag{3.2}$$

Já a distribuição de $S = S_1 + S_2$ pode ser obtida através do seguinte lema :

Lema 3.2.1 - Se M_1 tem distribuição $W_p(\Sigma, m_1)$, M_2 tem distribuição $W_p(\Sigma, m_2)$ e M_1 e M_2 são independentes, então $M_1 + M_2$ tem distribuição $W_p(\Sigma, m_1 + m_2)$.

Prova: (ver Mardia et all [1979])•

Assim, $S = S_1 + S_2$ tem distribuição $W_p(\Sigma, n_1 + n_2 - 2)$, e usando novamente o Lema 2.4.1 segue que $\underline{u}$ e S são independentes. Escrevendo agora $M = \Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}}$, pelo Lema 2.4.2 temos:

$$M \sim W_p(I, m).$$

Definindo λ por:

$$\lambda = m\underline{u}' M^{-1} \underline{u} = C^2 m(\underline{d} - \underline{\delta}_0)' S^{-1} (\underline{d} - \underline{\delta}_0), \tag{3.3}$$

e considerando H_0 como válida, segue pela Definição 2.4.2 que λ tem distribuição T^2 de Hotelling com parâmetros p e m.

A importância dessa nova definição de λ é que através dela podemos associá-la com a distância amostral D^2, vista no Capítulo 2, pois quando H_0 é verdadeira, ($\underline{\delta} = \underline{\delta}_0$), λ tem distribuição $T^2(p, m)$ e a principal consequência segue pelo Lema 2.4.4, que dá a relação entre as distribuições T^2 de Hotelling e a F de Snedecor, isto é,

$$\frac{m-p+1}{pm}\lambda \sim F_{p,m-p+1}.$$

Pode ser mostrado que a distribuição de λ sem a suposição de H_0 verdadeira (ver Kshirsagar [1972], pag.146) tem distribuição proporcional a T^2 de Hotelling com m graus de liberdade e parâmetro de proporcionalidade

$$\eta = C^2(\underline{\delta} - \underline{\delta}_0)'\Sigma^{-1}(\underline{\delta} - \underline{\delta}_0), \tag{3.4}$$

onde η também é chamado de parâmetro de não centralidade. Essa é uma informação adicional que poderá ser utilizada em futuras pesquisas. A princípio, o interesse deste trabalho está na suposição que H_0 é verdadeira para que possamos utilizá-la de forma equivalente, como é visto na próxima seção.

Considerando agora $\underline{\delta}_0 = 0$, λ se reduz a

$$\lambda = C^2 m\underline{d}' S^{-1}\underline{d} = C^2 D^2 \tag{3.5}$$

que corresponde a distância de Mahalanobis amostral, vista na Subseção 2.2.5, substituída pelos seus estimadores não viciados. Sendo, portanto, proporcional a distância D^2(ver

equação (2.23)). Isto é, a expressão de λ é a mesma que a expressão de D^2, exceto pela constante C^2. Assim a distribuição exata de D^2, pode ser obtida através da distribuição T^2 de Hotelling. Essa distribuição foi inicialmente estudada por Bose e Roy [1938], porém sem suposição sob H_0, sendo conhecida como a distribuição não nula de D^2. Eles mostraram (ver Kshirsagar[1972], pag.198) que o valor esperado de D^2 é:

$$E(D^2) = \frac{m}{m-p-1}(\Delta_p^2 + \frac{p}{C^2}), \tag{3.6}$$

onde

$$\Delta_p^2 = \underline{\delta}'\Sigma^{-1}\underline{\delta},$$

foi definida na Subseção 2.2.4 como a distância entre duas classes com distribuições Normais. D^2 portanto, é um estimador viciado para Δ_p^2. Um estimador não viciado seria:

$$\frac{m-p-1}{m}D_p^2 - \frac{p}{C^2}, \tag{3.7}$$

onde o subscrito p em D_p^2 indica que a distância é calculada com base em p variáveis de cada classe.

3.2.3 Escolha de variáveis discriminantes

O objetivo desta subseção é definir uma forma de avaliar a contribuição de cada variável na construção da distância apresentada na subseção anterior. Sem perda de generalidade, queremos avaliar a contribuição das $(p-k)$ variáveis $(x_{k+1}, x_{k+2}, \ldots, x_p)$ na distância D^2, feita

inicialmente com base nas primeiras k variáveis (x_1,x_2,...,x_k). Considere as quantidades $\underline{x}$,$^{(1)}$ $\underline{x}^{(2)}$, S_1 e S_2, definidas em (3.1) e as partições de $\underline{\delta}$, Σ, $\underline{d}$ e S definidas abaixo:

$$\underline{\delta} = \begin{bmatrix} \underline{\delta}_1 \\ \underline{\delta}_2 \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{21} \\ \cdot \\ \cdot \\ \cdot \\ \mu_{1k} - \mu_{2k} \\ \mu_{1(k+1)} - \mu_{2(k+1)} \\ \cdot \\ \cdot \\ \cdot \\ \mu_{1p} - \mu_{2p} \end{bmatrix}, \qquad \underline{d} = \begin{bmatrix} \underline{d}_1 \\ \underline{d}_2 \end{bmatrix} = \begin{bmatrix} \overline{x}_1^{(1)} - \overline{x}_1^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \overline{x}_k^{(1)} - \overline{x}_k^{(2)} \\ \overline{x}_{k+1}^{(1)} - \overline{x}_{k+1}^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \overline{x}_p^{(1)} - \overline{x}_p^{(2)} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad \text{e} \quad S = S_1 + S_2 = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \tag{3.8}$$

onde Σ é a matriz de covariância comum as duas classes, Σ_{11} é a matriz de covariância particionadas em relação ao conjunto das primeiras k variáveis, Σ_{12} é a matriz de covariância particionada do conjunto das primeiras k variáveis em relação ao cojuntos das $(p-k)$ variáveis restantes, Σ_{21} é a matriz de covariância particionada do conjunto das últimas $(p-k)$ variáveis em relação ao conjuntos das primeiras k variáveis, Σ_{22} é a matriz de covariância particionadas em relação ao conjunto das restantes $(p-k)$ variáveis e S e S_{ij} representam as quantidades amostrais em relação as matrizes de covariâncias Σ e Σ_{ij}. Sejam também,

$$\beta = \Sigma_{21}\Sigma_{11}^{-1}, \qquad B = S_{21}S_{11}^{-1},$$

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \qquad e \qquad S_{22.1} = S_{22} - S_{21}S_{11}^{-1}S_{12}. \tag{3.9}$$

Usando os resultados do Apêndice A-2, para particionar Σ^{-1} e um argumento similar para particionar S^{-1}, podemos escrever (ver Kshirsagar [1972], pg.199),

$$\Delta_p^2 = \Delta_k^2 + (\underline{\delta}_2 - \beta\underline{\delta}_1)' \, \Sigma_{22.1}^{-1} \, (\underline{\delta}_2 - \beta\underline{\delta}_1)$$

e

$$D_p^2 = D_k^2 + m \, (\underline{d}_2 - B\underline{d}_1)' \, S_{22.1}^{-1} \, (\underline{d}_2 - B\underline{d}_1) \, , \tag{3.10}$$

onde

$$\Delta_k^2 = \underline{\delta}_1' \Sigma_{11}^{-1} \underline{\delta}_1 \qquad e \qquad D_k^2 = m\underline{d}_1' \Sigma_{11}^{-1} \underline{d}_1, \tag{3.11}$$

são respectivamente, as distâncias verdadeira e a Studentizada entre as classes π_1 e π_2 baseadas nas k primeiras variáveis. O acréscimo na verdadeira distância devido a inclusão das variáveis x_{k+1},x_{k+2},...,x_p, é dado por:

$$\Delta_p^2 - \Delta_k^2 = (\underline{\delta}_2 - \beta\underline{\delta}_1)' \, \Sigma_{22.1}^{-1} \, (\underline{\delta}_2 - \beta\underline{\delta}_1) \, , \tag{3.12}$$

usando (3.7), o estimador não viciado de diferença $\Delta_p^2 - \Delta_k^2$ é:

$$\frac{1}{m}\left[(m-p-1)D_p^2 - (m-k-1)\, D_k^2\right] - \frac{1}{C^2}(p-k), \tag{3.13}$$

onde os estimadores não viciados de Δ_p^2 e Δ_k^2 são substituidos diretamente.

Uma forma de avaliar a contribuição da $(p-k)$ variáveis, pode ser feita através do teste de hipóteses:

$$H_0 : \Delta_p^2 = \Delta_k^2. \tag{3.14}$$

Se H_0 for aceita, significa que as $(p-k)$ variáveis não têm habilidade discriminante quando as k primeiras variáveis já tenham sido consideradas. Observe que H_0 definida em (3.14) é equivalente a supor que $\underline{\delta}_{2.1} = \underline{\delta}_2 - \beta\underline{\delta}_1 = 0$. O seguinte lema fornece a distribuição de $\left(D_p^2 - D_k^2\right)$, supondo $\underline{\delta}_{2.1} = 0$.

Lema 3.2.2 - Se D_p^2, D_k^2, m e C^2, definidos como acima e $\underline{\delta}_{2.1} = 0$, então

$$\frac{C^2\left(D_p^2 - D_k^2\right)}{m + C^2 D_k^2} \sim \frac{p-k}{m-p+1} F_{p-k,m-p+1}.$$

Prova:[ver Kshisagar[1972] pag.200)]•

A importância desse lema está em se poder avaliar a hipótese (3.14), e então decidir rejeitar H_0 se $F_c > F_t$, de acordo com um valor previamente fixado α, tal que:

$$\Pr(F_c > F_t) = \alpha, \tag{3.15}$$

onde

$$F_c = \frac{m-p+1}{p-k}\left[\frac{C^2\left(D_p^2 - D_k^2\right)}{m + C^2 D_k^2}\right],$$

e F_t, também denotado por $F_{p-k,m-p+1,\alpha}$, é um valor tabelado na Distribuição F de Snedecor com $(p-k)$ e $(m-p+1)$ graus de liberdade e α a probabilidade de rejeitar H_0, quando H_0 é verdadeira: O nível de significância do teste. Em outras palavras, α é a probabilidade de uma variável aleatória com distribuição F de Snedecor (F_c) exceder o limiar $F_{p-k,m-p+1,\alpha}$.

Observe que decidir por rejeitar H_0, significa decidir que a contribuição das $(p-k)$ variáveis é significativa na construção da distância Δ_p^2. Nesse sentido, nossa proposta é avaliar individualmente a contribuição de cada variável, num primeiro estágio, e em seguida avaliar se o conjunto das variáveis estatisticamente significante pode ser considerado como satisfatório. Isso pode ser feito realizando incialmente p testes sob H_0 (um teste para cada variável isolada), fazendo $k = p-1$. Nesse estágio seria definido um conjunto inicial de variáveis significativas (aquelas associadas à rejeição de H_0), onde digamos, k^*, variáveis significativas seriam escolhidas. Num segundo estágio seria realizado um único teste com $k = k^*$. Nesse segundo estágio, se H_0 for aceita, podemos confirmar que as variáveis excluídas no primeiro estágio, não contribuem de forma significativa na construção da distância Δ_p^2. Isto é, se H_0 for aceita, podemos concluir que $\Delta_p^2 = \Delta_{k*}^2$. E o conjunto de variáveis selecionadas seria composto dessas k^* variáveis. Caso contrário, se H_0 for rejeitada, poderíamos supor que, dentre as variáveis excluídas, existe pelo menos uma que teria uma contribuição significativa, sendo necessário incluir, uma a uma, no conjunto de variáveis selecionadas até que H_0 seja aceita. Um resumo desse método é apresentado a seguir:

i) Realize p testes usando (3.15) com $k = p-1$.

ii) Forme um novo conjunto com as k^* variáveis significativas do passo (i).

iii) Faça $k = k^*$ e realize novamente o teste usando (3.15).

iv) Se H_0 do passo (iii) for aceito o processo de seleção está terminado. Caso contrário, acrescente uma nova variável $(*)$ e retorne ao passo (iii) com $k = k^* + 1$.

(3.16)

$(*)$ A nova variável a ser incluída deve ser aquela que no passo (i) tiver F_c mais próximo de F_t.

Observe que esse método não inclui todas os conjuntos possíveis de combinações entre p e k. Mas esta proposta dá chances a uma variável que teve seu efeito isolado considerado não significativo, ser novamente avaliada em conjunto. Isso implica em estabelecer um valor para α (o nível de significância do teste) em (3.15) e depois aumentar esse valor para níveis aceitáveis. É comum adotar um valor inicial para α de no máximo $0,15$ e para o caso de rejeitar H_0, no passo (iv), estender esse valor para, no máximo, $0,20$.

3.3 Seleção de Variáveis Através de Modelos de Regressão Logística

3.3.1 Introdução

Um dos aspectos mais importantes da abordagem do problema de classificação através de modelos de regressão é a seleção de um conjunto de variáveis explicativas que melhor se ajuste ao modelo. A proposta é obter um modelo com um número mínimo de parâmetros que consiga descrever o comportamento da variável, dita resposta do modelo (variável dependente), em função de um conjunto de outras, ditas variáveis explicativas (variáveis independentes).

Em modelos de regressão linear, a seleção é feita através de técnicas bastante conhecidas. Uma das técnicas considera o ajuste de todos os submodelos possíveis, sendo a seleção feita com base em alguns coeficientes que medem o quanto a variável resposta está sendo

explicada pelo conjunto selecionado. Entretanto, quando o número de variáveis explicativas é muito grande esta técnica torna-se inviável. Outras técnicas fazem a escolha sistemática, onde apenas alguns modelos são verificados. Uma variável pode entrar ou sair do modelo dependendo de sua contribuição no ajuste, como por exemplo nas técnicas "Forward " e "Backward ", onde vai-se incluindo ou excluindo as variáveis dependendo de sua importância no ajuste do modelo (ver Draper e Smith[1981]). Um procedimento de escolha sistemática diminue consideravelmente o tempo computacional, especialmente no caso onde o número de variáveis explicativas é muito grande.

Em regressão logística, a seleção não pode ser feita, simplesmente, através de uma extensão ao caso de regressão linear (Hosme et al[1989]). Neste caso, a estimação dos parâmetros envolve soluções não lineares, necessitando de métodos iterativos. Vários métodos alternativos de estimação dos parâmetros têm sido propostos na literatura, sendo este o tema de várias pesquisas nos últimos anos (ver Aidar[1992]). Como este trabalho trata de um problema com um grande número de variáveis explicativas ($p > 30$), é inviável um estudo de todos os possíveis submodelos. Assim, destacamos nessa seção apenas os métodos sistemáticos, conhecidos como "Stepwise " e suas variações "Forward " e "Backward ".

A descrição dos procedimentos de seleção segue a teoria de teste de hipóteses de ajuste do modelo. Apresentaremos a seguir os critérios utilizados nessa escolha.

3.3.2 Teste para ajuste do Modelo de Regressão Logística

A suposição inicial para teste de hipóteses do ajuste do modelo é que as observações da variável resposta, y_i $(i = 1, 2, ..., n)$, sejam independentes e identicamente distribuídas de acordo com a distribuição de Bernoulli com parâmetro θ_i. A partir daí, um modelo contendo todas as p variáveis explicativas pode ser comparado com um modelo mais simples, por exemplo, com um modelo contendo apenas o termo constante β_0. O propósito é verificar se pelo menos uma das variáveis explicativas contribui de forma significativa no ajuste. Mas, o

interesse aqui é verificar a contribuição individual de cada variável no ajuste, de modo que se possa decidir, passo a passo, pela inclusão ou não da variável no modelo.

Uma estatística inicialmente indicada nesse caso é a estatística de Wald , onde o objetivo é testar a hipótese $H_0 : \beta_j = 0$, que é equivalente a testar se o efeito da j-ésima variável independente ($j = 1, 2, ..., p$) é significativo. Entretanto, Hauck e Dorner [1977] afirmam que esse teste sub-estima o nível de significância do teste (a probabilidade de rejeitar H_0, quando H_0 é verdadeira), não rejeitando H_0 para grandes valores de β_j. Eles sugerem que as conclusões finais, sobre o efeito de cada variável, devam ser feitas com base na razão de verossimilhança:

$$RV = -2 \log \left[\frac{L(\beta)}{L(\beta_{[j]})} \right] \tag{3.17}$$

onde $L(\beta)$ é a função de verossimilhança definida em (2.38) e o subescrito $[j]$ em $\beta_{[j]}$, indica que a verossimilhança não inclui a j-ésima variável. Na comparação entre os dois modelos, um com p e outro com $p - k$ variáveis, a razão de verossimilhanças (3.17) tem aproximadamente distribuição Qui-quadrado com $p - k$ graus de liberdade (ver Draper e Smith [1981]). Esse procedimento será repetido seguidamente no processo de seleção.

3.3.3 Procedimentos de seleção sistemática

O procedimento sistemático mais utilizado em MRL é o "Stepwise ". Basicamente esse procedimento consiste na avaliação da razão de verossimilhanças entre os modelos ajustados num determinado passo e o modelo do passo imediatamente anterior. Compara-se o valor dessa razão, cada vez que se inclui ou exclui uma variável do modelo. Os passos desse, e dos demais procedimentos, são apresentados a seguir:

No processo "Stepwise " o primeiro passo ajusta o modelo contendo apenas o termo constante, β_0. O segundo passo ajusta todos os p modelos ($\lambda = \beta_0 + \beta_{[i]} x_{[i]}$, $i = 1, 2, ..., p$), contendo uma variável. Fixa-se um nível de significância, α_e, para ser comparado com a variável que entra no modelo. O valor da razão de verossimilhanças entre cada modelo ajus-

tado e o modelo do passo anterior, é portanto, comparado com χ^2_{1,α_e}, onde $\chi^2_{1,\alpha}$ é um valor tabelado na distribuição Qui-quadrado com 1 grau de liberdade e nível de significância α. Se os valores das razões de verossimilhanças, associados acada uma das variáveis, que ainda não entraram no modelo resultarem em valores menores que χ^2_{1,α_e}, a seleção termina, indicando que nenhuma variável tem efeito significativo. Caso contrário, seleciona-se a variável associada com o maior valor entre os testes das razões de verossimilhanças. O terceiro passo, ajusta todos os $(p-1)$ modelos contendo duas variáveis (a variável do segundo passo e uma das outras $(p-1)$ que ainda não foram selecionadas). De modo similar, compara-se o valor do teste da razão de verossimilhança com χ^2_{1,α_e} e seleciona-se a variável associada à maior razão. Observe que a razão de verossimilhanças nesse passo é feita entre o modelo com duas varáveis e o modelo do passo anterior, constituído de apenas uma variável. Nesse momento verifica-se a permanência da primeira variável selecionada, através da razão de verossimilhança entre os modelos, com e sem a primeira variável. Em outras palavras, chamando, respectivamente, de $x_{[1]}$ e $x_{[2]}$ a primeira e a segunda variáveis selecionadas, compara-se os modelos $\lambda_{[12]} = \beta_0 + \beta_{[1]}x_{[1]} + \beta_{[2]}x_{[2]}$ e $\lambda_{[2]} = \beta_0 + \beta_{[2]}x_{[2]}$. Define-se um nível de significância, α_s, para ser comparado com a variável que sai do modelo. Se a razão de verossimilhança for inferior a χ^2_{1,α_s}, elimina-se $x_{[1]}$, indicando que o efeito de $x_{[1]}$ está contido em $x_{[2]}$. Os dois últimos passos são repetidos até que não exista razão de verossimilhanças maior que χ^2_{1,α_e}. A escolha dos valores de α_e e α_s é feita de modo que não exista dúvida em eliminar uma variável importante. É convencional definir $\alpha_e = 0,15$ e $\alpha_s = 0,20$.

No procedimento "Forward " o primeiro passo é idêntico ao procedimento "Stepwise ". Inicialmente ajusta-se o modelo apenas com o termo constante β_0, e nos passos seguintes vai-se incluindo as variáveis, uma a uma, de acordo com as razões de verossimilhanças. A diferença é que o procedimento "Forward " não propõe eliminar as variáveis que já foram incluídas nos passos anteriores. O processo termina quando todas as razões de verossimilhanças do passo em questão são inferiores a χ^2_{1,α_e}.

No processo "Backward " ocorre o contrário do processo "Forward ". O processo inicia com o ajuste do modelo contendo todas as variáveis, e vai-se eliminado, a cada passo, aquela variável associada com a menor razão de verossimilhanças. O processo termina quando todas as razões, do passo em questão, forem superiores a χ^2_{1,α_s}.

Uma seleção de variáveis mais cuidadosa poderia ser feita, se o pesquisador estivesse interessado nos efeitos combinados das variáveis explicativas, conhecido como modelos com interação. Assim como se o número de variáveis explicativas não for muito grande, de forma que seja viável uma comparação entre todos as possíveis combinações de p e k, sendo, neste caso, esperado uma seleção mais eficiente. Entretanto, a proposta desse trabalho é avaliar o desempenho da predição do MRL na sua forma linear, não pretendendo dar como concluídas as discussões sobre a melhor formar de selecionar variáveis explicativas. Existe uma vasta literatura dando ênfase especificamente a esse assunto (ver Aidar[1992]). No final deste capítulo, apresentamos um estudo simulado para avaliar alguns métodos alternativos de seleção de variáveis. A proposta deste trabalho é utilizar o conjunto de dados do capítulo anterior para medir o desempenho da seleção sistemática na comparação com outros métodos aqui apresentados.

3.4 Seleção de Variáveis Usando Observações Apenas de Uma das Classes

3.4.1 Introdução

Mais uma vez utilizamos as propriedades das distribuições, T^2 de Hotelling e Wishart, para avaliar as contribuições das variáveis explicativas na construção de uma medida de distância entre duas classes. Mas, agora, existe uma restrição ao uso das observações de uma das classes. A proposta, portanto, é estudar o desempenho de um conjunto com um número menor de variáveis, k, em relação a medida de distância de um objeto ao centro da

distribuição, definida na Seção 2.4, quando usamos todas as p ($p \geq k$) variáveis explicativas.

Nesta seção abordaremos o problema da seleção de variáveis de modo similar ao apresentado na Seção 3.2, onde a distância de Mahalanobis é usada na formulação da distribuição proporcional da medida de distância entre os elementos de uma classe com distribuição Normal e um ponto de referência. Inicialmente estudaremos a distribuição dessa medida de distância, e em seguida apresentaremos um método de seleção com base nessa distribuição.

3.4.2 Distribuição da distância D^2 restrita a uma das classes

A distância Studentizada D^2, estudada na Seção 3.2, leva em consideração o acesso irrestrito às observações de ambas as classes. Além da utilização das propriedades das distribuições T^2 de Hotelling e Wishart, inclui também a hipótese de que a diferença entre as médias é igual a uma certa constante, $\underline{\delta}_0$, na avaliação da contribuição de cada variável na construção da medida de distância. A proposta aqui é também avaliar a contribuição das variáveis, mas em relação a construção da medida de distância entre os elementos de uma das classes a um ponto de referência. Isto é, assumindo que a média da classe que temos acesso às observações é igual a uma certa constante, digamos $\underline{\mu}_0$, queremos avaliar a contribuição de cada variável na construção da medida de distância entre os elementos dessa classe e o valor constante $\underline{\mu}_0$.

Quando uma amostra aleatória de tamanho n de uma classe com distribuição $N_p(\underline{\mu}, \Sigma)$ é observada, e supondo Σ conhecido, uma forma de testarmos a hipótese:

$$H_0 : \underline{\mu} = \underline{\mu}_0 \tag{3.18}$$

pode ser feita através da estatística:

$$n\left(\overline{\underline{x}} - \underline{\mu}_0\right)' \Sigma^{-1} \left(\overline{\underline{x}} - \underline{\mu}_0\right) \tag{3.19}$$

que tem distribuição Qui-quadrado com p graus de liberdade (ver Kshirsagar[1972], pag.141). Entretanto quando não assumimos o conhecimento de Σ, podemos utilizar o Lema 2.4.1 e as propriedades da distribuição Normal para definir uma nova versão da distribuição T^2 de Hotelling, e isto é feito definindo inicialmente:

$$\underline{u} = \sqrt{n}\left(\underline{\overline{x}} - \underline{\mu_0}\right)$$

com

$$\underline{\overline{x}} = \frac{1}{n}X'\mathbf{1_{1,n}} \qquad e \qquad S = X'\left(I - \frac{1}{n}\mathbf{1_{n,n}}\right)X, \tag{3.20}$$

de modo que $\underline{u}$ tem distribuição $N_p(\sqrt{n}\left(\underline{\mu} - \underline{\mu_0}\right), \Sigma)$ e S tem distribuição $W_p(\Sigma, n-1)$. Se a hipótese H_0, como em (3.18), é verdadeira segue que $\underline{u}$ tem distribuição $N_p(\underline{0}, \Sigma)$, e usando o Lema 2.4.3, temos que

$$D_*^2 = (n-1)\underline{u}'S^{-1}\underline{u} = (n-1)n\left(\underline{\overline{x}} - \underline{\mu_0}\right)'S^{-1}\left(\underline{\overline{x}} - \underline{\mu_0}\right) \sim T^2(p, n-1), \tag{3.21}$$

onde o subescrito $*$ em D_*^2 é usado para diferenciar a distância Studentizada D^2, definida em (2.27). Pelo Lema 2.4.4 segue que

$$\frac{(n-1)-p+1}{np}D_*^2 = n\frac{n-p}{p}\left(\underline{\overline{x}} - \underline{\mu_0}\right)'S^{-1}\left(\underline{\overline{x}} - \underline{\mu_0}\right) \sim F_{p,n-p}. \tag{3.22}$$

Em resumo, D_*^2 como definido em (3.21), representa uma medida de distância entre os vetores pertencentes à uma classe com distribuição Normal ao centro da distribuição, baseado em p

variáveis explicativas. Bose e Roy [1938] também estudaram a distribuição dessa distância sem a suposição H_0. Nesse caso, D_*^2 teria uma distribuição proporcional a T^2 de Hotelling com $(n-1)$ graus de liberdade e parâmetro de não centralidade

$$\eta = n(\underline{\mu} - \underline{\mu}_0)'\Sigma^{-1}(\underline{\mu} - \underline{\mu}_0). \tag{3.23}$$

Mais detalhes sobre essa distribuição podem ser encontrados em Kshirsagar[1972]. A próxima subseção faz uma abordagem alternativa de como estudar a contribuição de cada variável na definição dessa distância.

3.4.3 Seleção de variáveis discriminantes

A distância de Mahalanobis de uma vetor aleatório com distribuição Normal multivariada até a origem é definida como:

$$\Delta_p^2 = \underline{\mu}'\Sigma^{-1}\underline{\mu}, \tag{3.24}$$

onde o subescrito p indica que a distância é calculada com base em p variáveis. O objetivo agora é avaliar essa distância com base em dois conjuntos de variáveis: um com k e outro com $(p-k)$. A proposta é fixar um dos conjuntos e usar a teoria de Distribuição Condicional (ver Apêndice A-5) no parcelamento dessa distância.

Particionando $\underline{\mu}' = \left(\underline{\mu}_1', \underline{\mu}_2'\right)$, onde sem perda de generalidade diremos que $\underline{\mu}_1$ está associado com as k primeiras variáveis, enquanto que $\underline{\mu}_2$ está associado com as $(p-k)$ variáveis restantes e usando os resultados do Apêndice A-2 para particionar Σ^{-1}, podemos escrever (3.24) (ver Mardia et al [1979]) como:

$$\begin{aligned}\Delta^2_{p*} &= \underline{\mu}_1'\Sigma_{11}^{-1}\underline{\mu}_1 + \underline{\mu}_{2.1}'\Sigma_{22.1}^{-1}\underline{\mu}_{2.1} \\ &= \Delta^2_{k*} + \underline{\mu}_{2.1}'\Sigma_{22.1}^{-1}\underline{\mu}_{2.1}, \end{aligned} \tag{3.25}$$

onde Δ^2_{k*} representa a distância de um vetor aleatório com distribuição Normal multivariada até a origem, com base nas k primeiras variáveis, o subescrito $*$ é usado para diferenciar a verdadeira distância entre duas classes, definida em (3.11), isto é, $\underline{\mu}_{2.1} = \underline{\mu}_2 - \Sigma_{21}\Sigma_{11}^{-1}\underline{\mu}_1$ e $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Uma forma de medir a contribuição das $(p-k)$ variáveis na construção da distância Δ^2_{k*}, quando as k primeiras variáveis estão fixadas, é avaliar a suposição que $\underline{\mu}_{2.1} = \underline{0}$. Em outras palavras, considerar $\underline{\mu}_{2.1} = \underline{0}$ é equivalente supor $\Delta^2_{p*} = \Delta^2_{k*}$, ou ainda, que a distância de Mahalanobis até a origem baseada em p variáveis é mesma que quando baseada em k variáveis.

Na prática, uma forma de avaliar a hipótese $\underline{\mu}_{2.1} = \underline{0}$ ou equivalentemente $\Delta^2_{p*} = \Delta^2_{k*}$, é feita através da distância D^2_*, definida em (3.21), considerando $\underline{\mu}_0 = \underline{0}$. Usando os resultados do Apêndice A-2, de modo similar, para particionar S^{-1}, podemos escrever (3.21) (ver Mardia et al [1979]) como:

$$\begin{aligned} D^2_{p*} &= (n-1)\underline{u}_1'S_{11}^{-1}\underline{u}_1 + (n-1)\underline{z}'S_{22.1}^{-1}\underline{z}, \\ &= D^2_{k*} + (n-1)\underline{z}'S_{22.1}^{-1}\underline{z}, \end{aligned} \tag{3.26}$$

onde D^2_{k*} é a distância amostral baseada nas k primeiras variáveis, $\underline{z} = \underline{u}_2 - S_{21}S_{11}^{-1}\underline{u}_1$, $S_{22.1} = S_{22} - S_{21}S_{11}^{-1}S_{12}$ e $\underline{u}' = (\underline{u}_1', \underline{u}_2')$, com $\underline{u}_i = \sqrt{n}\underline{\overline{x}}_i$, $i = 1, 2$. O lema a seguir fornece a distribuição proporcional de $\left(D^2_{p*} - D^2_{k*}\right)$ quando $\underline{\mu}_{2.1} = \underline{0}$.

Lema 3.4.1 - Se D^2_{p*} e D^2_{k*} são definidos na expressão (3.26) e $\underline{\mu}_{2.1} = \underline{0}$, então

$$\frac{\left(D^2_{p*} - D^2_{k*}\right)}{(n-1) + D^2_{k*}} \sim \frac{p-k}{n-p} F_{p-k,n-p},$$

e é independente de D^2_{k*}.

Prova: (ver Mardia et al [1979], pag.78)•

A importância do Lema 3.4.1 é que através dele podemos testar a hipótese:

$$H_0 : \underline{\mu}_{2.1} = \underline{0},$$

que é equivalente testar:

$$H_0 : \Delta^2_{p*} = \Delta^2_{k*}. \tag{3.27}$$

Em outras palavras, podemos testar se a contribuição das $(p-k)$ variáveis é significativa na construção da distância Δ^2_{p*}, quando fixadas as primeiras k variáveis. E decidir por rejeitar H_0, (3.27), se $F_c > F_t$, onde F_t, também denotado por $F_{p-k,n-p,\alpha}$, é escolhido de acordo com um valor previamente fixado de α, tal que:

$$\Pr(F_c > F_t) = \alpha,$$

onde

$$F_c = \left(\frac{n-p}{p-k}\right) \frac{\left(D^2_{p*} - D^2_{k*}\right)}{(n-1) + D^2_{k*}}, \tag{3.28}$$

sendo F_t um valor tabelado da distribuição F de Snedecor com $(p-k)$ e $(n-p)$ graus de liberdade no numerador e no denominador, respectivamente, e α a probabilidade de rejeitar H_0 quando esta hipótese é verdadeira: o nível de significância do teste. Isto é, α representa

a probabilidade de uma variável aleatória com distribuição F de Snedecor (F_c) exceder o limiar F_t.

A seguir apresentaremos um resumo do processo de seleção de modo equivalente aos passos de (3.16), quando o acesso às observações de ambas as classes foi suposto irrestrito. Mais uma vez, a nossa proposta é avaliar inicialmente a contribuição individual de cada variável, e em seguida verificar, de forma global, se o conjunto excluído não é, de fato, significante na construção da medida de distância.

i) Realize p testes usando (3.28), com $k = p - 1$.

ii) Forme um novo conjunto de variáveis com aquelas que forem significativas no passo (i) e chame o número de variáveis desse conjunto de k^*.

iii) Faça $k = k^*$ e realize novamente o teste usando (3.28).

iv) Se H_0 do passo (iii) for aceito o processo de seleção termina. Caso contrário acrescente uma nova variável (*) e retorne ao passo (iii) com $k = k^* + 1$.

(3.29)

(*) A nova variável a ser escolhida deve ser aquela que no passo (i) teve o valor de F_c mais próximo a F_t.

A vantagem desse método está em exigir pouco tempo computacional, além de possibilitar a inclusão de variáveis, que no passo inicial tenham sido excluídas. Como em (3.16) os valores de α recomendados são: 0,15 e 0,20, respectivamente, nos passos (i) e (iv). A seguir apresentaremos outros métodos de seleção de variáveis, baseados apenas nas observações de uma das classes, que servirão de alternativa a este.

3.5 Outros Métodos de Seleção de Variáveis

3.5.1 Introdução

Nesta seção apresentaremos três métodos de seleção de variáveis que consideram apenas as observações de uma das classes. Como nas últimas seções, o objetivo da seleção também é possibilitar uma forma eficiente de classificação. Entretanto, existem outros métodos na literatura, que podem servir de alternativa ao método apresentado na seção anterior.

Os critérios adotados em cada método apresentado a seguir, estão baseados em técnicas estatísticas, onde o objetivo é eliminar algumas variáveis, de modo que se mantenha a identidade original das variáveis que permanecem. Inicialmente introduzimos um método que avalia a variabilidade relativa de cada variável através do Coeficiente de Variação (CV). O critério se baseia no fato de que uma variável com grande dispersão relativa, interfere de modo significativo na definição da região de separação das classes. A proposta é que eliminando essas variáveis se reduzam as chances de classificar erroneamente uma observação que não pertença à classe em questão. Os dois outros métodos têm critérios semelhantes, onde o objetivo é eliminar as variáveis que possam ser explicadas por um conjunto de outras. Um deles utiliza a teoria de Análise de Componentes Principais (CP), de forma que o critério estabelecido é eliminar as variáveis associadas às componentes menos representativas. E, finalmente, o terceiro método avalia o Coeficiente de Correlação Múltipla (CR), entre uma variável e um conjunto de outras, considerados candidatas a permanecerem no conjunto após o processo de seleção. O objetivo é eliminar as variáveis mais correlacionadas com esse conjunto, consideradas independentes. A seguir apresentamos de forma resumida cada um dos métodos. Maiores detalhes sobre eles podem ser obtidos nas referências citadas durante a apresentação. Apenas uma modificação ao último método CR, foi sugerida neste trabalho, para viabilizar a operacionalidade do método, devido ao grande número de variáveis explicativas estudadas aqui.

3.5.2 Seleção de variáveis através do Coeficiente de Variação (CV)

A grande vantagem deste método está na sua forma clara de definir o critério de seleção: Elimina-se as variáveis com grande dispersão relativa. O argumento básico para se adotar esse critério se deve ao fato de que quanto maior a dispersão da variável, maiores as chances de interseção de seus pontos nas duas classes. Em outras palavras, em reconhecimento de padrões espera-se que as regiões onde estão os pontos de uma variável sejam distintos em relação as classes. Como na prática essa distinção dificilmente ocorre, a solução seria optar pela variáveis mais concentradas, em termos relativos, ao redor de sua média.

O Coeficiente de Variação (CV) é definido como a razão entre o desvio padrão e a média da variável em questão. Representa uma medida de variabilidade relativa de quão concentrada está a variável em relação a sua média. Em geral, é expresso em percentual e sua principal aplicação é servir de referência numa comparação entre variáveis de unidades diferentes. Denotando como μ_j e σ_j^2 a média e a variância da j-ésima variável ($j = 1, 2, ..., p$), respectivamente, o vetor de coeficientes de variação é definido como:

$$\underline{CV} = \left[\frac{\sigma_1}{\mu_1}, \frac{\sigma_2}{\mu_2}, ..., \frac{\sigma_p}{\mu_p}\right]. \tag{3.30}$$

O CV amostral é obtido substituindo os parâmetros pelos estimadores não viciados de cada variável, respectivamente, isto é,

$$\widehat{\underline{CV}} = \left[\frac{s_1}{\overline{\mathrm{x}}_1}, \frac{s_2}{\overline{\mathrm{x}}_2}, ..., \frac{s_p}{\overline{\mathrm{x}}_p}\right], \tag{3.31}$$

onde

$$\overline{\mathrm{x}}_j = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{ij} \qquad \text{e} \qquad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\mathrm{x}_{ij} - \overline{\mathrm{x}}_j)^2 \qquad j = 1, 2, ..., p. \tag{3.32}$$

O método consiste em ordenar os elementos de $\widehat{\underline{CV}}$ em (3.31) em ordem crescente, e selecionar as k primeiras variáveis que estão associadas a esses valores, reduzindo assim a dimensão da matriz de dados de (n x p) para (n x k), de acordo com

$$\widehat{\underline{CV}} = \left[cv_{[1]}, cv_{[2]}, ..., cv_{[k]}\right],$$

onde $cv_{[1]}$ é o menor valor de $cv_{[j]}$, $cv_{[2]}$ é o segundo menor, e assim por diante.

O problema desse método é estabelecer o valor de k. Não existe uma regra clara que possa estabelecer esse valor de forma definitiva. Em geral, um CV acima de 15% indica grande variabilidade relativa, significando alta dispersão absoluta. Assim, é razoável supor que uma variável com valor do CV nessa ordem, teria maior chance de se observar pontos em comum nas duas classes. Um critério mais rigoroso eliminaria a variável cujo CV amostral fosse superior a 10%.

Este critério será verificado através de dados simulados na última seção desse capítulo. Um critério adotado por Nelson et al [1994], para resolver o problema da definição do valor de k, foi feito para um estudo específico de verificação de assinaturas, onde alguns valores de k foram pré-estabelecidos arbitrariamente e depois verificados seus comportamentos.

3.5.3 Seleção de variáveis através de Análise de Componentes Principais (CP)

O primeiro objetivo da Análise de Componentes Principais (CP) é procurar uma combinação linear do vetor aleatório $\underline{x}$= (x_1,x_2, ...,x_p) que tenha máxima variância, onde as possíveis combinações lineares são padronizadas, de modo que a soma de seus coeficientes

seja igual a 1, para possibilitar uma comparação entre as combinações. De forma geral, o objetivo é estabelecer uma transformação ortogonal na matriz de dados X, em um conjunto de novas variáveis não correlacionadas. Essa tranformação é definida a seguir:

Definição 3.5.1 - Se $\underline{x}$ é um vetor aletório com média $\underline{\mu}$ e matriz de covariância Σ, então a transformação em componentes principais de $\underline{x}$ em $\underline{y}$ é definida pela transformação:

$$\underline{y} = \Gamma' \left(\underline{x} - \underline{\mu} \right),$$

onde Γ é uma matriz ortogonal, tal que $\Gamma'\Sigma\Gamma = \Lambda$ é uma matriz diagonal, cujos os elementos da diagonal principal são os auto-valores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ de Σ, onde Σ é uma matriz positiva definida.

A j-ésima componente do vetor $\underline{x}$ é definido como o j-ésimo elemento de $\underline{y}$,

$$\mathrm{y}_j = \underline{\gamma}_j' \left(\underline{x} - \underline{\mu} \right), \tag{3.33}$$

onde $\underline{\gamma}_j$ é a j-ésima coluna de Γ, sendo chamado de j-ésimo componente principal do vetor aleatório $\underline{x}$. Algumas propriedades decorrente da Definição 3.5.1 são apresentadas nos dois lemas a seguir:

Lema 3.5.1 - Se o vetor aleatório $\underline{x}$ é definido como na Definição 3.5.1, então:

a)$E(\mathrm{y}_j) = 0$.

b)$Var(\mathrm{y}_j) = \lambda_j$.

c)$Cov(\mathrm{y}_i,\mathrm{y}_j) = 0$, se $i \neq j$.

d)$Var(\mathrm{y}_1) \geq Var(\mathrm{y}_2) \geq ... \geq Var(\mathrm{y}_p) \geq 0.$

e)$\sum_{j=1}^{p} Var(\mathrm{y}_j) = tr\Sigma.$

f)$\prod_{j=1}^{p} Var(\mathrm{y}_j) = |\Sigma|.$

onde $tr\Sigma$ é o traço da matriz Σ definido como a soma dos elementos da diagonal principal e $|\Sigma|$ o seu determinante.

Prova: (ver Mardia et al[1979], pag.215)•

Lema 3.5.2 - Nenhuma combinação linear padronizada de $\underline{x}$ tem variância maior que λ_1.

Prova:(ver Mardia et al[1979], pag.215)•

Um item que merece destaque no Lema 3.5.1 é o (d) que estabelece que a variância de y_1 é maior ou igual a variância de qualquer outra componente. E a interpretação geométrica do Lema 3.5.2 é que a primeira componente tem o maior eixo no elipsóide de concentração dos pontos de $\underline{x}$, quando $\underline{x}$ é um vetor aleatório com distribuição normal multivariada. Isto é, os pontos de $\underline{y}$ estão mais dispersos em relação a y_1 de que a qualquer outra componente.

Na prática, pode ser feita uma analogia dos vetores aleatórios com os estimadores amostrais da média e da matriz de covariância (ver Mardia et al [1979]), através de

$$Y = (X - \mathbf{1}_{\mathbf{n},\mathbf{1}}\overline{x}')\,G, \tag{3.34}$$

com

$$\overline{x} = \frac{1}{n} X' \mathbf{1_{n,1}}.$$

onde G é uma matriz ortogonal, tal que $G'SG = L$, $S = X'(I - \frac{1}{n}\mathbf{1_{n,n}})X$ e L é uma matriz diagonal, cujos elementos da diagonal são $l_1 \geq l_2 \geq ... \geq l_p$. Isto é, as colunas de Y são não correlacionadas e a variância de $\underline{y}_j = l_j$. Desse modo podemos obter a matriz transformada Y por (3.34). E $\underline{y}_j$ pode ser obtido de modo similar, isto é:

$$\underline{y}_j = (X - \mathbf{1_{n,1}}\overline{x}')\underline{g}_j, \qquad j = 1, 2, ..., p,$$

onde g_j representa a j-ésima coluna da matriz G, denotando o j-ésimo auto-vetor padronizado associado ao j-ésimo auto-valor l_j de S

Outros resultados importantes podem ainda ser obtidos a partir dos Lemas 3.51 e 3.52 e da Definição 3.5.1. Mas o principal interesse está na aplicação da medida de variação total, $tr\Sigma$, e suas proporções. Através do item (e) do Lema 3.5.1, podemos construir uma medida quantitativa que associa a quantidade de informação perdida em se reduzir a dimensão do espaço de p para k ($k \leq p$). Mais precisamente: a soma dos k primeiros autos-valores dividido pela soma de todos os autos-valores, $(\lambda_1 + \lambda_2 + ... + \lambda_k)$ / $(\lambda_1 + \lambda_2 + ... + \lambda_p)$, representa a proporção da variação total explicada pelas k primeiras componentes.

O método de seleção apresentado a seguir, foi introduzido por Beale et al [1967]. O critério de seleção avalia inicialmente a proporção explicada por um número reduzido de componentes, digamos k ($k \leq p$), obtido pela contagem dos $\lambda_i \geq 0,7$. E em seguida, são eliminadas as variáveis associadas com as componentes excluídas. Isto é, são eliminadas as variáveis com maiores coeficientes (em valores absolutos) dos auto-vetores associados com as

últimas componentes principais. A justificativa é que são eliminadas as variáveis dominantes nas componentes menos importantes. O resumo do método segue:

i) Encontre as matrizes G e L, através da equação (3.34).

ii) Defina k como sendo o número de $\lambda_i \geq 0,7$ $(*)$.

iii) Considere o auto-vetor de G associado com o menor auto-valor de L. Elimine a variável associada com maior coeficiente deste auto-vetor (em valor absoluto). Em seguida, considere o auto-vetor associado com segundo menor auto-valor. Elimine a variável associada com maior coeficiente deste auto-vetor (também em valor absoluto, desde que a variável não tenha ainda sido eliminada). Execute esse passo até que $(p-k)$ auto-vetores tenham sido considerados.

(3.35)

$(*)$ O valor sugerido para definir o valor de k foi recomendado por Jolliffe (1972) depois de vários exemplos numéricos.

3.5.4 Seleção de variáveis através do Coeficiente de Correlação Múltipla (CR)

Quando alguma(s) coluna(s) da matriz de dados, X, apresentam alguma relação funcional, uma(s) com a(s) outra(s), ela(s) pode(m) ser eliminada(s) sem prejudicar a descrição dos dados. Isto é, quando alguma(s) coluna(s) da matriz de dados X, for quase ou totalmente dependente de outra(s), um procedimento de exclusão dessa(s) variável(is) pode ser adotado sem prejuízo da representatividade dos dados. Um procedimento similar adotado em modelos de regressão, utiliza o coeficiente de determinação, $R^2_{y,X}$, que resume uma medida de quanto um sub-conjunto de X , digamos, k variáveis (variáveis independentes), explicam a variação total da variável resposta do modelo, y (variável dependente) (ver Draper e Smith

[1981]). Nesse sentido, necessitamos de uma medida quantitativa que possa servir de base na definição do método de seleção das variáveis que melhor represente os dados X, numa dimensão reduzida.

O coeficiente de correlação múltipla é uma medida de associação entre uma variável e um conjunto de outras, indicando o quanto a variável está sendo explicada pelo conjunto. Morrison [1976], define o coeficiente de correlação múltipla entre x_1 e $\underline{x}_2$, onde x_1 representa uma única variável e $\underline{x}_2$ um conjunto de outras k variáveis, dentro um vetor p dimensional $(p > k)$ com distribuição $N_p(\underline{\mu},\Sigma)$, como sendo:

$$\rho_{1,(2,3,\ldots,k+1)} = \frac{\sqrt{\gamma_{12}\Sigma_{22}^{-1}\gamma_{21}}}{\sigma_1}, \tag{3.36}$$

onde γ_{12} é o vetor transposto de γ_{21}, cujas k componentes representam as covariâncias entre as variáveis x_1 e x_j $(j = 2, 3, \ldots, k+1)$, Σ_{22} é a matriz de covariância de $\underline{x}_2$ e σ_1^2 é a variância de x_1.

Um estimador de $\rho_{1,(2,3,\ldots,k+1)}$ é obtido substituindo os parâmetros em (3.36) pelos seus estimadores não viciados, isto é:

$$R_{1,(2,3,\ldots,k+1)} = \frac{\sqrt{g_{12}S_{22}^{-1}g_{21}}}{s_1},$$

ou pelo seu quadrado:

$$R^2_{1,(2,3,\ldots,k+1)} = \frac{g_{12}S_{22}^{-1}g_{21}}{s_1^2} \tag{3.37}$$

onde g_{12} é o vetor transposto de g_{21}, cujas k componentes representam as covariâncias amostrais entre x_1 e x_j $(j = 2, 3, ..., k)$, S_{22} é a covariância amostral de $\underline{x}_2$ e s_1^2 é variância amostral da variável x_1.

Um método de selecionar k variáveis, onde k é previamente fixado $(k < p)$, que utiliza (3.37) é apresentado em Mardia et al [1979], pag.179. Chamado de Análise de Interdependência, o método consiste em observar a explicação de cada uma das $(p - k)$ variáveis, que não estão sendo consideradas, através de um conjunto contendo k variáveis. Inicialmente forma-se um conjunto contendo k variáveis. Avalia-se, em seguida, cada variável que não faz parte desse conjunto através de:

$$\min_{i_1} R^2_{i_1,(i_2,i_3,...,i_{k+1})}, \tag{3.38}$$

onde i_1 percorre todas as $(p-k)$ variáveis que não fazem parte do conjunto fixado. Escolhendo-se aquela que foi menos explicada por esse conjunto. Considera-se outro conjunto contendo k variáveis, e avalia-se novamente aquelas que não fazem parte desse conjunto. Escolhe-se, novamente, aquela que foi menos explicada por esse novo conjunto através de (3.38). Esse procedimento é repetido para todos os possíveis conjuntos de k variáveis. Então a seleção das variáveis é obtida maximizando todos os resultados obtidos através de (3.38) e associando esse máximo ao conjunto de k variáveis. Isto é, escolhe-se o máximo entre todos os subconjuntos contendo k variáveis. Em resumo, fixa-se um conjunto de k variáveis e associa-se a esse conjunto o menor valor do CCM (Coeficiente de Correlação Múltipla) entre esse conjunto e de todas as variáveis que não fazem parte dele. Repete-se o mesmo procedimento para todos os conjuntos de k variáveis, associando sempre o menor valor do CCM entre esse conjunto e as variáveis que não fazem parte dele. Finalmente, opta-se pelo conjunto associado ao maior dentre os menores valores do CCM.

A justificativa para a adoção desse método é que em (3.38) escolhe-se o pior dos casos em cada uma dos conjuntos de k variáveis. Maximizar (3.38), significa que optou-se pelo conjunto que melhor explica a variação das $(p-k)$ variáveis que qualquer outro conjunto de k variáveis. Maiores detalhes desse método pode ser encontrado em Beale et al [1967].

O método acima, não considera a escolha do valor de k. Um critério estabelecido em Mardia et al [1979], é que o valor em (3.38) seja pelo menos $0,5$ em qualquer variável que não faça parte do conjunto previamente fixado, por exemplo, fixa-se $k = k_0$, e verifica-se se o valor de (3.38) é superior a $0,5$, para qualquer conjunto contendo k_0 variáveis (k_0 pode assumir qualquer valor inteiro positivo menor que p). A justificativa para adotar-se esse valor mínimo é que alguns autores consideram um coeficiente de correlação simples acima de $0,7$ como significativo e consequentemente seu valor quadrado próximo a $0,5$.

Um problema deste método de seleção de variáveis aparece quando o conjunto total de variáveis explicativas, p, é muito grande. Nesse caso uma comparação entre todas as combinações entre p e k $(k = 1, 2, .., p)$, tornaria o processo de seleção inviável em termos de tempo computacional. Por exemplo, considerando $p = 20$ e $k = 10$, seriam necessárias 184756 comparações para se obter a escolha do melhor conjunto de 10 variáveis, onde a seleção final ainda dependeria de que o valor em (3.38) fosse superior a $0,5$, para qualquer uma das possíveis comparações.

Devido a esse fato, será proposto neste trabalho um método alternativo, onde generaliza-se a expressão (3.37) para todos as possíveis combinações entre p e k. O objetivo é eliminar, passo a passo, uma variável que esteja associada com o maior grau de explicação dado pelas variáveis que permanecem no conjunto, no passo corrente. O processo termina quando não houver coeficiente de correlação múltipla superior a $0,5$. Isso é feito calculando-se repetidamente:

$$\max_i R^2_{i,(j_1,...,j_k)} = \max_i \frac{g_{iX_k} S^{-1}_{kk} g_{X_k i}}{s_i^2}, \tag{3.39}$$

onde as quantidades foram definidas em (3.37), e os subescritos $\{i,(j_1,...,j_k)\}$ representam que as comparações estão sendo feitas entre uma variável e um conjunto de k variáveis fixadas, assim que i percorre todas as variáveis que não faça parte do conjunto fixado.

Um resumo desse método alternativo é apresentado a seguir:

i) Considere p como sendo o número inicial de variáveis. Faça $k = p-1$, e realize p vezes a operação dada pela expressão (3.39). Se o valor máximo for menor que $0,5$, o processo termina. Caso contrário, elimine a variável associada com o valor máximo.

ii) Defina o conjunto de variáveis selecionadas como sendo todas do passo anterior, com exceção daquela que foi eliminada.

iii) Repita sucessivamente os passos (i) e (ii), considerando a redução do número de variáveis selecionadas. Observe que se (3.39) for menor que $0,5$ o processo termina e o conjunto de variáveis selecionadas é aquele estabelecido no passo anterior.

(3.40)

3.6 Testes Experimentais com Dados Simulados

3.6.1 Introdução

O objetivo desta Seção é avaliar o desempenho de cada método de seleção, apresentado neste capítulo, através da simulação apresentada no Capítulo 2. O desempenho será avaliado em termos de erros de classificação de acordo com os procedimentos apresentados

anteriormente, onde inicialmente são desenvolvidos os procedimentos de seleção e em seguida aplicado os procedimentos de classificação.

Lembrando que o nosso banco de dados simulado é formado de dois conjuntos: o primeiro com duas classes com distribuições Normais multivariadas bastante diferentes e o segundo, também com distribuições Normais, mas com algumas semelhanças (ver Seção 2.5.4). Os resultados desse estudo são apresentados a seguir.

3.6.2 Resultados da seleção e classificação com dados simulados

As Tabelas 3.1, 3.2, 3.3, 3.4 e 3.5 apresentam as taxas de erros de classificação estimadas de acordo com os resumos descritos em: (2.21), (2.43), (2.52), (2.54) e (2.56), respectivamente, em relação aos classificadores: Análise Discriminante - (AD), Modelos de Regressão Logística - (RL), Aproximação pela T^2 de Hotteling - (T^2), Classificador de Bayes (CB) e Aproximação pela Qui-quadrado (χ^2). Esses procedimentos foram executados após serem realizados os procedimentos de seleção, onde apenas as características selecionadas foram utilizadas. Cada método de seleção foi desenvolvido de acordo com os procedimentos descritos nas seções: 3.2.3, 3.3.3, 3.4.3, 3.5.2, 3.5.3 e 3.5.4, respectivamente em relação a: distância D^2 sem restrição (D^2), procedimento " Stepwise " (SW), distância D^2 restrita a apenas uma das classes (D^2_*), coeficiente de variação (CV), análise de componentes principais (CP) e análise de interdependência através do coeficiente de correlação múltipla (CR).

Cada tabela a seguir representa um estudo com base nos dois conjuntos amostrais descritos na Seção 2.5.4 e apresentados em forma de subtabelas. As amostras usadas inicialmente na modelagem, são também testadas. Assim, os valores que aparecem nos corpos das subtabelas representam as taxas de erros (em percentual) de acordo com os métodos de seleção e classificação, e também se a amostra foi utilizada para modelagem, que denotamos aqui por amostra de treinamento. As amostras de teste foram usadas somente para validação dos procedimentos. Por exemplo, na Tabela 3.1, a simulação acusou 44% de erros

do Tipo I e 55,8% de erros do Tipo II, em relação ao método de seleção D^2 no segundo conjunto usadas apenas para teste, e foram classificadas de acordo com o Classificador de Bayes (CB). O valor que aparece subescrito, abaixo de cada procedimento, indica o número de características selecionadas por cada método.

As Tabelas 3.4 e 3.5 não apresentam resultados para a amostra de treinamento, em relação a classe 2, pois estas observações não foram utilizadas na construção dos respectivos modelos de classificação.

Tabela 3.1 - Taxas de erros de classificação (em %) estimadas pelo procedimento CB de acordo com os respectivos métodos de Seleção

Conj.1 $_{p}$		D^2_{16}	SW_{10}	Conj.2	D^2_{16}	SW_{18}
Amostras						
Treinamento	π_1	0,0	0,0		35,0	4,0
	π_2	22,0	0,0		40,0	24,0
Teste	π_1	0,0	0,0		44,0	6,4
	π_2	21,3	0,4		55,8	60,4

Tabela 3.2 - Taxas de erros de classificação (em %) estimadas pelo procedimento AD de acordo com os respectivos métodos de Seleção

Conj.1 $_{p}$		D^2_{16}	SW_{10}	Conj.2	D^2_{16}	SW_{18}
Amostras						
Treinamento	π_1	0,0	1,0		3,0	2,0
	π_2	0,0	8,0		28,0	26,0
Teste	π_1	0,0	0,4		7,8	5,7
	π_2	3,8	5,8		60,3	63,1

Tabela 3.3 - Taxas de erros de classificação (em %) estimadas pelo procedimento RL de acordo com os respectivos métodos de Seleção

Conj.1 p		D^2_{16}	SW_{10}	Conj.2	D^2_{16}	SW_{18}
Amostras						
Treinamento	π_1	0,0	0,0		0,0	1,0
	π_2	0,0	0,0		20,0	20,0
Teste	π_1	0,0	0,0		1,7	3,2
	π_2	2,6	0,1		64,8	62,6

Tabela 3.4 - Taxas de erros de classificação (em %) estimadas pelo procedimento T^2 de acordo com os respectivos métodos de Seleção (0.05)

Conj.1 p		D^2_{*18}	CV_{27}	CP_{15}	CR_{18}	Conj.2	D^2_{*16}	CV_{37}	CP_{15}	CR_{18}
Amostras										
Treinamento	π_1	0,0	0,0	0,0	0,0		0,0	0,0	2,0	0,0
	π_2	-	-	-	-		-	-	-	-
Teste	π_1	8,2	7,7	4,8	3,9		2,5	2,8	7,6	3,1
	π_2	0,0	0,0	0,0	0,0		0,0	0,0	0,0	0,0

Tabela 3.5 - Taxas de erros de classificação (em %) estimadas pelo procedimento χ^2 de acordo com os respectivos métodos de Seleção (0.05)

Conj.1 p		D^2_* 18	CV 27	CP 15	CR 18	Conj.2	D^2_* 16	CV 37	CP 15	CR 18
Amostr.										
Treinamento	π_1	5,0	3,0	5,0	10,0		2,0	2,0	3,0	4,0
	π_2	-	-	-	-		-	-	-	-
Teste	π_1	35,9	43,3	13,8	25,7		14,3	37,7	16,3	11,8
	π_2	0,0	0,0	0,0	0,0		0,0	0,0	0,0	0,0

De acordo com essa simulação não houve destaque especial para qualquer método de seleção, quando observados independentemente dos métodos de classificação. Entretanto, podemos observar um bom desempenho dos métodos de seleção quando associados com o seu procedimento de classificação natural em relação ao primeiro conjunto (Conj.1). Isto é, quando associamos os procedimentos: AD de classificação e D^2 de seleção, RL de classificação e SW de seleção, assim como T^2 de classificação com D^2_* de seleção, supondo que as classes têm distribuições Normais com parâmetros distintos, o método de classificação AD associado com o método de seleção D^2 e RL associado com o SW, produzem taxas de erros bastantes satisfatórias, onde os métodos combinados apenas apresentam erros de estimação do Tipo II, em relação as amostras de teste igual a 3,8% para os métodos combinados AD/D^2 (ver Tabela 3.2) e 0,1% para os métodos RL/SW (ver Tabela 3.3). Enquanto o método de classificação T^2 associado com o método de seleção D^2_* produz taxas de erros de 8,2%, mas em relação ao erro estimado do Tipo I (ver Tabela 3.4). Entretanto, em relação ao segundo conjunto de amostras (Conj.2), observa-se um excelente desempenho do método T^2 associado com D^2_*,onde apenas apresenta erro estimado do Tipo I igual 2,5% (ver novamente Tabela 3.4). O que não acontece com as outras combinações citadas acimas, onde as taxas de erros do Tipo II são consideradas altas, especialmente para o erro estimado do Tipo II com

taxas de 60,3% para os métodos combinados AD/D^2 (ver novamente Tabela 3.2) e 62,6% para os métodos RL/SW (ver Tabela 3.3). Isso confirma os resultados obtidos no capítulo anterior, principalmente em relação ao segundo conjunto onde as médias das duas classes são supostamente iguais, mas com matrizes de covariância proporcionais. Podemos observar, portanto, um ganho em relação ao custo da construção de um sistema de reconhecimento utilizando previamente um recomendado método de seleção, pois o mesmo desempenho do classificador foi obtido com um número reduzido de variáveis utilizadas. A Tabela a seguir faz um resumo dos métodos de classificação combinado com o seu método natural de seleção.

Tabela 3.6 - Taxas de erros de classificação (em %) estimadas pelos procedimentos de classificação combinado com o seu método de seleção natural

*Conj.*1		AD/D^2	RL/SW	T^2/D^2_*	Conj.2	AD/D^2	RL/SW	T^2/D^2_*
Amostras								
Treinamento	π_1	0,0	0,0	0,0		3,0	0,0	2,0
	π_2	0,0	0,0	-		28,0	20,0	-
Teste	π_1	0,0	0,0	8,2		7,8	3,2	2,5
	π_2	3,8	2,6	0,0		60,3	62,6	0,0

Quando comparamos apenas os métodos de seleção temos que dividir nossa atenção em dois grupos distintos de tabelas: as Tabelas 3.1, 3.2 e 3.3 referem-se aos métodos que utilizam as observações de ambas as classes na construção dos métodos de seleção, enquanto que nas Tabelas 3.4 e 3.5 os métodos utilizam apenas as observações da classe π_1. Portanto essas duas últimas tabelas não apresentam resultados quando se refere aos dados de treinamento da classe π_2.

Nas Tabelas 3.1, 3.2 e 3.3, podemos observar um equilíbrio entre os métodos, parti-

cularmente em relação ao primeiro conjunto de amostras, onde as distribuições simuladas apresentam diferenças entre os parâmetros. Uma pequena vantagem ao método SW pode ser observada, pois apresenta taxas de erros, nesse primeiro conjunto, de excelente a aceitável independente do classificador, pois a maior taxa de erro estimada por esse método de seleção foi de 8% quando combinado com o método de classificação AD, enquanto o método D^2 apresenta altas taxas de erros do Tipo II em relação ao classificador de Bayes (21,2%).

Nas Tabelas 3.4 e 3.5 podemos também observar o equilíbrio entre os métodos que utilizam apenas as observações da classe π_1. Observe que o erros do Tipo I estimados por ambos os classificadores são sempre superiores aos erros do Tipo II. Isso pode ser justificado pelo controle exercido apenas ao erro do Tipo I (5% nesse caso), devido ao suposto não conhecimento das observações da classe π_2. Exceto na Tabela 3.5, onde as taxas de erros foram estimadas pelo classificador usando aproximação pela distribuição Qui-quadrado, esses erros podem ser considerados altos, e essa justificativa foi apresentada no capítulo anterior, onde sugere-se uma correção nesse classificador devido a subtituição dos verdadeiros parâmetros pelo estimadores (ver teorema 2.4.1). Assim, observando apenas a Tabela 3.4, podemos destacar o métodos CR e CP, no primeiro conjunto, com todas as taxas de erros estimadas, inclusive do Tipo I, inferiores ao proposto (5%), e no segundo conjunto, apenas o método CP apresentou taxas de erro do Tipo I estimada superior a 5%.

Em relação ao número de variáveis selecionadas, p, podemos destacar os métodos SW quando utilizamos as observações de ambas as classes e CV quando utilizamos apenas as observações da classe π_1. Esses dois métodos selecionaram o valor de p diferentemente de acordo com o conjunto de amostras (10 e 27 no primeiro e 18 e 37 no segundo, respectivamente, SW e CV - veja índice em subscritos nas tabelas, abaixo de cada procedimento). Embora estejamos diante de um estudo simulado, todos os outros métodos selecionaram características diferentes em relação ao conjunto, mas pelo menos o número delas manteve-se aproximadamente o mesmo, indicando uma maior sensibilidade dos métodos, citados acima,

em relação ao tipo de amostra utilizada.

Uma conclusão que podemos obter com base nessa simulação é que ocorrendo 1) a impossibilidade de se observar ambas as classes, para a construção do modelo de classificação, ou 2) se as médias das classes não são significativamente diferentes, um eficiente método de seleção pode ser adotado . Nesse caso sugerimos os métodos CR e D_*^2 no primeiro e segundo caso, respectivamente.

3.7 Comentários e Conclusões

Neste capítulo foram apresentadas algumas técnicas estatísticas de seleção de variáveis. O objetivo comum entre essas técnicas foi manter a identidade original da variável (característica) selecionada sem prejudicar o poder de discriminação do classificador. Em geral, para cada procedimento de classificação pode-se derivar um procedimento de seleção, esperando-se um bom desempenho quando aplicados conjuntamente. Entretanto, nada impede de se combinar um método de classificação com outro de seleção que utiliza critério particular de escolha das variáveis importantes.

O objetivo principal deste capítulo foi apresentar novas alternativas, propostas por este trabalho, na seleção de variáveis, principalmente quando não temos acesso as observações de ambas as classes na construção dos modelos de classificação. Através da utilização das principais variáveis na construção da distância de Mahalanobis ao centro da distribuição de uma das classes e através da viabilidade da utilização da análise de interdependência, introduzimos duas novas técnicas de seleção de variáveis. E a simulação mostrou que elas podem servir de opção nessas condicões de restrição com as observações de uma das classes.

Entre as técnicas que utilizam as observações de ambas as classes, a simulação indica que

elas devem ser combinadas de acordo com suas derivações, embora pesquisas nesse sentido ainda sejam necessárias para uma conclusão mais precisa.

Capítulo 4

Aplicação em Verificação de Assinaturas

4.1 Introdução

Verificação automática de assinaturas tem recebido cada vez mais atenção dos pesquisadores da área de reconhecimento de padrões. Plamondon e Lorette [1989] apresentam uma pesquisa sobre os trabalhos que utilizam as mais diversas técnicas desenvolvidas até então. Mais recentemente, Leclerc e Plamondon [1994], apresentam uma nova pesquisa sobre o tema, com os trabalhos desenvolvido no período entre 1989 a 1993. O ponto em comum entre as discussões é que uma assinatura contém muito mais informação do que aparenta. Por mais cuidadoso que um indivíduo seja, ele próprio não será capaz de reproduzir duas assinaturas perfeitamente idênticas, pois no momento da reprodução, estão associados fatores pessoais e emocionais característico de um particular instante. Entretanto, um conjunto de características pode ser formado a partir dessas reproduções e utilizadas na construção de um modelo que irá representar um padrão básico para as assinaturas daquele indivíduo.

O planejamento de um sistema de verificação de assinaturas é feito em vários estágios, entre eles destacamos: Aquisição dos Dados, Pré-processamento do Sinal, Extração (Seleção) de Característica e Classificação da Assinatura. Tipos diferentes de abordagens do problema

da verificação podem ser definidos em relação ao estágio da aquisição dos dados. Um sistema pode ser considerado Dinâmico ou Estático, dependo se as assinaturas foram coletadas através de mecanismo que registram os sinais da assinatura, no momento em que a assinatura estiver sendo coletada, simultaneamente com o tempo, considerada como pertencente à categoria das dinâmicas, ou se foram coletadas através de mecanismo que registram os sinais da assinatura apenas por uma imagem fixa daquela assinatura, onde nesse caso será considerada dentro da categoria das estáticas.

Outro estágio que também define uma abordagem diferente do problema é o processo de extração de características. Em geral, os sistemas estáticos adotam critérios próprios de escolha de parâmetros que vão representar a imagem recebida em um vetor, onde cada coordenada representa uma particular característica da imagem da assinatura do indivíduo. Em sistemas dinâmicos, essa transformação também pode ser adotada. Entretanto, é possível utilizar transformações que representem completamente o sinal recebido em função do tempo, onde os valores que essas funções assumem, em cada instante de tempo, podem ser considerados como uma característica. Em geral, espera-se que sistemas de verificação de assinaturas que utilizam esse último tipo de procedimento de extração de características, apresentem melhor desempenho que aqueles que utilizam algum critério específico de escolha de parâmetros, pois acredita-se que esses critérios de escolha podem vir a desprezar algumas características importantes (Hastie et al [1991]). Entretanto, os custos operacionais associados aos sistemas que utilizam funções para representar os sinais recebidos, são muito mais altos, principalmente em relação as dificuldades de planejamento e recursos de memória computacional. Portanto, consideramos neste trabalho apenas os sistemas que utilizam algum critério de escolha das caracteríticas, onde é aplicado uma transformação os dados recebidos para um vetor composto das principais características do indivíduo.

De modo geral, este trabalho se preocupa apenas com os dois últimos estágios citados acima. Em particular, supõe que as informações recebidas, sobre uma assinatura de um

indivíduo, estejam representadas em um vetor com dimensão p e depois, todas as reproduções dessa assinatura são agrupadas em uma matriz com dimensão (n x p), onde n representa o número de observações (reproduções) daquela assinatura e p o número de características importantes da assinatura. A partir daí é aplicado um procedimento de seleção dessas características, e então um modelo de classificação, usando apenas as características mais importantes, é desenvolvido. Esses modelos irão responder, com certa margem de erro, se uma assinatura desconhecida pertence ou não ao indivíduo em questão.

Nesse capítulo apresentamos inicialmente um estudo envolvendo a aplicação das principais técnicas apresentadas nos capítulos anteriores, tanto para seleção de variáveis, como para classificação de padrões, aplicados a um problema de verificação de assinaturas. São utilizados dois bancos de dados que variam segundo a natureza dos dados: Assinatura Dinâmica e Assinatura Estática, cujas observações foram apresentadas em vetores contendo as respectivas características. Em seguida apresentaremos um estudo envolvendo uma comparação entre as taxas de erros empírica, através de curvas operacionais de referências, de modo a se estabelecer limites de separação de classes com base em taxas aceitáveis.

4.2 Aplicação de Técnicas de Seleção de Variáveis e Classificação de Assinaturas: Dinâmicas e Estáticas

4.2.1 Introdução

Nesta seção utilizaremos os bancos de dados adquiridos segundo os mecanismos dinâmico e estático apresentados em Lee e Berger [1996] e Lizárraga [1996], respectivamente, na seleção de variáveis e classificação de assinaturas representadas por vetores de características.

Os dados apresentados em Lee e Berger [1996] referem-se a um conjunto composto de 1.825 assinaturas dinâmicas de um indivíduo, sendo 1.000 genuínas e 825 falsificadas. Elas

foram adquiridas através de um digitalizador, onde as saídas dos dados são representadas em termos de par de funções temporais $(X(t_i), Y(t_i))$, indicando a posição da caneta no espaço R^2, em cada instante de tempo, com relação a uma origem arbitrária. A extração de características transformou os dados primitivos $(X(t_i), Y(t_i))$, em vetores compostos de 42 características. A lista completa da definição destas 42 características é apresentada no apêndice do referido trabalho. Os detalhes da obtenção das assinaturas genuínas e falsificadas também são apresentadas nesse trabalho. Uma relação com a definição de cada característica dinâmica é apresentada no Apêndice-B.

Os dados apresentados em Lizárraga [1996] referem-se a um conjunto de 950 assinaturas estáticas em relação a 5 indivíduos, sendo 550 genuínas, 100 falsificadas aleatoriamente e 300 falsificadas após um treinamento habilitado. A diferença básica entre a falsificação aleatória e a chamada habilitada, é que as habilitadas seguiram um procedimento mais específico, onde o falsificador conhecia pelo menos a imagem da assinatura genuína, enquanto que as falsas aleatórias são consideradas aquelas assinaturas que são totalmente alheias ao indivíduo em questão. Utilizamos apenas as assinaturas referentes a um dos indivíduos, escolhido ao acaso, sendo, portanto, nosso conjunto de assinaturas estática composto de 270 assinaturas, com 110 genuínas, 100 falsas aleatórias e 60 falsas habilitadas. Elas foram adquiridas através de um "scanner " de mesa, que gera imagens digitais de assinaturas. O processo de extração de características transformou cada imagem em um vetor composto de 62 características escolhidas pelos autores, que representavam basicamente as características de inclinação e contorno dos traços de uma assinatura estática. A forma de obtenção das assinaturas genuínas e falsificadas é apresentada com detalhes no referido trabalho. Uma relação com a definição de cada característica estática é também apresentada no Apêndice-B.

4.2.2 Descrição dos resultados

As Tabelas 4.1, 4.2, 4.3 e 4.4 apresentam as taxas de erros (em percentual) estimadas pelos métodos de classificação: Análise Discriminante - (AD), Modelos de Regressão

Logística - (MRL), e Aproximação pela T^2 de Hotelling - (T^2), após a execução dos métodos de seleção: Distância D^2 Studentizada (sem restrição) - (D^2), "Stepwise" - (SW), respectivamente para os métodos de classificação (AD) e (MRL), e Distância D^2 Studentizada com restrição - (D^2_*), Coeficiente de Variação - (CV), Análise de Componentes Principais - (CP) e Análise de interdependência através do Coeficiente de Correlação Múltipla - (CR), em relação ao método de classificação T^2, como descritos nos Capítulos 2 e 3.

Os resultados apresentados nas tabelas a seguir se baseiam em dois conjuntos, cujas amostras de treinamento, são compostas, respectivamente, de $n_1 = n_2 = 30$ e $n_1 = n_2 = 50$ observações para os dois tipos de assinaturas (dinâmica e estática), enquanto que são utilizadas 900 observações genuínas e 725 observações falsificadas para a validação das assinaturas dinâmicas (Tabelas 4.1 e 4.2), e 60 observações genuínas, 30 falsas habilitadas e 80 falsas aleatórias para a validação das assinaturas estáticas (Tabelas 4.3 e 4.4). Todos os testes seguiram exatamente o proposto nos referidos resumos, apresentados nos Capítulos 2 e 3, respectivamente para classificação e seleção. Lembrando que os métodos de classificação AD e MRL não apresentam limites de separação de classes (limiar) que dependa de uma específica distribuição de probabilidade, enquanto que o método aproximado pela T^2 utiliza a distribuição F de Snedecor e seus resultados podem variar de acordo com as exigências do pesquisador. Nesse estudo utilizamos 5% como taxa de erro do Tipo I esperado.

Entre as 1.825 observações de assinaturas dinâmicas, foi realizado incialmente um sorteio para a escolha de 200 observações usadas para treinamento (100 genuínas e 100 falsificadas), e 1.625 para teste (900 genuínas e 725 falsificadas). Dentre as observações de treinamento, realizamos novos sorteios para a escolha dos conjuntos com 30 e 50 observações, tanto para genuínas como falsificadas, respectivamente. Todas as observações de teste foram utilizadas para validação dos modelos.

Entre as 270 assinaturas estáticas, também foi realizado inicialmente um sorteio para

escolha de 100 observações de treinamento (50 dentre as 110 genuínas, 30 dentre as 60 falsas habilitadas e 20 dentre as 100 falsas aleatórias). Outro sorteio foi realizado para a escolha das 30 observações genuínas e 30 falsas habilitadas do primeiro conjunto de treinamento, enquanto que o segundo conjunto utilizou todas as observações de treinamento (50 genuínas, 30 falsas habilitadas e 20 falsas aleatórias). Ainda em relação aos dados de assinaturas estáticas, o número de características escolhidas inicialmente pelo processo de extração de características realizado pelos autores, era composto de 62 variáveis, entretanto nosso trabalho detectou singularidade entre esse conjunto. Mais precisamente, 21 das 62 variáveis iniciais eram formadas por combinação linear das restantes, isso reduziu nosso número inicial de possíveis características discriminantes para 41 variáveis.

Observamos também que o método de seleção D_*^2 não se aplica quando o número de observações amostrais , entre as genuínas ($n = n_1$), é inferior ao número de características iniciais (p), por problemas de identificabilidade dos parâmetros. A rigor todos os métodos de seleção também impõe essa condição, embora possamos utilizar recursos matemáticos, tais como a utilização da pseudo-inversa, para decomposição de matrizes singulares. Entretanto, apenas o método D_*^2 supõe explicitamente a condição ($n > p$). Com exceção dos métodos CV e SW, todos os demais métodos utilizaram esse artifício matemático na obtenção dos resultados quando o número de observações foi inferior ao número de variáveis. É importante destacar o fato das nossas conclusões e comentários se basearem apenas nas amostras com 50 observações, pois é pouco esperado adotar um procedimento estatístico condicões contrárias. Entretanto, usando outros métodos de aproximação dos resultados, tais como Redes Neurais e Sintática, essa condição não é imposta.

Em resumo, os resultados a seguir contém as taxas de erros estimadas através de cada método de classificação combinado com o respectivo método de seleção. O número de variáveis selecionadas por cada método (k) aparece logo abaixo da descrição do modelo combinado. Por exemplo, na Tabela 4.1, observamos que o método de classificação AD,

combinado com o método de seleção D^2 (AD/D^2), estimou uma taxa de erro do Tipo I de 0,00% (não rejeitou qualquer das assinaturas genuínas dentres as 900 testadas), enquanto que estimou em 32,14% a taxa de erro do Tipo II (aceitou 233 dentre as 725 assinaturas falsificadas testadas), sendo $k = 9$ o número de características selecionadas, entre as $p = 42$ possíveis, pelo método D^2. Na Tabela 4.3, o mesmo método estimou uma taxa de 16,67% para o erro do Tipo I (rejeitou 10 das 60 assinaturas genuínas testadas), enquanto que estimou em 26,67% a taxa de erro do Tipo II habilitada (aceitou 8 das 30 assinaturas falsas habilitadas) e estimou em 37,5% a taxa de erro do Tipo II aleatória (aceitou 30 das 80 assinaturas falsas aleatórias), sendo $k = 10$ o número de características selecionadas pelo mesmo método de seleção.

Tabela 4.1 - Taxa estimada de erro de classificação após a seleção de variáveis em assinaturas dinâmicas com número de observações de treinamento - 30

Tipo de Ass. p=42	AD/D^2 k=9	RL/SW k=14	T^2/D^2_* –	T^2/CV k=27	T^2/CP k=14	T^2/CR k=16
Genuína	0,00	0,00	-	28,33	6,22	28,00
Falsa	32,14	30,90	-	1,52	0,83	0,42

Tabela 4.2 - Taxa estimada de erro de classificação após a seleção de variáveis em assinaturas dinâmicas com número de observações de treinamento - 50

Tipo de Ass. p=42	AD/D^2 k=14	RL/SW k=14	T^2/D^2_* k=17	T^2/CV k=25	T^2/CP k=16	T^2/CR k=24
Genuína	0,00	0,00	12,22	41,11	8,33	10,89
Falsa	33,10	32,83	0,00	0,00	0,00	0,00

Tabela 4.3 – Taxa estimada de erro de classificação após a seleçao de variáveis em assinaturas estáticas com número de observações de treinamento - 30

Tipo de Ass. p=41	AD/D^2 k=10	RL/SW k=38	T^2/D^2_* –	T^2/CV k=7	T^2/CP k=29	T^2/CR k=13
Genuína	16,67	0,00	-	16,67	18,33	1,67
Fals. Habilit.	26,67	30,00	-	66,67	60,00	56,67
Fals. Aleatória	37,50	47,50	-	0,00	0,00	0,00

Tabela 4.4 - Taxa estimada de erro de classificação após a seleção de variáveis em assinaturas estáticas com número de observações de treinamento - 50

Tipo de Ass. p=41	AD/D^2 k=21	RL/SW k=38	T^2/D^2_* k=13	T^2/CV k=7	T^2/CP k=22	T^2/CR k=13
Genuína	6,67	13,33	11,67	0,00	5,00	8,33
Fals. Habilit.	10,00	10,00	10,00	50,00	36,67	30,00
Fals. Aleatória	5,00	2,50	0,00	0,00	0,00	0,00

4.2.3 Análise dos resultados

Através das tabelas acima, podemos observar um comportamento semelhante entre os métodos de classificação que utilizam as observações de ambas as classes na construção dos modelos de classificação (AD e RL), com pequena vantagem (menor taxa de erro) para os métodos combinados RL/SW em relação as assinaturas dinâmicas, e pequena vantagem para os métodos combinados AD/D^2 em relação as assinaturas estáticas, principalmente quando o número de observações na amostra de treinamento é maior ($n_1 = n_2 = 50$).Vale salientar que nas assinaturas estáticas, os métodos combinados AD/D^2 utilizaram um número bem menor de variáveis.

Quando o método de classificação usa a aproximação pela T^2 de Hotelling (T^2), podemos destacar o bom desempenho do método combinado com a seleção via análise de componentes principais (CP) em relação aos dados de assinaturas dinâmicas. Entretanto, essa combinação de métodos (T^2/CP) não mantém o mesmo desempenho quando o tipo de assinatura é estática, principalmente quando o número de observações na amostra de treinamento é menor ($n_1 = n_2 = 30$). De modo geral, ainda podemos destacar o bom comportamento dos métodos combinados (T^2/D_*^2), pois embora não apresente taxas de erros inferiores em nenhum dos casos, apresenta uma certa regularidade, sempre com taxas aceitáveis (inferiores a 13%).

A presença das assinaturas falsas aleatórias, no caso das assinaturas estáticas, indicam a necessidade do uso de um grande número de observações de treinamento (pelo menos $n > p$), para a obtenção de melhores desempenho entre os métodos, principalmente em relação aos métodos de classificação que utilizam as observações de ambas as classes (RL e AD), pois apresentam taxas para esse tipo de erro elevadas. De fato, esse têm sido um problema prático dos procedimentos que utilizam técnicas estatísticas, em comparação as outras técnicas aplicadas aos sistemas de verificação de assinaturas, como por exemplo redes neurais, que necessitam de bem menos observações como suposição inicial. Em geral, as técnicas estatísticas são desenvolvidas assintoticamente, enquanto que grande parte dos sistemas propostos na literatura tomam como base amostras de treinamentos com um número de observações que variam entre 5 e 10 observações da assinatura (ver por exemplo Hastie et al [1991] e Nelson et al [1994]).

4.3 Estudo Empírico para Escolha dos Limites de Separação de Classes

4.3.1 Introdução

Os procedimentos adotados na seção anterior consideram exatamente os passos sugeridos pelos resumos descritos em cada método e apresentados anteriormente. Isso dificulta uma possível comparação entre os mesmos, já que cada um adota critério próprio de separação entre as classes. Por exemplo, o critério de classificação via AD utiliza uma função discriminante que depende da observação a ser testada, e compara seu valor com uma constante fixada através dos dados de treinamento: se essa função estimada para essa observação for superior a uma constante fixada é razoável classificá-la como genuína. O MRL utiliza o modelo como preditor da probabilidade de sucesso, isto é, o modelo logístico estima a probabilidade da observação a ser testada pertencer a classe das genuínas. Nesse sentido é natural classificar como genuína a observação cuja probabilidade estimada pela resposta do modelo for maior ou igual a 0,5. Entretanto, esse valor pode ser modificado de acordo com o ajuste do modelo, sendo necessário um estudo mais detalhado, a cerca dos parâmetros do vetor $\underline{\beta}$, caso o pesquisador queira basear suas conclusões finais através deste método. Finalmente, a classificação via aproximação pela distribuição T^2 de Hotelling compara a distância estimada de uma observação ao centro da distribuição da classe das genuínas e compara essa distância com um valor tabelado pela distribuição F de Snedecor. Se a distância estimada para essa observação não for significativamente grande (não for superior a um fixado valor tabelado), decide-se por aceitá-la como genuína.

Entretanto, esse limite de separação pode ser adaptado a uma determinada realidade. Em outras palavras, podemos relaxar esses valores para tomar como base uma linha de separação que estabeleça critérios de escolha de acordo com o pesquisador. Por exemplo, em sistemas de segurança máxima, seria natural o pesquisador procurar controlar o erro do Tipo II, onde a aceitação de uma observação falsa poderia causar sérios prejuízos. Entretanto,

isso exigiria maior rigor também em relação ao erro do Tipo I, pois aumentaria as chances de se rejeitar uma observação genuína. Em sistemas bancários, esse critério certamente iria aumentar o tempo de atendimento de cada cliente além de aumentar também o tempo de permanência dos clientes em filas de espera.

Nessa seção avaliaremos o comportamento de cada método em relação aos limites de separação das classes em cada um dos tipos de assinaturas: dinâmica e estática.

4.3.2 Apresentação dos resultados empíricos

Os gráficos a seguir representam as taxas de de erros dos Tipos I e II, estimadas através dos respectivos procedimentos de classificação e seleção combinados, em relação aos diversos limites de separação de classes. Isto é, depois de fixado um determinado limite de separação de classes, os métodos combinados estimam as probabilidades de erros dos Tipos I e II, respectivamente.

A rigor, o valor ótimo de separação entre classes é aquele cujos erros do Tipo I e Tipo II são iguais a zero. Veja, por exemplo, o Gráfico 4.4a (a seguir), com limites entre 9 e 10, e o Gráfico 4.5a, com limites entre 2,5 e 5. Esses gráficos indicam que, estabelecendo-se um limite nas respectivas faixas, os métodos combinados T^2/CV e T^2/CP estimam taxas de erros dos dois tipos iguais a zero (não rejeitaria qualquer assinatura genuína, assim como não aceitaria qualquer assinatura falsa). Esses limites foram obtidos após a execução dos procedimentos de classificação/seleção combinados e aplicados às observações de teste, considerando os modelos construídos com base na amostra de treinamento com 50 observações. Observe que esses limites variam de um procedimento para o outro dependendo do classificador. Os limites obtidos através do método de classificação AD dependem da função discriminante aplicada as observações de teste, os limites do método RL dependem da probabilidade de sucesso e finalmente os limites da T^2 dependem da distribuição F de Snedecor.

Os Gráficos 4.7 e 4.8 representam as Curvas Operacionais de Referências (COR), que avaliam os dois tipos de erros estimados sem expressar explicitamente os limites de separação. É natural que exista um compromisso entre eles, isto é, quando um erro cresce o outro decresce. Enquanto que a situação ótima é aquela onde a curva se aproxima, dos eixos principais. Esta situação é apresentada no Gráfico 4.7, em relação aos métodos combinados T^2/CV e T^2/CP, reproduzindo o que ocorreu nos Gráficos 4.4a e 4.5a, quando as linhas que representam os erros estimados se cruzam com o eixo das abscissas. É comum esses gráficos virem acompanhados de linhas verticais e horizontais para identificarem regiões onde os métodos podem ser considerados aceitáveis.

Gráfico 4.1a - Taxa de erro estimada por limite de separação em assinaturas dinâmicas através do método AD/D^2

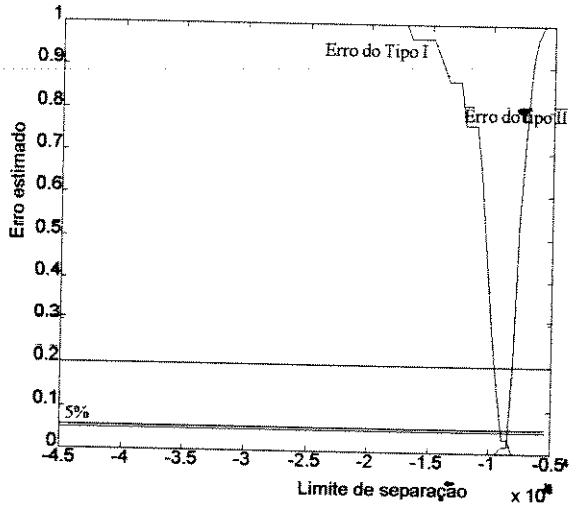


Gráfico 4.1b - Taxa de erro estimada por limite de separação em assinaturas estáticas através do método AD/D^2

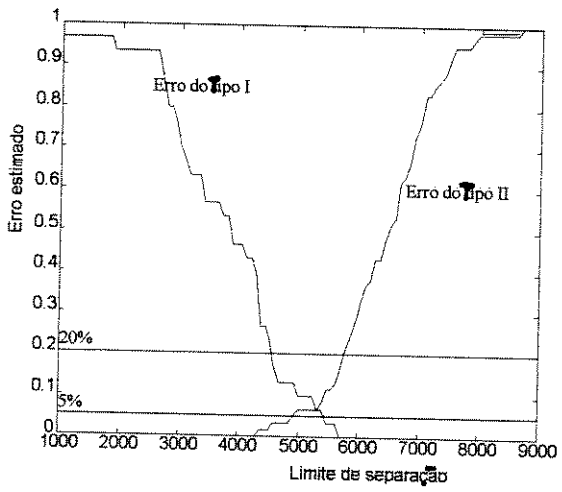


Gráfico 4.2a - Taxa de erro estimada por limite de separação em assinaturas dinâmicas através do método RL/SW

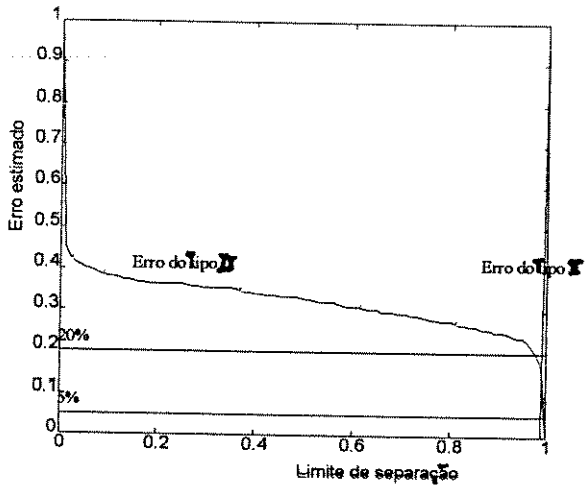


Gráfico 4.2b - Taxa de erro estimada por limite de separação em assinaturas estáticas através do método RL/SW

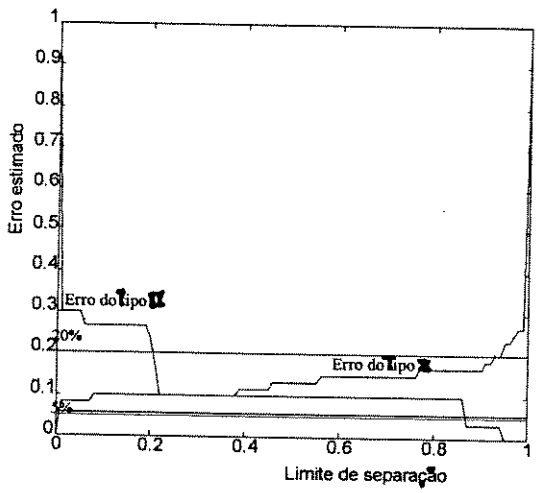


Gráfico 4.3a - Taxa de erro estimada por limite de separação em assinaturas dinâmicas através do método T^2/D_*^2

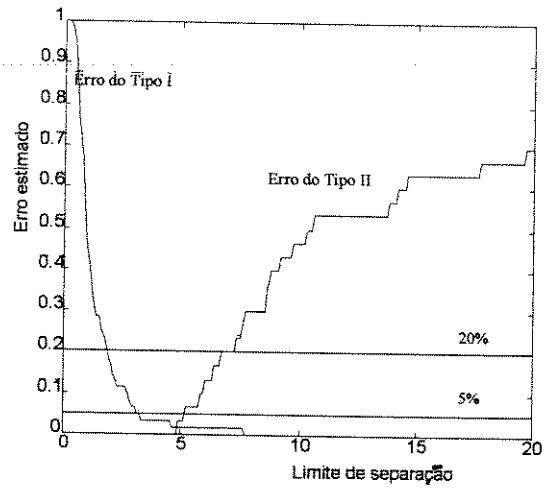


Gráfico 4.3b - Taxa de erro estimada por limite de separação em assinaturas estáticas através do método T^2/D_*^2

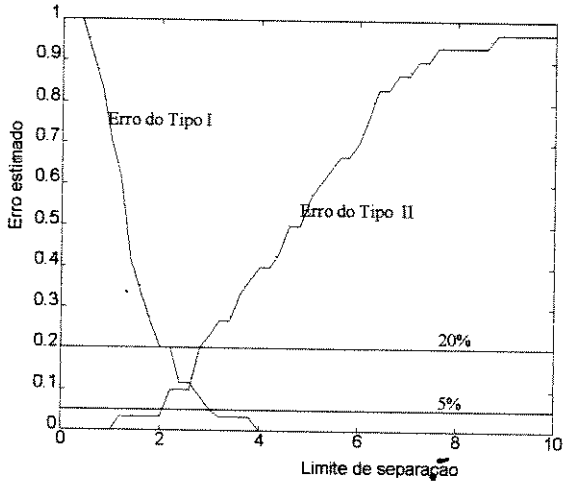


Gráfico 4.4a - Taxa de erro estimada por limite de separação em assinaturas dinâmicas através do método T^2/CV

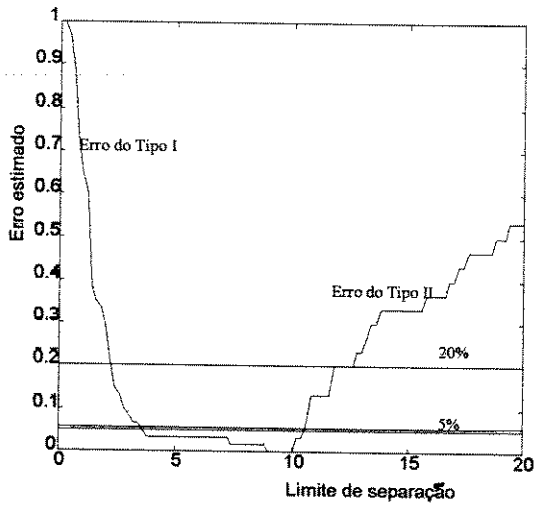


Gráfico 4.4b - Taxa de erro estimada por limite de separação em assinaturas estáticas através do método T^2/CV

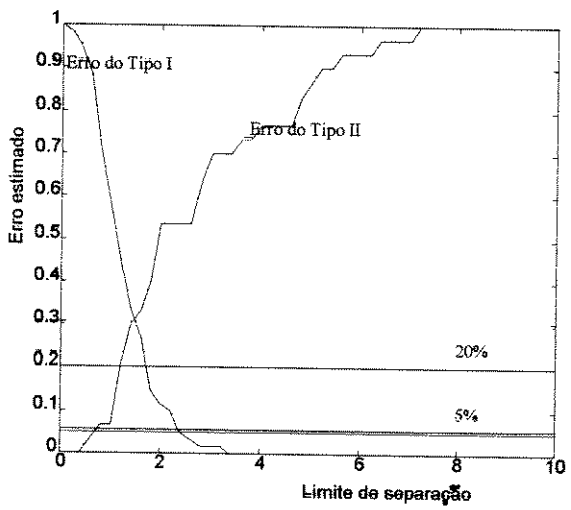


Gráfico 4.5a - Taxa de erro estimada por limite de separação em assinaturas dinâmicas através do método T^2/CP

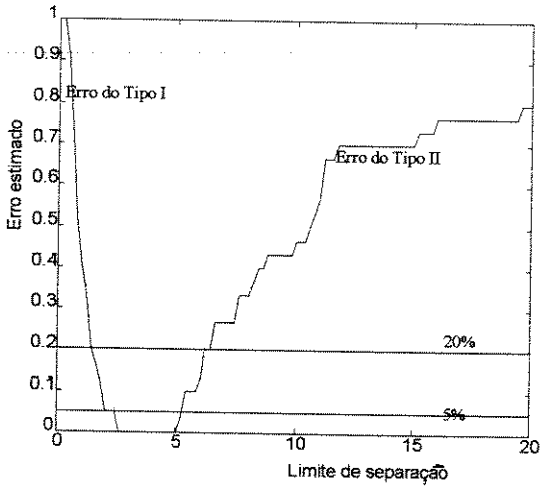


Gráfico 4.5b - Taxa de erro estimada por limite de separação em assinaturas estáticas através do método T^2/CP

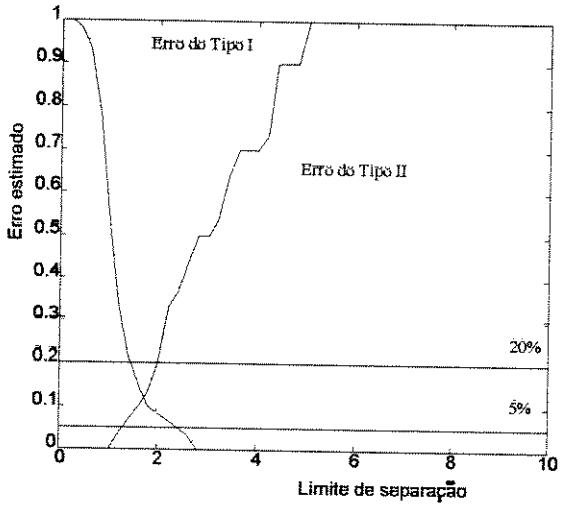


Gráfico 4.6a - Taxa de erro estimada por limite de separação em assinaturas dinâmicas através do método T^2/CR

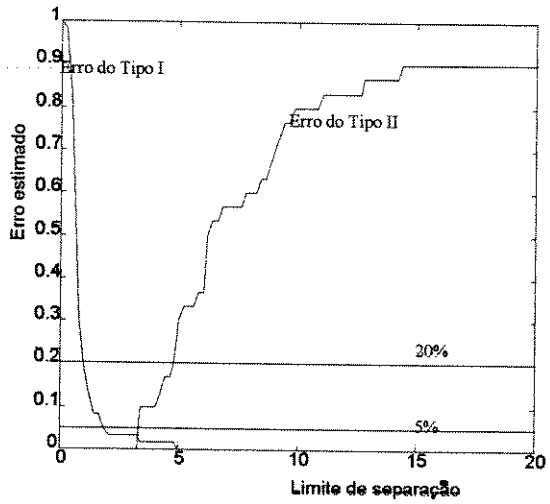
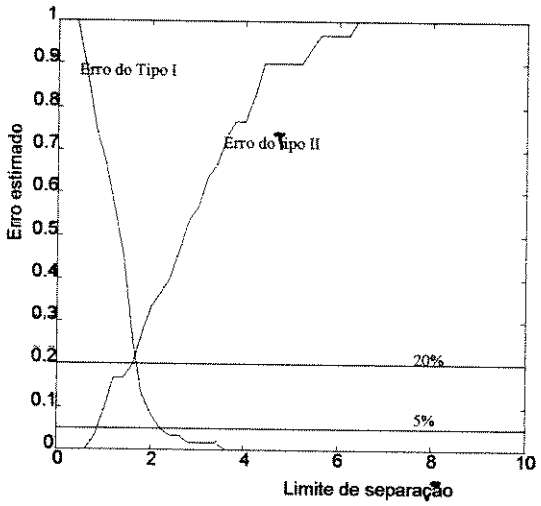


Gráfico 4.6b - Taxa de erro estimada por limite de separação em assinaturas estáticas através do método T^2/CR



Os Gráficos 4.7 e 4.8 resumem o comportamento de cada procedimento combinado em comparação com os demais. Através desses gráficos podemos avaliar o método que obteve melhor desempenho em relação aos seus próprios critérios de separação.

Gráfico 4.7 - Taxas de erros estimadas para assinaturas dinâmicas

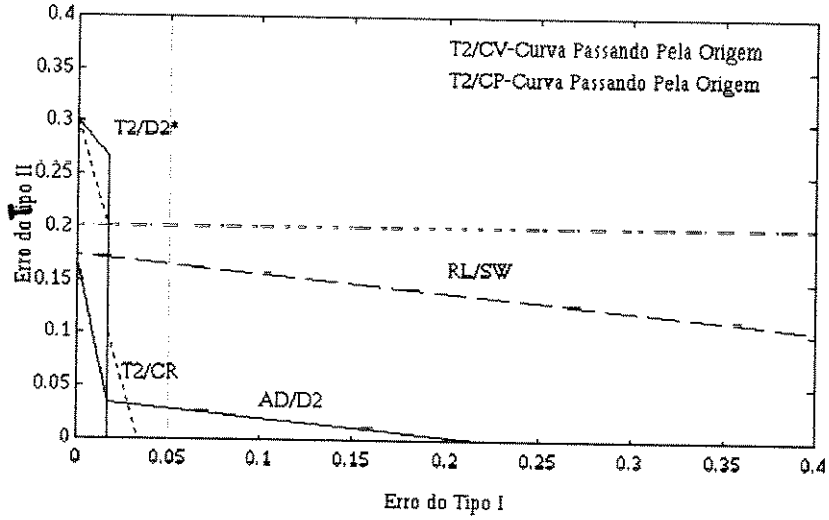
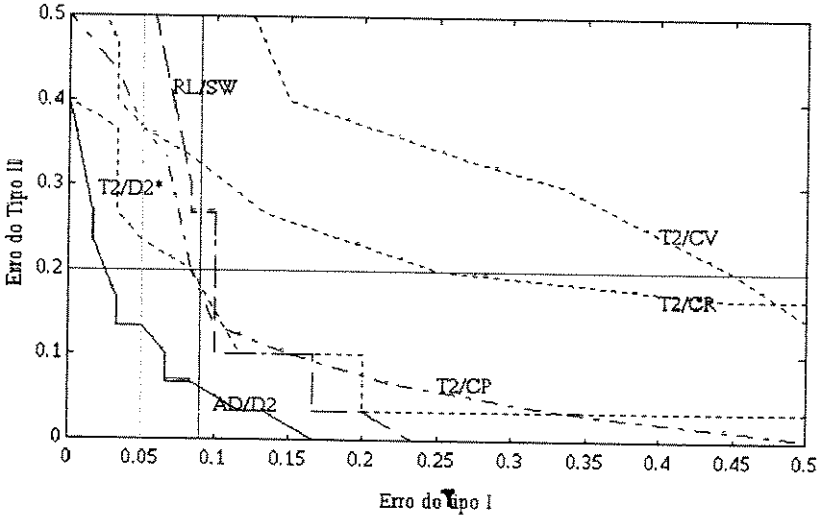


Gráfico 4.8 - Taxas de erros estimadas para assinaturas estáticas



4.3.3 Análise dos resultados empíricos

De acordo com os Gráficos 4.1 a 4.8, podemos destacar os procedimentos T^2/CV e T^2/CP, pois proporcionaram taxas de erros iguais a zero em relação às assinaturas dinâmicas. Isso quer dizer que se tivessemos optado pelos limites de separação de classes nas respectivas faixas (qualquer valor entre 9 e 10 para o método T^2/CV e qualquer valor entre 2,5 e 5 para o método T^2/CP não teríamos erros de classificação estimados para as assinaturas dinâmicas. Entretanto, o método T^2/CV não obteve o mesmo desempenho para as assinaturas estáticas. Nesse tipo de assinatura o método AD/D^2 obteve melhor desempenho.

De modo geral, todos os métodos, com exceção da combinação RL/SW tiveram excelentes desempenhos em relação as assinaturas dinâmicas, pois todos tiveram taxas estimadas abaixo de 5%, para o erro do Tipo I igual ao erro do Tipo II. Observe que, em todos os gráficos relacionados com as assinaturas dinâmicas o ponto de cruzamento entre as linhas ficou abaixo da margem de 5%, com excerção do Gráfico 4.2a. Em relação as assinaturas estáticas, apenas a combinação AD/D^2, obteve excelente desempenho, com taxas estimadas para erros iguais, próximas a 6%. Entretanto, nesse tipo de assinaturas os métodos RL/SW, T^2/D^2_* e T^2/CP tiveram desempenhos aceitáveis, com taxas estimadas para erros iguais, dadas aproximdamentes por 10%, 11,5% e 13%, respectivamente.

Em relação ao método de classificação T^2, podemos confirmar o excelente desempenho, principalmente em relação às assinaturas dinâmicas, onde combinado com os métodos de seleção CV e CP obteve taxas estimadas para erros iguais zero, além do razoável desempenho com assinaturas estáticas através da combinação com os métodos de seleção D^2_* e CP.

4.4 Comentários e Conclusão

Existe na literatura diversos estudos sobre sistemas de verificação de assinaturas, cada um adotando bancos de dados quase que exclusivos. Isso tem dificultado possíveis com-

parações entre eles. Em geral, admite-se que a taxa de erro do Tipo I não seja superior a 5%, enquanto que a taxa de erro do Tipo II não seja superior a 20% (ver Nelson et al [1994]). Nesse campo de validação, todos o métodos aqui apresentados, poderiam servir como opção em relação as assinaturas dinâmicas, enquanto que apenas a combinação AD/D^2 pertenceria a esse campo, em relação às assinaturas estáticas. Entretanto, admitindo um erro do Tipo I de até 7,5%, a combinação T^2/D_*^2 e T^2/CP, poderia servir como opção, com a vantagem de se utilizar apenas as observações genuínas para treinamento, reduzindo assim o custo operacional do sistema.

Capítulo 5

Conclusões Finais e Sugestões para Futuros Trabalhos

5.1 Introdução

Este trabalho foi direcionado no sentido de dar apoio técnico aos sistemas de RP através de métodos estatísticos na solução dos problemas. A principal aplicação foi voltada para os sistemas de verificação de assinaturas, mas sua aplicação pode ser generalizada, dependendo dos objetivos do pesquisador.

O desenvolvimento das técnicas aqui apresentadas, supõe que o banco de dados seja apresentado em forma de matrizes, cujas linhas representam as observações e as colunas as características principais de um determinado padrão ou classe.

Sendo assim, o objetivo principal deste trabalho foi desenvolver novas técnicas nas áreas de seleção de variáveis e classificação de padrões, e avaliar o desempenho em relação as principais técnicas existentes na literatura.

5.2 Contribuição do Trabalho

Os procedimentos de classificação que utilizam métodos estatísticos têm sido desenvolvidos principalmente para situações onde é possível treinar o classificador utilizando as observações de ambas as classes. Quando esta situação não pode ser considerada, o método utilizando a aproximação pela distribuição Qui-quadrado tem sido empregado na construção dos modelos de classificação. Esse trabalho apresenta o desenvolvimento de uma nova técnica de classificação, utilizando a aproximação pela distribuição T^2 de Hotelling e suas relações com a distribuição F de Snedecor, ao invés da aproximação pela distribuição Qui-quadrado.

Um estudo simulado foi realizado para identificar uma situação onde o novo classificador fosse preferível mesmo quando comparado com aqueles que utilizam as observações de ambas as classes na construção dos modelos. Este é o caso quando os vetores de médias das classes não são significativamente diferentes, e as matrizes de covariância são proporcionais. Inclusive confirma sua vantagem em relação ao classificador que utiliza a aproximação pela distribuição Qui-quadrado.

Dois novos métodos de seleção de variáveis foram introduzidos. O primeiro desenvolvido a partir do procedimento de classificação aproximado pela distribuição T^2 de Hotelling e o segundo viabilizando a utilização da análise de interdependência pela correlação múltipla. A simulação mostrou que em situações especiais, como suposição de normalidade na distribuições de cada classe com médias iguais, os dois métodos propostos podem ser considerados de excelente desempenho, sendo preferíveis, mesmo em comparação aos métodos que utilizam as obsevações de ambas as classes.

Destacam-se nas aplicações em sistemas de verificação de assinaturas os métodos combinados AD/D^2 pela seu excelente desempenho em ambos os tipos de assinaturas: dinâmicas e estáticas. Enquanto que, de forma geral, duas combinações de métodos T^2/D^2_* e T^2/CP podem servir de opção em situações onde apenas as observações genuínas possam ser utilizadas

no treinamento. Enquanto que, no caso particular de assinaturas dinâmicas, os métodos T^2/CV e T^2/CP obtiveram taxas de erros iguais, estimadas em 0%, sugerindo sua aplicação em situações semelhante.

5.3 Sugestões para Futuros Trabalhos

A título de sugestão para futuros trabalhos, podemos destacar a implementação de uma sistema automático, utilizando esses procedimentos de seleção e classificação, aplicados em sistemas com mais de duas classes possíveis. Uma extenção das técnicas aplicadas a duas classes poderia ser feita por exemplo em reconhecimento de caracter.

Em termos de classificadores que supõem o acesso irrestrito às observações de ambas as classes, existe a necessidade de um estudo teórico sobre a construção de modelos via AD utilizando taxas de erros esperadas do Tipo I diferentes das taxas de erros esperadas para o Tipo II. Um estudo mais detalhado sobre métodos de seleção que possam ser aplicados aos modelos logísticos, também poderão ser desenvolvidos.

Em termos de aproximação estatística, o número de observações necessárias para treinamento do classificador têm sido um dos principais problemas, principalmente em sistemas de verificação de assinaturas. Entretanto, existem técnicas de reamostragem, tais como Bootsptrap e Jackknife (Efron e Gong [1970]), que podem ser investigadas para a solução desse problema. Uma sugestão seria aproveitar as observações testadas no passado e reaproveitá-las para compor o conjunto de treinamento futuro. Uma avaliação periódica poderia servir de referência na obtenção do número ótimo de observações desse novo conjunto de treinamento.

Apêndice - A

Neste Apêndice apresentamos alguns resumos de teoremas, definições e resultados essenciais ao desenvolvimento do nosso trabalho, que podem ser encontrados na literatura. Detalhes das provas de alguns lemas apresentados são indicados com referências. Dividimos a apresentação desse Apêndice em 5 partes classificadas de Apêndice A-1 até A-5.

Apêndice A-1 Indepêndencia

O lema a seguir, embora não tenha sido usado rigorosamente, é de fundamental importância para a demonstração do Teorema 2.4.1.

Lema A.1 : Se X é uma matriz (n x p) de dados, cujas linhas são observações de um vetor aleatório com distribuição $N_p(\underline{\mu}, \Sigma)$, e se definirmos $Y = AXB$ e $Z = CXD$, onde A, B, C e D são matrizes com dimensões adequadas com X, então os elementos de Y são independentes dos elementos de Z se e somente se ocorrerem uma das seguintes situações:

a)$B'\Sigma D = 0$

b)$AC' = 0$

Prova: (ver Mardia et al [1979], pag.457.)•

Apêndice A-2 Inversa e Partição de Matrizes

Apresentamos nesse Apêndice algumas definições e lemas fundamentais à Álgebra de Matrizes.

Definição A.2 : A inversa de uma matriz A é uma matriz única A^{-1} satisfazendo

$$AA^{-1} = A^{-1}A = I,$$

A inversa de uma matriz A existe se e somente se A é não singular, isto é, se e somente se $|A| \neq 0$, onde $|A|$ representa o determinante da matriz A.

Lema A.2 : Para as matrizes A e A^{-1} particionadas, isto é,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{e} \quad A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix},$$

as seguintes propriedades são válidas:

a) $(cA)^{-1} = c^{-1}A^{-1}$

b)$(AB)^{-1} = B^{-1}A^{-1}$

c)A única solução de $A\underline{x} = \underline{b}$ é $\underline{x} = A^{-1}\underline{b}$

d)Se as inversas de A e C existem, então para $A(p \text{ x } p)$, $B(p \text{ x } n)$, $C(n \text{ x } n)$ e $D(n \text{ x } p)$ tem-se que:

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$$

e)Se as inversa de A_{11} e A_{22} existem, então os elementos de A^{-1} são:

$$A^{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} \qquad A^{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$
$$A^{12} = -A^{11}A_{12}A_{22} \qquad A^{21} = -A_{22}^{-1}A_{21}A^{11}$$

Alternativamente A^{12} e A^{21} podem ser escritos como

$$A^{12} = -A_{11}^{-1}A_{12}A^{22} \qquad A^{21} = -A^{22}A_{21}A_{11}^{-1}$$

Prova: (ver Mardia et al [1979])•

Apêndice A-3 Diferenciação e Maximização de Matrizes

De acordo com as derivadas de $f(\mathbf{X})$, onde $\mathbf{X}$ é uma matriz (n x p) e f é uma função dos elementos de $\mathbf{X}$, representada por:

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left(\frac{\partial f(\mathbf{X})}{\partial x_{ij}}\right),$$

onde x_{ij} representam o elemento (i, j) da matriz $\mathbf{X}$, e $\underline{x}$ representa um vetor aleatório, temos os seguintes resultados:

a) $\frac{\partial \underline{a}'\underline{x}}{\partial \underline{x}} = \underline{a}$, onde $\underline{a}$ é um vetor constante.

b) $\frac{\partial \underline{x}'\underline{x}}{\partial \underline{x}} = \underline{x}$

c) $\frac{\partial \underline{x}' A \underline{x}}{\partial \underline{x}} = 2A\underline{x}$, onde A é uma matriz com dimensão compatível com o vetor $\underline{x}$

d) $\frac{\partial \underline{x}' A \underline{y}}{\partial \underline{x}} = A\underline{y}$

e) $\frac{\partial |\mathbf{X}|}{\partial x_{ij}} = \begin{cases} X_{ij} \text{ se todos os elementos de } \mathbf{X}(n \text{ x } n) \text{ são distintos} \\ X_{ij} \text{ se } \mathbf{X} \text{ é simétrica e } i = j \\ 2X_{ij} \text{ se } \mathbf{X} \text{ é simétrica e } i \neq j \end{cases}$

onde X_{ij} é o elemento (i, j) da matriz de cofatores de $\mathbf{X}$.

O valor de $\underline{x}$ que minimiza

$$f(\underline{x}) = (\underline{y} - A\underline{x})'(\underline{y} - A\underline{x})$$

é dado por:

$$A'A\underline{x} = A\underline{y}$$

Apêndice A-4 Distribuições de Probabilidades

Apresentaremos a seguir as definições de algumas funções de distribuições descritas no desenvolvimento do nosso trabalho e que foram citadas sem as devidas expressões.

- Se x é uma variável aleatória com distribuição normal com média μ e variância σ^2 , então sua *pdf* é dada por:

$$f(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^2/\sigma^2\} \qquad -\infty \leq x \leq \infty,$$
$$-\infty \leq \mu \leq \infty, \qquad \text{e} \qquad \sigma > 0$$

se $\mu = 0$ e $\sigma^2 = 1$, x é dita ter distribuição normal padrão.

- Sejam $x_1, x_2, ..., x_p$ variáveis aleatórias independentes e identicamente distribuídas segundo uma distribuição normal padrão, então

$$y = x_1^2 + x_2^2 + ... + x_p^2$$

tem distribuição Qui-quadrado com p graus de liberdade e sua *pdf* é dada por:

$$f(y) = \{2^{\frac{p}{2}}\Gamma(\frac{1}{2}p)\}^{-1} y^{(\frac{p}{2}-1)} \exp\{-\frac{y}{2}\} \qquad y \geq 0, p \geq 0$$

onde

$\Gamma(x+1) = x\Gamma(x) \qquad x \geq 0 \qquad e \qquad \Gamma(n+1) = n!$ onde n é um inteiro positivo.

- Sejam u e v variáveis aleatórias independentes com u e v distribuídas segundo uma distribuição Qui-quadrado com p e q graus de liberdade, respectivamente, então

$$x = \frac{u/p}{v/q}$$

tem distribuição F de Snedecor com p graus de liberdade no numerador e q graus de liberdade no denominador.

- Seja x uma variável aletória com distribuição normal padrão, independente de u sendo esta uma variável aleatória com distribuição Qui-quadrado com p graus de liberdade, então

$$t = \frac{x}{(u/p)^{\frac{1}{2}}}$$

tem distribuição t-Student com p graus de liberdade e sua *pdf* é dada por:

$$f(t) = \left[\Gamma(\frac{1}{2}p + \frac{1}{2})/\{(p\pi)^{\frac{1}{2}}\Gamma(\frac{1}{2}p)\}\right]\left(1 + t^2/p\right)^{-(p+1)/2} \qquad -\infty \leq t \leq \infty.$$

Observe que se t tem distribuição t-Student com p graus de liberdade, então t^2 tem distribuição F de Snedecor com 1 grau de liberdade no numerador e p graus de liberdade no denominador.

- Seja $\underline{x}$ um vetor aleatório p-dimensional com distribuição normal multivariada, com vetor de média $\underline{\mu}$ e matriz de covariância Σ, então sua *pdf* é dada por:

$$f(\underline{x}) = (2\pi\,|\Sigma|)^{-\frac{1}{2}} \exp\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{x}-\underline{\mu}\right)\}$$

onde Σ é uma matriz definida positiva e $|\Sigma|$ denota seu determinante.

- Seja $\underline{t}$ um vetor aleatório p-dimensional com distribuição t-Student multivariada, com vetor de média $\underline{\mu}$ e matriz de covariância Σ, então sua *pdf* é dada por:

$$f(\underline{t}) = \frac{C_p\,|\Sigma|^{-\frac{1}{2}}}{\left[1+\upsilon^{-1}\left(\underline{t}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{t}-\underline{\mu}\right)\right]^{(\upsilon+p)/2}},$$

onde $C_p = \frac{\Gamma((\upsilon+p)/2)}{(\pi\upsilon)^{p/2}\Gamma(\upsilon/2)}$ e υ o número de graus de liberdade.

- Seja M uma matriz aleatória com distribuição Wishart, com parâmetro de escala Σ, e m graus de liberdade, então sua *pdf* é dada por:

$$f(M) = \frac{|M|^{(m-p-1)/2}\exp(-\frac{1}{2}tr\Sigma^{-1}M)}{2^{mp/2}\pi^{p(p-1)/4}\,|M|^{m/2}\prod_{i=1}^{p}\Gamma(\frac{1}{2}(m+1-i))},$$

onde $tr\Sigma$, denota o traço da matriz Σ.

Apêndice A-5 Distribuição Condicional

Os lemas apresentados a seguir são generalizações dos casos univariados envolvendo a distribuição condicional de vetores com distribuição normal multivariada.

Lema A.5.1 : Se $\underline{x}= (\underline{x}_1', \underline{x}_2')'$ tem distribuição $N_p(\underline{\mu}, \Sigma)$, então $\underline{x}_1$ e $\underline{x}_{2.1} = \underline{x}_2 - \Sigma_{21}\Sigma_{11}^{-1}\underline{x}_1$ tem as seguintes distribuições e são estatisticamente independentes:

$$\underline{x}_1 \sim N_r(\underline{\mu}_1, \Sigma_{11}), \qquad \underline{x}_{2.1} \sim N_s(\underline{\mu}_{2.1}, \Sigma_{22.1})$$

onde

$$\underline{\mu}_{2.1} = \underline{\mu}_2 - \Sigma_{21}\Sigma_{11}^{-1}\underline{\mu}_1 \qquad e \qquad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Prova: (ver Mardia et al [1979], pag. 63)•

Lema A.5.2 : Usando as suposições e notações do Lema A.5.1, a distribuição condicional de $\underline{x}_2$ para um dado valor de $\underline{x}_1$ é:

$$\underline{x}_2/\underline{x}_1 \sim N_s(\underline{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1 - \underline{\mu}_1), \Sigma_{22.1}),$$

isto é, a distribuição condicional de $\underline{x}_2/\underline{x}_1$ é Normal, e sua média condicional é dada por:

$$E\{\underline{x}_2/\underline{x}_1\} = \underline{\mu}_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}\underline{x}_1 = \underline{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1 - \underline{\mu}_1),$$

e a matriz de covariância condicional $\underline{x}_2$ é a mesma que $\underline{x}_{2.1}$ $(\Sigma_{22.1})$.

Prova: (ver Mardia et al [1979], pag. 63)•

Apêndice - B

Neste Apêndice apresentamos a descrições referentes aos dois conjuntos de assinaturas (Dinâmica e Estática) de acordo com sua características citadas neste trabalho.

Apêndice - B-1 Assinaturas Dinâmicas

Consideramos que o conjunto de dados de saída do processo de aquisição de dados é formado por uma sequência tríplece (t, x, y), representando o tempo, a coordenada horizontal e a coordenada vertical, respectivamente, para um determinado ponto da assinatura. Um pre-processamento foi aplicado ao cálculo final da característica, como por exemplo normalização dos dados, para mimimizar o efeito da resolução básica da tabela gráfica. Após o pré-processamento a extração das 42 características dinâmicas foi a seguinte:

1 - Velocidade média de escrita.

2 - Velocidade máxima de escrita.

3 - Tempo até ocorrer a velocidade máxima de escrita.

4 - Tempo entre o *pen down* e o movimento inicial da caneta.

5 - Tempo de duração da escrita da assinatura (tempo de assinatura).

6 - Duração total do tempo de escrita (tempo de escrita).

7 - Velocidade mínima de escrita na direção horizontal.

8 - Tempo até ocorrer a velocidade mínima na direção horizontal.

9 - Total de pontos (pingos) na assinatura (pingos na letra *i*, por exemplo).

10 - Tempo de duração médio dos pingos.

11 - Número total de componentes desconectadas.

12 - Tempo inicial da segunda componente.

13 - Direção inicial da assinatura.

14 - Direção da linha que conecta o ponto inicial da primeira componente e o ponto inicial da segunda componentes.

15 - Direção da linha que conecta o ponto inicial da primeira componete e o ponto final da segunda componente.

16 - Direção inicial da segunda componente.

17 - Direção da linha que conecta o inicio e o fim da assinatura.

18 - Direção da velocidade positiva na direção horizontal.

19 - Direção da velocidade negativa na direção horizontal.

20 - Direção da velocidade positiva na direção vertical.

21 - Direção da velocidade negativa na direção vertical.

22 - Duração média da velocidade positiva na direção horizontal.

23 - Duração média da velocidade negativa na direção horizontal.

24 - Duração média da velocidade positiva na direção vertical.

25 - Duração média da velocidade negativa na direção vertical.

26 - Número de vezes que a velocidade cruza o zero na direção horizontal.

27 - Número de vezes que a velocidade cruza o zero na direção vertical.

28 - Diferença entre as velocidades máxima e média na direção horizontal.

29 - Diferença entre as velocidades máxima e média na direção vertical.

30 - Diferença entre as velocidades máxima e mínima na direção horizontal.

31 - Diferença entre as velocidades máxima na direção horizontal e mínima na direção vertical.

32 - Diferença entre as velocidades máxima e mínima na direção vertical.

33 - Razão entre o tempo em que a assinatura é máxima no eixo vertical e o tempo total da escrita.

34 - Razão entre o tempo em que a assinatura é mínima no eixo vertical e o tempo total da escrita.

35 - Área mínima do retângulo completamente coberto pela assinatura.

36 - Razão entre o comprimento total da assinatura e área mínima do retângulo completamente coberto pela assinatura.

37 - Diferença entre a coordenada horizontal inicial e a coordenada horizontal mínima.

38 - Diferença entre a coordenada horizontal final e a coordenada horizontal máxima.

39 - Diferença entre a coordenada horizontal final e a coordenada horizontal mínima.

40 - Razão entre a diferença entre as coordenadas máxima e mínima na direção horizontal e a diferença entre as coordenadas máxima e mínima na direção vertical.

41 - Desvio padrão horizontal.

42 - Desvio padrão vertical.

A seguir apresentaremos os vetores de médias ($\overline{x}_d^{(1)}$ e $\overline{x}_d^{(2)}$) e de desvios padrões ($s_d^{(1)}$ e $s_d^{(2)}$) baseados em 100 observarções genuínas (1) e 100 observações falsas (2), de um indivíduo, coletadas na Universidade de Cornell (EUA).

$$\overline{x}_d^{(1)} = [$$

44,0	343,7	5,3	0,0	5,5	4,9	-112,7
4,7	1,0	0,1	5,0	0,9	4,2	5,5
5,7	5,6	6,2	1,9	1,0	1,8	1,7
38,0	-34,0	44,8	-47,9	20,2	37,5	233,6
172,3	494,3	69,6	424,7	1,0	0,8	1054,1
0,2	4,4	0,0	52,2	2,6	14,4	3,0

]

$$s_d^{(1)} = [$$

2,75	141,27	0,16	0,03	0,16	0,13	21,9
1,54	0,22	0,02	0,65	0,14	0,10	0,29
0,91	1,44	0,62	0,08	0,07	0,15	0,15
2,36	1,82	2,49	2,64	1,51	3,24	90,51
21,22	191,23	91,78	111,76	0,01	0,16	94,81
0,02	0,65	0,10	3,05	0,28	0,74	0,16

]

$$\overline{x}_d^{(2)} = [$$

30,75	215,68	7,24	0,11	7,78	5,66	-61,33
2,69	1,20	0,09	9,27	0,91	3,72	4,74
4,83	4,13	5,73	1,87	1,02	1,95	1,62
31,13	-26,34	34,14	-36,53	21,33	35,38	140,47
113,75	311,36	30,95	280,41	0,95	0,29	756,16
0,28	3,21	-0,04	48,07	3,16	14,29	2,50

]

$$\underline{s}_d^{(2)} = [$$

5,06	129,77	1,69	0,13	1,40	0,93	17,36
2,82	0,74	0,04	2,91	0,41	2,39	1,65
1,27	2,41	1,69	0,26	0,12	0,26	0,25
3,62	2,42	4,20	6,94	2,42	4,00	86,79
20,04	177,81	81,12	110,56	0,09	0,24	156,61
0,05	0,75	0,22	3,97	0,57	1,05	0,28

$]$

Apêndice - B-2 Assinaturas Estáticas

São consideradas aqui dois tipos básicos de extrações de características de uma imagem da assinatura: caracteríticas de inclinação dos traços (32 características) e características de contorno dos traços (30 características).

Os 32 elementos básicos iniciais são seguimentos de retas compostas de cinco pixels. Cada elemento representa uma inclinação diferente em relação ao seu sucessor de aproximadamente 11 graus. Após uma operação, chamada pelo autor de erosão da assinatura, é feita uma contagem de todos os pixels para compor as 32 coordenadas iniciais.

Para os 30 elementos básicos restantes, define-se inicialmente um elemento básico como sendo um quadrado de 3 x 3 pixels pretos com origem no centro do quadrado. Seguindo o mesmo procedimento de erosão, é feita uma contagem para detectar os seguimentos de retas verticais e horizontais, além da variação de pixels na direção diagonal, formando 6 novas coordenadas. Esse procedimento é repetido cinco vezes para uma nova imagem dilatada do ciclo anterior.

A seguir apresentaremos os vetores de médias ($\underline{\overline{x}}_e^{(1)}$ e $\underline{\overline{x}}_e^{(2)}$) e de desvios padrões ($\underline{s}_e^{(1)}$ e $\underline{s}_e^{(2)}$) baseados em 110 observarções genuínas (1) e 60 observações falsas (2), de um indivíduo, coletadas na Universidade de Campinas, UNICAMP-DECOM-FEEC.

$$\overline{\underline{x}}_e^{(1)} = [$$

587,15	504,22	473,29	462,25	454,01	453,48	486,15
570,34	691,01	752,94	754,93	710,55	649,49	706,95
720,03	668,37	587,15	504,23	473,29	462,25	454,01
453,48	486,15	570,34	691,01	572,94	574,93	710,55
649,49	706,95	720,03	668,37	163,46	114,58	5,26
5,26	411,73	7,86	182,05	186,35	3,46	3,46
317,25	7,33	189,38	254,72	1,77	1,77	225,85
7,52	196,94	304,03	1,40	1,40	164,68	8,05
208,58	333,09	0,74	0,74	126,13	8,47	

]

$$\underline{s}_e^{(1)} = [$$

103,55	90,74	83,71	80,95	80,56	81,15	87,66
101,09	113,05	124,76	126,24	121,68	113,71	118,93
117,94	114,00	103,55	90,74	83,71	80,95	80,56
81,15	87,66	101,09	113,05	124,76	126,24	121,68
113,71	118,93	117,94	114,00	39,69	18,97	2,97
2,97	38,80	2,38	32,96	16,57	1,98	1,98
32,34	2,47	25,72	16,31	1,38	1,38	27,57
2,49	21,49	16,43	1,35	1,35	24,20	2,41
20,39	17,53	1,02	1,02	20,96	2,31	

]

$$\overline{x}_e^{(2)} = [$$

492,78	457,05	441,37	424,98	405,13	406,42	417,12
447,07	526,13	610,73	668,83	651,68	588,35	603,60
582,20	534,25	492,78	457,05	441,37	424,98	405,13
406,42	417,12	447,07	526,13	610,73	668,83	651,68
588,35	603,60	582,20	534,25	105,62	85,30	8,25
8,25	412,40	5,28	129,02	160,35	3,83	3,83
281,95	4,92	139,05	225,00	1,43	1,43	197,37
4,83	146,55	269,57	0,98	0,98	151,50	4,83
154,25	297,43	0,40	0,40	123,75	4,93	]

$$\underline{s}_e^{(2)} = [$$

130,30	132,21	140,84	144,48	143,91	140,84	136,87
132,67	150,32	177,65	201,73	224,10	228,03	238,61
207,94	156,94	130,30	132,21	140,84	144,48	143,91
140,84	136,87	132,67	150,32	177,65	201,73	224,10
228,03	238,61	207,94	256,94	51,12	26,33	6,14
6,14	85,06	3,18	48,14	23,15	2,41	2,41
46,12	3,39	45,34	19,17	1,21	1,21	35,56
3,55	42,45	15,81	1,21	1,21	30,33	3,36
39,89	15,07	0,87	0,87	24,65	3,14	]

Bibliografia

- Anderson [1958] - Anderson, T. W. - *Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, New York, 1958.

- Aidar [1992] - Aidar, T. - Regressão Linear Ponderada na Seleção de Variáveis Independentes em Modelos de Regressão Logística. *Dissertação de Mestrado, IMECC-UNICAMP*, Campinas, 1992.

- Aitchison et al [1977] - Aitchison, J., Habbema, J. D. F. and Kay, J. W. - A critical comparison of two methods of discrimination. *J. Roy. Statist. Soc. C*, **26**, 15-25, 1977.

- Andrews [1971] - Andrews, H. C. - Multidimensional rotations in feature selection. *IEEE trans. Comput.*, Short Notes, 1045-1051,1971.

- Anderson e Bahadur [1962] - Anderson, T.W and Bahadur, R. R. - Classification in two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.*, **33**, 420-431, 1962.

- Beale et al [1967] - Beale, B. M. L., Kendal, M. G. and Mann, D. W. - The discarding of variable in multivariate analysis. *Biometrika*, **54**, 357-366, 1967.

- Bose e Roy [1938] - Bose, R. C. and Roy, J. N. - The distribution of Studentized D^2 - Statististic. *Sankhyar*, **4**, 19-23, 1938.

- Breslow e Day [1980] - Breslow, N. E. and Day, N. E. - Statistical methods in cancer research, vol. 1: The analysis of case-control studies. *International Agency for Research on Cancer*, Lyon, 1980.

- Cornfield [1962] - Cornfield, J. - Join dependent of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc.*, **21**, 58-61, 1962.

- Cox [1970] - Cox, C. R. - *The Analysis of Binary Data.* Methuen & Co. Ltda, London, 1970.

- Day e Kerridge [1967] - Day, N. E. and Kerridge, D. E. - A general maximum likelihood discriminant. *Biometrics*, **23**, 313-323, 1967.

- Dorbson [1990] - Dorbson, A. J. - *An Introduction to Generalized Linear Models.* Chapman and Hall, London, 1990.

- Draper e Smith [1981] - Draper, N. R. and Smith, H. - *Applied Regression Analysis.* 2nd ed. Wiley, New York, 1981.

- Duda e Hart [1973] - Duada, R.O. and Hart, P.E. - *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

- Efron [1975] - Efron, B. - The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistic Association*, **70**, 892-898, 1975.

- Efron e Gong [1983] - Efron, B. and Gong, G. - A leisurely look at the Bootstrap, the Jackknife and Cross-Validation. *The American Statistician*, **37**, 36-48, 1983.

- Fisher [1936] - Fisher, R.A. - The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179-188, 1936.

- Fisher e van Ness [1973] - Fisher, L. and van Ness, J. W. -Admissible discriminant analysis. *Journal of the American Statistic Association*, **68**, 603-607, 1973.

- Gnanadesikam e Gupta [1970] - Gnanadesikan, M. and Gupta, S. S. - A selection procedure for multivariate normal distributions in terms of the generalized variance. *Technometrics*, **12**, 103-117, 1970.

- Gupta et al [1992] - Gupta, L., Wang, J., Charles, A. and Kisatsky, P. - Prototype selection rules for neural network training. *Pattern Recognition*, **25**, 1401- 1408, 1992.

- Hastie et al [1991] - Hastie, T., Kishon, E., Clark, M. and Fan, J. - A statistical model for signature verification. *Proc. IEEE Conf. on Sistems, Man and Cybernetics*, Charlotteville, VA. oct.-1991.

- Hauck e Doner [1977] - Hauck, W. W. Jr. and Doner, A. - Wald's test as applied to hypoteses in logit analysis. *Journal of the American Statistic Association*, **72**, 851-853, 1977.

- Heydorn [1970]- Heydorn,R.P. - Redundancy in feature extraction. *IEEE Symposium on Feature Extraction and Selection in Pattern Recognition* - Argonne, Ill, 1970.

- Hosme et al [1989] - Hosme, W. D., Jovanovic, B. and Lameshow, S. - Best subsets logistic regression. *Biometrics*, **45**, 1265-1270, 1989.

- Hotelling [1931] - Hoteling, H. - The generalization of Student ratio. *Ann. Math. Statist.*, **2**, 360-378, 1931.

- James [1981] - James, B.R. - *Probabilidade: Um Curso em Nível Intermediário.* Projeto Euclides-CNPq, 1981.

- Jeffreys [1961] - Jeffreys, H. - *Theory of Probability.* Cledon Press, Oxford, 1961.

- Jolliffe [1972] - Jolliffe, I. T. - Discarding variables in principal component analysis I: Artificial data. *Appl. Statist.*, **21**, 160-173, 1972.

- Johnson e Wichern [1982] - Johnson, R. A. and Wichern, D. W. - *Applied Multivariate Statistical Analysis.* Printice-Hall, New Jersey, 1982

- Kshirsagar [1972] - Kshirsagar, A. M. - *Multivariate Analysis* - Marcel Dekker, New York, 1972.

- Leclerc e Plamondon [1994] - Leclerc, F. and Plamondon, R. - Automatic signature verification: The state of the art. *International Journal of Pattern Recognition and Artificial Inteligence*, **8**, 643-660, 1994.

- Lee [1991] - Lee,L. L. - On-line Systems for Human Signature Verification. *Ph.D. Thesis*, Cornell University, 1991.

- Lee e Berger [1996] - Lee,L. L. and Berger, T. - Reliable on-line signature verification systems for point-of-sales applications. *IEEE Trans. Pattern Anal. Machine Intell.*, **18**, 643-647, 1996.

- Lee e Landgrebe [1993] - Lee, C. and Landgrebe, D. A. - Feature extraction based on decision boundaries. *IEEE Trans. Pattern Anal. Machine Intell.*, **15**, 388-400, 1993.

- Lizárraga [1996] - Lizárraga, M. G. - Um sistema automático de consulta e verificação de assinaturas estáticas. *Dissertação de Mestrado. DECOM- FEEC - UNICAMP*, Campinas, 1996.

- Malina [1981] - Malina, W. - On an extrended Fischer criterion for feature selection. *IEEE Trans. Pattern Anal. Machine Intell.*, **5**, 611-614, 1981.

- Mardia et al [1979] - Mardia, .V., Kent, J. T. and Bibby, J.M. - *Multivariate Analysis.* Academic Press, London, 1979.

- McFadden [1976] - McFadden, D. - A comment on discriminant analysis 'versus ' logit analysis. *Annals of Economic and Social Measurement*, **5**, 511-523, 1976

- McCullagh e Nelder [1989] - McCullagh, P. and Nelder, J. A. - *Generalized Linear Models*, 2nd ed., Chapman and Hall, London, 1989.

- Morrison [1976] - Morrison, D. F. - *Multivariate Analysis Methods*, 2nd ed. McGraw Hill, New York, 1976.

- Mucciardi e Gose [1970] - Mucciardi, A. N. and Gose, A. E. -A comparison of seven techniques for choosing subset for pattern recognition properties. *Symposium on Feature Extraction and Selection in Pattern Recognition*, Argonne, Ill., 1970.

- Nelson et al [1994] - Nelson,W., Turin, W. and Hastie, T. - Statistical methods for on-line signature verification. *International Journal of Pattern Recognition and Artificial Inteligence*, **8**, 749-769, 1994.

- Plamondon e Lorette [1989] - Plamondon, R. and Lorette, G. - Automatic signature verification and writer identification: The state of the art. *Pattern Recogn.*, **22**, 107-131, 1989.

- Press e Wilson [1978] - Press, S. J. and Wilson, S. L. - Choosing between regression and discriminant analysis. *Journal of the American Statistic Association*, **73**, 699-705, 1978.

- Rao [1973] - Rao,C.R. - *Linear Statistical Inference and Its Application.* Wiley, New York, 1973.

- Searn [1980] - Stearn,S.D. - On selecting features for pattern classifiers. *Tecnical Report - Security Systems Departament - GTE*, Sylvania Mountain View, California 94042, 1980.

- Schalkoff [1992] - Shalkoff, R. - *Pattern Recognition Statistical, Structural and Neural Approaches.* Wiley, 1992.

- Truett et al [1967] - Truett, J. Cornefield, J. and Kannel, W. - A multivariate analysis of the risk of coronary heart disease in framing. *Journal of Chronic Disease*, **20**, 511-524,1967.

- Walker e Duncan [1967] - Walker, S. H. and Duncan, D. B. - Estimation of the probability of an event as a function of several independent variables. *Biometrika*, **54**, 167-179, 1967.

- Welch [1939] - The generalization of Student's problem when several population are involved. *Biometrika*, **34**, 28-35.