Universidade Técnica de Lisboa
Instituto Superior de Economia e Gestão

Mestrado em Matemática Aplicada à Economia e Gestão

Métodos não paramétricos em Análise Discriminante

Regina Maria Agostinho Soares

Orientação: Doutora Ana Maria Pires de Melo Parente

Júri:
Presidente: Doutor António Luís Silvestre
Vogais: Doutora Ana Maria Pires de Melo Parente
Doutora Maria Antónia da Conceição Abrantes Amaral Turkman
Dra. Maria Filomena de Faria Santos Gonçalves Pimenta

Lisboa, Outubro de 1997

Universidade Técnica de Lisboa
Instituto Superior de Economia e Gestão

Mestrado em Matemática Aplicada à Economia e Gestão

Métodos não paramétricos em Análise Discriminante

Regina Maria Agostinho Soares

Orientação: Doutora Ana Maria Pires de Melo Parente

Júri:
Presidente:
Doutor António Luís Silvestre, professor associado do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa.

Vogais:
Doutora Ana Maria Pires de Melo Parente, professora auxiliar do Instituto Superior Técnico da Universidade Técnica de Lisboa.
Doutora Maria Antónia da Conceição Abrantes Amaral Turkman, professora catedrática da Faculdade de Ciências da Universidade Nova de Lisboa.
Dra. Maria Filomena de Faria Santos Gonçalves Pimenta, professora auxiliar convidada do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa.

Lisboa, Outubro de 1997

MÉTODOS NÃO PARAMÉTRICOS EM ANÁLISE DISCRIMINANTE

RESUMO:

Neste trabalho pretendeu-se fazer uma exposição sumária dos métodos não paramétricos mais populares de estimativas de densidade e apresentar a discriminação não paramétrica.

Após a apresentação dos principais aspectos teóricos, fez-se uma aplicação de quatro regras discriminantes - duas obtidas pelos métodos clássicos e duas pelos métodos não paramétricos - a um conjunto de exemplos ilustrativos.

Foi ainda efectuado um pequeno estudo de simulação para averiguar o comportamento dos quatro métodos, quando aplicados a um conjunto de distribuições representativas de situações próximas da realidade, sendo a comparação do desempenho dos vários métodos feito através das estimativas das taxas de erro.

Palavras chave: análise discriminante, métodos não paramétricos, taxas de erro, regras de classificação, núcleo, vizinhos mais próximos.

Agradecimentos

Desejo agradecer em primeiro lugar à Prof. Doutora Ana Maria Pires de Melo Parente, orientadora desta dissertação, pelos preciosos comentários e sugestões que efectuou ao longo da elaboração do presente trabalho e pela forma dedicada como acompanhou a sua realização.

Agradeço também o apoio e simpatia das minhas amigas e colegas do INE, Carla Martins, Patrícia Valadas, Isabel Apolinário, Fátima Cardoso, Cristina Sousa e Carla Grosa. Gostaria também de deixar uma palavra de agradecimento ao Prof. Doutor João Manuel C. Santos Silva pelo seu incentivo e pela disponibilidade que sempre demonstrou. Gostaria também de deixar uma palavra de agradecimento a todos os meus amigos e familiares, que me apoiaram e incentivaram ao longo deste ano.

Desejo ainda agradecer ao Quim, à Catarina e à Beta pelo carinho, apoio e compreensão sempre presentes.

Ao Quim e à
Catarina

Índice

1 Introdução **6**

1.1 Notação, formulação e critérios de decisão 9

1.1.1 Notação e formulação 9

1.1.2 Critérios de decisão 10

1.2 Discriminação com modelos normais 15

1.2.1 Modelo homocedástico 16

1.2.2 Modelo heterocedástico 17

1.3 Taxas de erro das regras de classificação 18

1.3.1 Taxa de erro óptima 18

1.3.2 Taxa de erro actual 18

1.3.3 Taxa de erro actual esperada 19

1.4 Estimação das taxas de erro 19

1.4.1 Taxas de erro de substituição 19

1.4.2 Taxas de erro aparentes 19

1.4.3 Taxas de validação cruzada 20

1.4.4 Taxas obtidas por *"bootstrap"* 21

1.5 Métodos paramétricos parciais 21

1.5.1 Discriminante logística 22

1.5.2 Discriminante logística quadrática 23

2 Métodos não paramétricos **25**

2.1 Introdução 25

2.1.1 Propriedades estatísticas dos estimadores de densidade 25

2.1.2 Histogramas 27
2.1.3 Estimadores do núcleo 29
2.1.4 Estimador dos k-vizinhos mais próximos 32
2.1.5 Estimador generalizado dos k-vizinhos mais próximos 33
2.1.6 Estimador do núcleo adaptativo 34
2.1.7 Estimador das séries ortogonais 35
2.1.8 Estimador de funções peso gerais 37
2.2 Escolha do parâmetro de alisamento 37
2.2.1 Método do núcleo 37
2.3 Estimadores não paramétricos para dados multivariados 43
2.3.1 Histogramas multivariados 43
2.3.2 Estimadores do produto de núcleos 45
2.3.3 Estimadores do núcleo 46
2.3.4 Estimadores do núcleo para dados binários 48
2.3.5 Estimador do núcleo adaptativo 49
2.3.6 Estimador dos k-vizinhos mais próximos 49
2.3.7 Estimador generalizado dos k-vizinhos mais próximos 50
2.4 Análise discriminante não paramétrica 50
2.4.1 Regras de classificação 50

3 Exemplos ilustrativos 55
3.1 Descrição dos exemplos 55
3.1.1 Mulheres portadoras do gene da hemofilia 57
3.1.2 Dados Iris 61
3.1.3 Dados "Grupo" 62
3.1.4 Dados "Outro" 65
3.2 Aspectos computacionais 68
3.3 Experiências de simulação 70
3.3.1 Análise dos resultados 72

4 Conclusões 81

A Dados dos exemplos ilustrativos 83

B Programas 87

C Referências Bibliográficas 93

Lista de Quadros

3.1 Hemofilia - Resultados considerando iguais probabilidades a priori 59

3.2 Hemofilia - Resultados após introdução das probabilidades a priori estimadas 60

3.3 Resultados do exemplo da NSC 64

3.4 Resultados do exemplo com 'outliers' 67

3.5 Taxas de erro aparentes, por grupo e totais, para n_1=n_2=50 76

3.6 Taxas de erro de validação cruzada, por grupo e totais, para n_1=n_2=50 . . . 76

3.7 Taxas de erro aparentes, por grupo e totais, para n_1=25, n_2=75 77

3.8 Taxas de erro de validação cruzada, por grupo e totais, para n_1=25, n_2=75 . 77

3.9 Taxas de erro aparentes, por grupo e totais, para n_1=100, n_2=100 78

3.10 Taxas de erro de validação cruzada, por grupo e totais, para n_1=100, n_2=100 78

Lista de Figuras

3.1 Dados bidimensionais do exemplo das mulheres portadoras do gene da hemofilia 58
3.2 Função de densidade estimada do grupo A - Mulheres não portadoras do gene da hemofilia 58
3.3 Função de densidade estimada do grupo B - Mulheres portadoras do gene da hemofilia 59
3.4 Dados do exemplo Iris 61
3.5 Observações bidimensionais do exemplo da NSC 62
3.6 Função de densidade estimada para o grupo A, do exemplo da NSC . . . 63
3.7 Função de densidade estimada para o grupo B, do exemplo da NSC . . . 63
3.8 Dados bidimensionais do exemplo da Normal com "outliers" 65
3.9 Função de densidade estimada para o grupo A, do exemplo com *"outliers"* 66
3.10 Função de densidade estimada para o grupo B, do exemplo com *"outliers"* 66
3.11 Rácio das taxas de erro: e_a/e_{opt} 79
3.12 Rácio das taxas de erro: e_c/e_{opt} 80

Capítulo 1

Introdução

O objectivo da análise discriminante é obter uma regra de classificação que permita classificar um indivíduo num de vários grupos possíveis. Essa regra é encontrada, com base num conjunto de p variáveis, a partir de indivíduos já classificados.

A análise discriminante tem aplicação em áreas muito diversas, fora das estatísticas tradicionais, podendo-se destacar a medicina, a biologia, a arqueologia, a engenharia e as ciências sociais. Como exemplos ilustrativos da aplicação destas técnicas considerem-se os seguintes: i)-Arqueologia: classificação de um esqueleto de uma espécie arqueológica numa de duas raças; ii)-Medicina: a análise discriminante revela-se particularmente útil no diagnóstico de várias doenças como, por exemplo, nos casos em que é perigoso sujeitar o indivíduo a determinados exames, utilizando-se as características do indivíduo, exames clínicos e testes laboratoriais para prever se esse indivíduo sofre ou não de determinada doença; iii)-Metereologia, Geologia, Agricultura, etc.: reconhecimento automático de imagens enviadas pelos satélites.

Muitos outros exemplos poderão ser encontrados na extensa bibliografia existente sobre análise discriminante.[1]

Embora a teoria da análise multivariada tenha o seu início nos anos de 1930, as razões que levaram ao desenvolvimento de métodos estatísticos para classificação estão intimamente ligadas às facilidades computacionais, que permitem a análise de grandes

[1] McLachlan (1992) dá-nos alguns exemplos dos principais resultados, principalmente dos desenvolvimentos mais recentes e indicação de referências adicionais.

conjuntos de dados e a automatização dos processos repetitivos para classificação. Também o facto de ser uma forma objectiva de classificar um indivíduo e de se poder conhecer qual o desempenho da regra de classificação têm grande importância.

Para encontrar uma regra de classificação é habitual assumir que a função geradora dos dados na população é normal multivariada e, como os parâmetros da população são desconhecidos, utilizam-se as estimativas amostrais desses parâmetros. Uma forma de libertar a análise discriminante desta hipótese rígida acerca da distribuição dos dados observados é a abordagem não paramétrica que tem hipóteses menos rígidas acerca dessa distribuição pois embora se assuma que a distribuição tem uma função de densidade f, permite-se que os dados falem por si próprios na sua determinação.

Outra forma é a utilização de métodos robustos em que, embora ainda se assuma um modelo paramétrico, há uma certa tolerância em relação a desvios desse modelo. Não sendo objectivo deste trabalho o estudo destes métodos, chama-se a atenção para o trabalho recentemente desenvolvido nesta área por Pires, A. M. (1995).

A abordagem não paramétrica para obter uma regra de classificação é muito importante quando as hipóteses em que os outros métodos se baseiam não se revelam adequadas. Para utilizar esta abordagem é necessário decidir qual o método de estimação da densidade que vai ser utilizado e especificar os parâmetros de alisamento.

O presente trabalho tem como objectivos gerais:

(a) fazer uma exposição sumária dos métodos não paramétricos de estimação de densidade mais populares para o caso univariado e da sua generalização ao caso multivariado;

(b) apresentar a discriminação não paramétrica e avaliar o seu desempenho, comparando-o com o desempenho dos métodos clássicos.

Neste primeiro capítulo faz-se uma breve apresentação de alguns métodos tradicionais de classificação, dos critérios de decisão existentes para obtenção de regras discriminantes e da forma de avaliar essas regras através das taxas de erro e suas estimativas.

No segundo capítulo são apresentados os principais estimadores não paramétricos da função de densidade de probabilidade, as suas propriedades e a forma de escolher o parâmetro de alisamento mais adequado. O capítulo termina com algumas regras de classificação obtidas no contexto das estimativas de densidade não paramétrica.

O terceiro capítulo começa com a aplicação de quatro das regras discriminantes a um conjunto de exemplos ilustrativos (discriminante linear, discriminante quadrática, discriminante obtida pelo método do produto do núcleo e discriminante obtida pelo método dos k-vizinhos-mais próximos). Segue-se um pequeno estudo de simulação destinado a averiguar o comportamento dos quatro métodos acima referidos quando aplicados a um conjunto de distribuições, representativas de situações mais ou menos reais, sendo a comparação do desempenho dos vários métodos feita através de estimativas das taxas de erro. No último capítulo encontram-se as principais conclusões.

1.1 Notação, formulação e critérios de decisão

1.1.1 Notação e formulação

O problema de encontrar um esquema de classificação a partir de amostras de objectos já classificados e transformá-lo numa regra prática de classificação pode ser formulado da seguinte forma:

- Dado um indivíduo (ou objecto) pertencente a um de g grupos diferentes G_i $(i = 1, ..., g)$ de uma população $\mathcal{P}$, pretende-se afectar o indivíduo a um desses grupos com base em p características observáveis, $\mathbf{x} = \{x_1, ..., x_p\}$, associadas ao indivíduo.

A regra de classificação define uma partição do espaço $\mathcal{R}^p$ em regiões Ω_i, $(i = 1, ..., g)$, que correspondem ao número de grupos. Um indivíduo é classificado como pertencendo ao grupo G_k se a representação do vector $\mathbf{x}$ correspondente ficar na região Ω_k.

Quando se procura uma regra de classificação podemos distinguir quatro situações diferentes:

1- A distribuição de $\mathbf{x}$ (em cada grupo, $f_i(\mathbf{x}), i = 1, ..., g$) é completamente conhecida;

2- A distribuição de $\mathbf{x}$ é conhecida excepto para alguns parâmetros;

3- A distribuição de $\mathbf{x}$ é parcialmente conhecida;

4- A distribuição de $\mathbf{x}$ é totalmente desconhecida;

Para amostras grandes pode-se ignorar a variância das estimativas e, nesse caso, as situações 1 e 2 são equivalentes e os métodos paramétricos podem ser adequados.

Na situação 3 pode ser adequada uma abordagem intermédia entre a paramétrica e a não paramétrica, que consiste em admitir uma forma paramétrica para o rácio $f_i(\mathbf{x})/f_j(\mathbf{x})$. Quando admitimos que o logaritmo deste rácio é linear somos conduzidos à regra discriminante logística.

Na situação 4 parece mais adequado estimar a densidade condicional dos grupos por um dos métodos não paramétricos.

1.1.2 Critérios de decisão

Conhecida a forma da distribuição pode-se definir a regra de classificação ideal para o problema em análise, que nos permita estimar a probabilidade de que um $\mathbf{x}$ observado pertença ao grupo G_i.

Probabilidades *a priori*

Em alguns casos é possível ter informação adicional, empírica ou com base em resultados de experiências anteriores, de tipo probabilístico, habitualmente designada por *probabilidades a priori*, $\pi_i = P(G_i)$, que são independentes de $\mathbf{x}$, e conhecidas antes de qualquer observação.

Quando não existe mais informação para além das probabilidades *a priori* é admissível a seguinte regra de classificação, que minimiza a probabilidade de cometer um erro:

- Classifica-se um indivíduo no grupo G_k se

$$\pi_k > \pi_i \quad \forall i \neq k.$$

Neste caso é irrelevante a observação de $\mathbf{x}$.

Máxima verosimilhança

No caso de não haver informação sobre as probabilidades *a priori*, nem sobre os custos de uma classificação incorrecta, uma forma intuitiva de obter uma regra de classificação é escolher o grupo G_i que *maximiza a função de verosimilhança* para a observação $\mathbf{x}$. Este critério conduz à seguinte região de classificação:

$$\mathbf{x} \in G_i \quad \text{se} \quad f_i(\mathbf{x}) > f_j(\mathbf{x}) \quad \forall i \neq j \quad i,j = 1,...,g.$$

Minimização da probabilidade total de classificação incorrecta

Se tivermos uma regra de classificação tal que se classifica um indivíduo no grupo G_i se $\mathbf{x}$ pertence a Ω_i, a probabilidade do indivíduo pertencente de facto ao grupo G_i ser mal classificado no grupo G_j é dada por

$$P(j \mid i) = \int_{\Omega_j} f_i(\mathbf{x})\ d\mathbf{x}$$

sendo a probabilidade total de má classificação dada por

$$P(\Omega, f) = \sum_{i=1}^{g} pr\,(\text{classificar erradamente } \mathbf{x} \text{ em } G_i)\,.$$

Se $g = 2$

$$P(\Omega, f) = P(1 \mid 2)\,\pi_2 + P(2 \mid 1)\,\pi_1.$$

A regra de classificação que se obtém para *minimizar a probabilidade total de classificação incorrecta*, devida a Welch (1939), é a seguinte:

$$\mathbf{x} \in G_i \ \text{ se } \ \pi_i f_i(\mathbf{x}) > \pi_j f_j(\mathbf{x}) \quad \forall i \neq j \quad i, j = 1, ..., g. \tag{1.1}$$

Minimização do custo total de classificação incorrecta

Os custos de classificação incorrecta de um indivíduo do grupo G_i no grupo G_j podem ser diferentes consoante a situação em análise, podendo ser formalizados utilizando uma matriz de custos com elemento genérico $C_{ij} = C(i|j)$, que representa o custo de classificar incorrectamente no grupo G_i um indivíduo pertencente ao grupo G_j. Como é evidente $C(i|i) = 0$.

Se $\mathbf{x} \in G_i$, o custo esperado por classificação incorrecta no grupo G_j é dado pela seguinte expressão:

$$C(i) = \sum_{j=1}^{g} C(j \mid i) \int_{\Omega_j} f_i(\mathbf{x})\, d\mathbf{x} = \sum_{j=1}^{g} C(j \mid i)\, P(j \mid i)$$

onde $P(j \mid i)$ é a probabilidade de classificar no grupo G_j um indivíduo do grupo G_i, sendo o custo total esperado dado por

$$C_T = \sum_{i=1}^{g} \pi_i C(i).$$

Este custo é minimizado quando Ω_i é escolhido tal que[2]

$$\mathbf{x} \in \Omega_i \text{ se } \sum_{k=1}^{g} C(i|k)\,\pi_k f_k(\mathbf{x}) < \sum_{k=1}^{g} C(j|k)\,\pi_k f_k(\mathbf{x}) \quad \forall j \neq i \quad i,j = 1,...,g$$

Para $g = 2$ obtem-se (note-se que $C_{11} = C_{22} = 0$)

$$\mathbf{x} \in \Omega_1 \text{ se } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2 C_{21}}{\pi_1 C_{12}}.$$

Esta regra de decisão é conhecida por *regra de decisão do custo mínimo de Bayes.*[3]

Se os custos forem iguais esta regra é equivalente à regra (1.1). Para $g = 2$, em termos matemáticos, podemos introduzir os custos na regra (1.1) da seguinte forma:

$$\pi_1^* = \frac{C(2 \mid 1)\,\pi_1}{C(2 \mid 1)\,\pi_1 + C(1 \mid 2)\,\pi_2}$$

$$\pi_2^* = 1 - \pi_1^*.$$

As probabilidades *a priori* são assim transformadas pela razão dos custos.[4] Esta transformação para mais de dois grupos só é possível se os custos forem dependentes apenas dos grupos de origem, isto é, se $C(j \mid i) = C(\cdot \mid i), \forall j \neq i$. Neste caso teremos

$$\pi_i^* = \frac{\pi_i C(\cdot \mid i)}{\sum_i \pi_i C(\cdot \mid i)}.$$

Maximização das probabilidades *a posteriori*

Se tivermos um vector $\mathbf{x}$ de observações do indivíduo a ser classificado, podem-se comparar as probabilidades de pertencer a um grupo condicionais a $\mathbf{x}$.

[2]Hand (1981), cap. 1, pag.6

[3]Anderson, (1958), cap.6

[4]Para uma informação mais detalhada ver Pires (1995)

Pelo teorema de Bayes a *probabilidade a posteriori* do grupo G_i é dada por

$$P(G_i \mid \mathbf{x}) = \frac{\pi_i f_i(\mathbf{x})}{\sum_{i=1}^{g} \pi_i f_i(\mathbf{x})}.$$

A regra de classificação conhecida por *regra do erro mínimo de Bayes*, segundo Hand (1981), é a seguinte

$$\mathbf{x} \in \Omega_i \quad \text{se} \quad P(G_i \mid \mathbf{x}) > P(G_j \mid \mathbf{x}) \quad \forall i \neq j \quad i,j = 1,...,g.$$

Por aplicação do teorema de Bayes podemos reformular esta regra:

$$\mathbf{x} \in \Omega_i \quad se \quad \pi_i f_i(\mathbf{x}) > \pi_j f_j(\mathbf{x}) \quad \forall i \neq j \quad i,j = 1,...,g.$$

a qual coincide com (1.1)

Classificação minimax

Uma regra que minimize as probabilidades totais pode conduzir a probabilidades de má classificação demasiado elevadas para um determinado grupo G_i. Isso acontece geralmente no caso em que a probabilidade *a priori*, π_i, desse grupo é pequena. Nessa situação podemos utilizar a regra de classificação minimax, que classifica $\mathbf{x}$ de forma a minimizar a maior probabilidade individual de classificação incorrecta, isto é, no caso de $g = 2$ minimiza o maior entre $P(2|1)$ e $P(1|2)$.[5]

A regra de classificação será então, classificar $\mathbf{x}$ no grupo G_1 se

$$f_1(\mathbf{x})/f_2(\mathbf{x}) > c$$

onde c é tal que $P(2|1) = P(1|2)$.

Para mais de dois grupos a solução[6] é dada por

$$\mathbf{x} \in \Omega_i \text{ se } \tilde{\pi}_i f_i(\mathbf{x}) > \tilde{\pi}_j f_j(\mathbf{x}) \quad \forall i \neq j \quad i,j = 1,...,g$$

onde $\tilde{\pi}_i$ são probabilidades *a priori* fictícias, calculadas de forma a obter probabilidades de classificação incorrecta iguais para todos os grupos.

[5] Seber (1984), cap. 6, pag.216

[6] Pires (1995), cap. 1, pag. 18

Maximização da separação entre os grupos

Este critério é também conhecido por critério de Fisher e impõe como restrição que a fronteira entre duas regiões seja linear. Fisher definiu a separação entre dois grupos numa direcção particular, como sendo a distância entre as médias dos dois grupos nessa direcção, estandardizada pela variância dentro de cada grupo na direcção especificada. Na pesquisa da melhor direcção de separação entre os grupos pretende-se encontrar $\mathbf{v}$ tal que $\left(\mathbf{v}^T\bar{\mathbf{x}}_1-\mathbf{v}^T\bar{\mathbf{x}}_2\right)$ seja maximizada relativamente ao desvio padrão $\sqrt{\mathbf{v}^T\mathbf{S}\,\mathbf{v}}$ nessa direcção, onde $\bar{\mathbf{x}}_i$ é a média do grupo G_i, $(i=1,2)$ e $\mathbf{S}$ é a matriz de covariâncias, que se assume comum para os dois grupos.

O que se pretende maximizar é assim:

$$q=\frac{\left(\mathbf{v}^T\bar{\mathbf{x}}_1-\mathbf{v}^T\bar{\mathbf{x}}_2\right)}{\sqrt{\mathbf{v}^T\mathbf{S}\,\mathbf{v}}}.$$

Diferenciando a expressão acima em ordem a $\mathbf{v}$ e igualando a zero, temos[7]

$$\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2=\frac{\hat{\mathbf{v}}^T\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\mathbf{S}\hat{\mathbf{v}}}{\hat{\mathbf{v}}^T\mathbf{S}\,\hat{\mathbf{v}}}\Longrightarrow\hat{\mathbf{v}}\propto\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right).$$

Esta é a abordagem clássica da análise discriminante. Se os dois grupos tiverem distribuição normal com matriz de covariâncias igual a solução de Fisher é assimptoticamente equivalente à solução óptima de Bayes.

A separação entre os grupos é assim definida pela equação $q\left(\mathbf{x}\right)=\mathbf{v}_0+\mathbf{v}^T\mathbf{x}=0$ onde $\mathbf{v}_0=-\frac{1}{2}\mathbf{v}^T\left(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2\right)$ é o ponto médio entre as projecções das médias dos grupos e a regra de classificação será:

$$\begin{aligned}\mathbf{x}\in\Omega_1 \quad &\text{se} \quad q\left(\mathbf{x}\right)>0\\ \mathbf{x}\in\Omega_2 \quad &\text{se} \quad q\left(\mathbf{x}\right)<0\end{aligned}.$$

Uma das objecções à utilização do critério vem da impossibilidade de ter em conta as probabilidades *a priori* e/ou os custos a não ser que se admita uma forma para a densidade da população unidimensional projectada. Outra questão frequentemente levantada, por ser considerada demasiado restritiva, é a hipótese da igualdade das matrizes de covariâncias para os dois grupos.

[7]Hand (1981), cap. 4, pag.83

As vantagens apontadas à utilização deste critério são a simplicidade de análise dos resultados e a sua vasta aplicabilidade, principalmente quando a dimensão de $\mathbf{x}$ é grande.

Na prática raramente se conhecem as distribuições condicionais dos grupos, sendo por isso necessário estimá-las a partir de um conjunto de indivíduos de que se conhece a classificação. Este conjunto é conhecido por amostra de treino e é a partir dela que se estimam os parâmetros necessários para obter a regra de classificação ou se estima a densidade de probabilidade de cada um dos grupos.

Existem dois tipos de amostragem habitualmente utilizados para obter a amostra de treino: amostragem mista ou conjunta e a amostragem condicional ou separada.

Na primeira, são escolhidos aleatoriamente n indivíduos do total da população $\mathcal{P}$, formada pela mistura dos g grupos possíveis, sendo $n = n_1 + ... + n_g$. Desta forma os n_i $(i = 1, ..., g)$ são variáveis aleatórias [Day e Kerridge (1967)] e a estimação das probabilidades *a priori* pode ser feita com base no tamanho relativo das amostras dos grupos: $\hat{\pi}_i = \frac{n_i}{n}$.

Na segunda, o número de observações n_i é fixado à partida, sendo os indivíduos escolhidos aleatoriamente dentro de cada um dos grupos G_i, $(i = 1, ..., g)$ [Anderson (1972), Prentice e Pyke (1979)]. Note-se que por este tipo de amostragem ser condicional ao grupo G_i, não é possível obter directamente estimativas das probabilidades *a priori*, π_i, para os grupos.[8]

Os custos quando não são conhecidos levam à impossibilidade da utilização da regra de decisão do custo mínimo de Bayes.

1.2 Discriminação com modelos normais

É frequente admitir algumas hipóteses sobre a forma das distribuições condicionais aos grupos como, por exemplo, admitir que elas pertencem a uma determinada família cujos parâmetros são desconhecidos e utilizar a amostra de treino para estimar esses parâmetros.

[8] McLachlan (1992), pag. 11 e pag. 31

A distribuição normal multivariada e a sua generalização de misturas de normais são as famílias mais populares. A escolha da distribuição normal tem a ver com a boa aproximação que esta dá a muitos fenómenos que ocorrem na natureza e à grande quantidade de teoria estatística existente baseada nela. A escolha da mistura de normais permite introduzir flexibilidade adicional ao estimador, apesar de introduzir um custo adicional por necessitar estimar mais parâmetros, podendo ser uma alternativa realista quando os dados mostram claramente não terem distribuição normal.

Conhecida a forma da distribuição pode-se definir a regra de classificação ideal para o problema em análise, que nos permita estimar a probabilidade de que um $\mathbf{x}$ observado pertença ao grupo G_i .

1.2.1 Modelo homocedástico

Para um modelo normal homocedástico, isto é, um modelo em que as matrizes de covariâncias dos grupos são todas iguais ($\Sigma_i = \Sigma$) e assumindo que a distribuição condicional do grupo G_i é $\mathbf{X}|G_i \sim N_p(\mu_i, \Sigma)$, para $i = 1, ..., g$, com

$$f_i(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i)\right\}$$

a regra de decisão, que será sempre definida a partir de $f_i(\mathbf{x}) / f_j(\mathbf{x}) > k_{ij}$, é linear:

$$(\mu_i - \mu_j)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i + \mu_j) > \log\ k_{ij}$$

onde k_{ij} é escolhido de acordo com o critério de decisão adoptado.

Se o critério de decisão for o critério de máxima verosimilhança então $k_{ij} = 1$.

Se se pretender maximizar as probabilidades *a posteriori* ou minimizar a probabilidade total de classificação incorrecta então $k_{ij} = \pi_i / \pi_j$, ou seja, será igual ao rácio das probabilidades *a priori*. Quando se pretende utilizar um critério de classificação minimax $k_{ij} = \tilde{\pi}_i / \tilde{\pi}_j$. Para minimizar o custo total de classificação incorrecta, para $g = 2$, $k_{ij} = \pi_2 C_{21} / \pi_1 C_{12}$.

Na prática os parâmetros da população μ_i e Σ são geralmente desconhecidos e utilizam-se habitualmente as estimativas de máxima verosimilhança destes parâmetros.

Dada uma amostra $\mathbf{X} = \{\mathrm{X}_1, ..., \mathrm{X}_n\}$ para o total de indivíduos, se associarmos à realização $\mathbf{x}_j = \{x_{j1}, ..., x_{jn}\}$ para um indivíduo, um vector $\mathbf{z}_j$, g-dimensional, que nos indique o grupo de origem, tal que a i-ésima componente de $\mathbf{z}_j$ seja

$$z_{ij} = 1 \quad se \quad \mathbf{x}_j \in G_i$$
$$z_{ij} = 0 \quad se \quad \mathbf{x}_j \notin G_i$$

estas estimativas são:

$$\bar{\mathbf{x}}_i = \sum_{j=1}^{n} z_{ij}\mathbf{x}_j / n_i$$

$$\hat{\Sigma} = \sum_{i=1}^{g} \frac{n_i}{n} \hat{\Sigma}_i = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \frac{(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T}{n}$$

para $i = 1, ..., g$ e $n_i = \sum_{j=1}^{n} z_{ij}$. Pode-se também utilizar o estimador centrado de Σ,

$$\mathbf{S} = \frac{n\hat{\Sigma}}{n-g}.$$

1.2.2 Modelo heterocedástico

Para um modelo normal heterocedástico a distribuição condicional do grupo G_i é $\mathbf{X} \mid G_i \sim N_p(\mu_i, \Sigma_i)$, para $i = 1, ..., g$, e a densidade condicional do grupo é

$$f_i(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right\}.$$

A regra discriminante define fronteiras quadráticas entre as regiões Ω_i e é por isso chamada *discriminante quadrática*, podendo a condição $\frac{f_i(\mathbf{x})}{f_j(\mathbf{x})} > k_{ij}$, escrever-se na seguinte forma:

$$\log \frac{|\Sigma_j|}{|\Sigma_i|} - \mu_i^T \Sigma_i^{-1} \mu_i + \mu_j^T \Sigma_j^{-1} \mu_j - \mathbf{x}^T \left(\Sigma_i^{-1} - \Sigma_j^{-1}\right) \mathbf{x} + 2\mathbf{x}^T \left(\Sigma_i^{-1} \mu_i - \Sigma_j^{-1} \mu_j\right) > 2 \log k_{ij}.$$

As estimativas de máxima verosimilhança de Σ_i e μ_i são:

$$\bar{\mathbf{x}}_i = \sum_{j=1}^{n} z_{ij}\mathbf{x}_j / n_i, \tag{1.2}$$

$$\hat{\Sigma}_i = \sum_{j=1}^{n} z_{ij} \frac{(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T}{n_i}$$

para $i = 1, ..., g$ e $n_i = \sum_{j=1}^{n} z_{ij}$. O estimador centrado de Σ_i é

$$\mathbf{S}_i = \frac{n_i \hat{\Sigma}_i}{n_i - 1}.$$

1.3 Taxas de erro das regras de classificação

Para avaliar o desempenho de uma regra de classificação, é necessário conhecer a sua capacidade de bem ou mal classificar. Há várias probabilidades de erro associadas a essa regra que são chamadas as taxas de erro e que devem ser consideradas.[9]

Quase todas as regras de classificação tomam a seguinte forma: classifica-se um indivíduo no grupo G_i se $\frac{f_i(\mathbf{x})}{f_j(\mathbf{x})} > c$. A região óptima será a que optimiza um determinado critério, tomando a seguinte forma: $\Omega_{0i} = \left\{\mathbf{x} : \frac{f_i(\mathbf{x})}{f_j(\mathbf{x})} > c\right\}$.

1.3.1 Taxa de erro óptima

A taxa de erro óptima para o grupo G_i é:

$$e_{i,opt} = \int_{\bar{\Omega}_{0i}} f_i(\mathbf{x})\, d\mathbf{x}$$

sendo a taxa de erro óptima total para $g = 2$

$$e_{opt} = \pi_1 e_{1,opt} + \pi_2 e_{2,opt}.$$

1.3.2 Taxa de erro actual

A região óptima estimada é $\hat{\Omega}_{0i}$ e a taxa de erro actual para o grupo G_i é dada por:

$$e_{i,act} = \int_{\bar{\hat{\Omega}}_{0i}} f_i(\mathbf{x})\, dx$$

sendo a taxa de erro actual total, para $g = 2$,

$$e_{act} = \pi_1 e_{1,act} + \pi_2 e_{2,act}.$$

Note-se que, por definição, $e_{opt} \leq e_{act}$.

[9] Ver Hills (1966) para uma abordagem teórica e as definições em Lachenbruch (1975; pag.30)

1.3.3 Taxa de erro actual esperada

A taxa de erro actual esperada, $E[e_{i,act}]$, é obtida pelo valor esperado no espaço amostral das taxas anteriores e para $g = 2$

$$E[e_{act}] = \pi_1 E[e_{1,act}] + \pi_2 E[e_{2,act}].$$

1.4 Estimação das taxas de erro

Para estimar a taxa de erro actual, são utilizadas as taxas de erro descritas em seguida.

1.4.1 Taxas de erro de substituição

$$\hat{e}_{i,act} = \int_{\hat{\Omega}_{0i}} \hat{f}_i(\mathbf{x})\, dx$$

em que se substituiu f_i por $\hat{f}_i$ obtida utilizando as estimativas dos parâmetros no caso de utilização de um método paramétrico ou uma estimativa não paramétrica. Para $g = 2$:

$$\hat{e}_{act} = \pi_1 \hat{e}_{1,act} + \pi_2 \hat{e}_{2,act}.$$

Alguns autores chamam a estas estimativas as taxas de erro aparentes [ex: Hills (1966), Glick (1973), Goldstein e Dillon (1978)].

1.4.2 Taxas de erro aparentes

As taxas de erro aparentes são obtidas por reclassificação da amostra, aplicando a regra discriminante obtida com essa amostra.

A taxa de erro aparente do grupo G_i é a seguinte:

$$e_{i,ap} = m_i/n_i$$

e para $g = 2$

$$e_{ap} = \pi_1 e_{1,ap} + \pi_2 e_{2,ap}$$

onde m_i é o número de observações classificadas incorrectamente ao aplicar a regra discriminante.

No caso de amostragem mista, se as probabilidades *a priori* forem desconhecidas podemos estimá-las:

$$\hat{\pi}_i = \frac{n_i}{n_1 + n_2} \tag{1.3}$$

e

$$\hat{e}_{ap} = \hat{\pi}_1 \hat{e}_{1,ap} + \hat{\pi}_2 \hat{e}_{2,ap} = \frac{n_1}{n_1 + n_2}\frac{m_1}{n_1} + \frac{n_2}{n_1 + n_2}\frac{m_2}{n_2} = \frac{m_1 + m_2}{n_1 + n_2}.$$

As taxas de erro aparentes são também conhecidas por taxas de ressubstituição.

1.4.3 Taxas de validação cruzada

Este método foi proposto por Lachenbruch e Mickey (1968) e a forma de obter estas taxas é a seguinte:

- Determina-se a regra discriminante utilizando $n - k$ observações da amostra e depois utiliza-se essa regra para classificar as k observações omitidas. Quando $k = 1$ este método é conhecido por *"leaving-one-out"*.

O erro para o grupo G_i é obtido por cálculo da proporção de observações mal classificadas:

$$e_{i,c} = \frac{a_i}{n_i}$$

e, para $g = 2$, $e_c = \pi_1 e_{1,c} + \pi_2 e_{2,c}$, que, no caso da amostragem mista, pode ser estimado por $\hat{e}_c = (a_1 + a_2)/(n_1 + n_2)$.

Estas estimativas são também conhecidas por *"jacknife"*, embora a denominação seja incorrecta.

1.4.4 Taxas obtidas por *"bootstrap"*

Esta estimativa foi proposta por Efron (1979, 1981), que sugeriu a técnica de *"bootstrap"* para estimar o enviesamento de $e_{i,ap}$.[10]

Para calcular estas taxas, a partir das n_i observações iniciais, é extraída uma nova amostra de tamanho n_i $(i = 1, 2)$ com reposição. Com esta nova amostra obtém-se uma nova regra discriminante, utilizando o mesmo método que tinha sido utilizado anteriormente. Suponha-se que com esta nova regra se obtêm m_i^* e m_i^{**} observações mal classificadas, respectivamente da nova mostra e da amostra original. Seja $d_i = (m_i^{**} - m_i^*)/n_i$ em cada réplica e $\bar{d}$ a média dos d_i obtidos num largo número de réplicas.

As estimativas *"bootstrap"* são:

$$e_{i,boot} = \frac{m_i}{n_i} + \bar{d}_i \tag{1.4}$$

onde m_i/n_i são as taxas de erro aparentes e $e_{boot} = \pi_1 e_{1,boot} + \pi_2 e_{2,boot}$.

Note-se que, nos métodos paramétricos, com estimadores adequados dos parâmetros e para grandes amostras[11]

$$E\left[\hat{e}_{act}\right] \leq e_{opt} \leq E\left[e_{act}\right].$$

Espera-se que $\hat{e}_{act}$ subestime e_{act}, podendo o enviesamento ser grande, principalmente se $f_i(\mathbf{x})$ não estiver correctamente especificada. Também $e_{i,ap}$ e e_{ap} tendem a subestimar as correspondentes taxas actuais e só devem ser usadas quando as amostras são grandes. Para amostras pequenas é preferível utilizar $e_{i,c}$ e $e_{i,boot}$ porque são estimativas quase sem enviesamento de $e_{i,act}$.

1.5 Métodos paramétricos parciais

As regras de classificação desenvolvidas até agora dependem apenas do rácio das funções de densidade. Por essa razão podemos simplesmente modelar este rácio sem especificar

[10] Ver também McLachlan (1980)

[11] Seber (1984) - pag.284

as funções de densidade individuais $f_i(\mathbf{x})$.

1.5.1 Discriminante logística

No modelo logístico assume-se que, para $g = 2$,

$$\log\left[\frac{f_i(\mathbf{x})}{f_j(\mathbf{x})}\right] = \alpha + \beta^T\mathbf{x}$$

onde α e $\beta = (\beta_1, ..., \beta_p)$ são os $p+1$ parâmetros a estimar.

A regra discriminante é a seguinte:

- Classifica-se o indivíduo no grupo G_1 se $\alpha + \beta^T\mathbf{x} > \log k$.

Se $k = \pi_2/\pi_1$, estamos a maximizar as probabilidades *a posteriori* ou minimizar a probabilidade total de classificação incorrecta, e a regra discriminante será:

- Classifica-se o indivíduo no grupo G_1 se $\alpha_0 + \beta^T\mathbf{x} > 0$ onde $\alpha_0 = \alpha + \log\left[\frac{\pi_1}{\pi_2}\right]$.

Para definir o modelo logístico para $g > 2$, seja $\beta_i = (\beta_{1i}, ..., \beta_{pi})$, para $i = 1, ..., g-1$, um vector de p parâmetros. Sendo G_g o grupo base, o modelo logístico assume que

$$\log\left[\frac{f_i(\mathbf{x})}{f_g(\mathbf{x})}\right] = \alpha_0 + \beta_i^T\mathbf{x}.$$

Os parâmetros deste modelo podem ser estimados por máxima verosimilhança.

A discriminação logística é aplicável a uma grande variedade de distribuições[12] [ver Lachenbruch (1975), Anderson (1982) e também Anderson e Blair (1982)], nomeadamente i) distribuições normais multivariadas com matrizes de covariâncias iguais; ii) distribuições multivariadas discretas que sigam um modelo log-linear com iterações iguais em cada grupo; iii) distribuições conjuntas de variáveis discretas e contínuas que sigam i) ou ii), não necessariamente independentes; iv) versões truncadas das anteriores; v) variáveis multivariadas dicotómicas; vi) versões das anteriores onde $\mathbf{x}$ é substituído por uma função vectorial de $\mathbf{x}$.

[12] Pode-se encontrar um grande número de referências sobre o assunto em McLachlan (1992), cap.8, pag.268

São possíveis outras escolhas, para além da transformação logística, que relacionem $\mathbf{x}$ com a sua probabilidade *a posteriori*. Albert e Anderson (1981) consideraram o conceito de discriminação *probit*:

$$probit\left\{\tau_1(\mathbf{x})\right\} = \Phi^{-1}\left\{\tau_1(\mathbf{x})\right\}$$

onde $\Phi(\cdot)$ é a distribuição normal estandardizada. Na prática é quase impossível distinguir entre estes dois modelos, uma vez que

$$logit\left\{\tau_1(\mathbf{x})\right\} \cong c\,\Phi^{-1}\left\{\tau_1(\mathbf{x})\right\}.$$

Habitualmente prefere-se o modelo logístico devido às facilidades computacionais.

Uma abordagem mais geral do que a discriminação logística ou probit é a utilização da classe dos modelos aditivos generalizados proposta por Hastie e Tibshirani (1986, 1990). Nestes modelos a forma linear $\alpha_0 + \beta^T\mathbf{x}$ é generalizada a uma soma de funções alisadas $\sum_{j=1}^{p} s_j\left((\mathbf{x})_j\right)$ onde $s_j(\cdot)$ são funções não especificadas que são estimadas utilizando um alisamento gráfico ("scatter plot") num procedimento iterativo chamado "scoring algorithm" local.

1.5.2 Discriminante logística quadrática

Como vimos o modelo logístico não exige linearidade nas variáveis básicas, sendo possível incluir qualquer função específica delas [Anderson (1975 e 1982)]. As mais vulgares são as transformações logarítmicas, o quadrado e a raiz quadrada.

Se as variáveis são binárias e as interacções de primeira ordem na escala loglinear não são iguais em cada grupo, existe uma transformação que permite resolver o problema desde que as interacções de ordem mais elevada sejam as mesmas para todos os grupos.

Supondo que a distribuição condicional do grupo G_i é $\mathbf{X}\,|G_i \backsim N_p(\mu_i, \Sigma_i)$, para $i = 1, ..., g$, então

$$\log\left[\frac{f_i(\mathbf{x})}{f_g(\mathbf{x})}\right] = \alpha_{0i} + \beta_i^T\mathbf{x} + \mathbf{x}^T\Lambda_i\mathbf{x}$$

onde Λ_i é uma matriz simétrica $p \times p$.

Esta função é linear nos coeficientes α_{0i}, β_i e nos elementos distintos de Λ_i, mas tem $p+1+\frac{1}{2}p(p+1)$ parâmetros a estimar, o que dificulta a sua utilização para $p > 4$ ou 5.

Foram sugeridas algumas aproximações para a forma quadrática $\mathbf{x}^T\Lambda_i\mathbf{x}$, para situações em que isso acontece, sendo a mais simples

$$\Lambda_i \approx \lambda_{1j}\psi_{1i}\psi_{1i}^T$$

onde λ_{1j} é o maior valor próprio de Λ_i e ψ_{1i} $(i = 1, ..., g-1)$ é o correspondente vector próprio.

Capítulo 2

Métodos não paramétricos

2.1 Introdução

Quando nada se sabe sobre a distribuição condicional dos grupos, nem sequer se tem conhecimento teórico da sua forma paramétrica, podemos recorrer aos métodos não paramétricos.

Existem quatro tipos principais de estimadores não paramétricos da função de densidade de probabilidade: o histograma, o método do núcleo *(kernel)*, o método dos k-vizinhos mais próximos *(k-Nearest Neighbour ou k-NN)* e o método das séries.

Embora para a análise discriminante a estimação de densidades multivariadas seja mais importante, como os métodos multivariados são generalizações dos métodos univariados, começa-se por abordar estes últimos.

2.1.1 Propriedades estatísticas dos estimadores de densidade

Tal como no caso paramétrico é necessário conhecer as propriedades estatísticas dos estimadores de densidade não paramétrica, de forma a assegurar que estes têm as propriedades desejáveis.

Uma propriedade importante dos estimadores é a consistência. Diz-se que um estimador de densidade é fracamente consistente pontualmente para f se $\hat{f}(x)$ converge para $f(x)$ em probabilidade, para qualquer x, e fortemente consistente pontualmente

se a convergência for quase certa. Outros tipos de consistência dependem do critério de erro. Um critério de erro habitualmente utilizado é o *erro quadrático médio (MSE)* definido por

$$MSE_x\left(\hat{f}\right) = E\left\{\hat{f}(x) - f(x)\right\}^2. \tag{2.1}$$

Pelas propriedades da média e da variância, temos

$$MSE_x\left(\hat{f}\right) = \left\{E\left[\hat{f}(x)\right] - f(x)\right\}^2 + var\,\hat{f}(x) \tag{2.2}$$

ou seja, a soma do quadrado do enviesamento e da variância em x, onde $var\hat{f}(x) = E\left\{\hat{f}(x) - E[\hat{f}(x)]\right\}^2$ e o enviesamento $bias\,\hat{f}(x) = E[\hat{f}(x)] - f(x)$. Se $MSE_x \to 0$, para qualquer x, quando $n \to \infty$, diz-se que $\hat{f}$ é um estimador pontual de f consistente em média quadrática.

Outra medida habitualmente utilizada é o *erro quadrático integrado (ISE)* também conhecido por *norma* L_2:

$$\int \left\{\hat{f}(x) - f(x)\right\}^2 dx. \tag{2.3}$$

Um outro critério importante, a medida mais conhecida para a discrepância global [Rosenblatt (1956)], que mede o ajustamento da curva $\hat{f}$ a f, é o *erro quadrático médio integrado (MISE)* definido por

$$MISE\left(\hat{f}\right) = E\int \left\{\hat{f}(x) - f(x)\right\}^2 dx \tag{2.4}$$

que, por o integrando ser não negativo, se pode escrever nas seguintes formas alternativas:

$$\begin{aligned} MISE\left(\hat{f}\right) &= \int E\left\{\hat{f}(x) - f(x)\right\}^2 dx && (2.5)\\ &= \int MSE_x\left(\hat{f}\right) dx && (2.6)\\ &= \int \left\{E\left[\hat{f}(x)\right] - f(x)\right\}^2 dx + \int var\hat{f}(x)\,dx. && (2.7) \end{aligned}$$

Note-se que $MISE = E(ISE)$.

Uma objecção à utilização destes critérios na estimação da densidade não paramétrica é que o comportamento da cauda da distribuição torna-se pouco importante, pelo que podem surgir algumas peculiaridades nas caudas da esimativa da densidade.[1]

[1] Izenman (1991)

Uma alternativa a esta abordagem é a utilização do critério do *erro integrado absoluto (IAE)* também conhecido como *variação total* [2] ou *norma* L_1:

$$IAE = \int_{-\infty}^{+\infty} \left| \hat{f}(x) - f(x) \right| dx \tag{2.8}$$

que é invariante sob transformações monótonas e $0 \leq IAE \leq 2$. O valor esperado de (2.8) em todas as densidades $\hat{f}$ é o *erro integrado absoluto médio (MIAE).* A desvantagem deste critério é que, para obter resultados análogos ao da *norma* L_2, este é muito mais difícil e trabalhoso.

2.1.2 Histogramas

O histograma permite estimar a função de densidade de probabilidade pela proporção de pontos da amostra que caiem em cada célula.

A diferença entre um histograma de frequências e um histograma de densidade é que o último sofre uma transformação de forma a que o seu integral seja 1. Suponha-se que o suporte da função de densidade é $\Omega = [a, b]$. Faça-se uma partição deste intervalo em m rectângulos numa grelha, onde $a = t_1 < t_2 < ... < t_{m+1} = b$, e seja $B_k = [t_k,\ t_{k+1})$ $(k = 1, ..., m)$, o $k - ésimo$ rectângulo. Suponha-se que $t_{k+1} - t_k = h$ e seja v_k o número de X_i no mesmo rectângulo B_k $(\sum_k v_k = n)$.

O histograma é então definido por:

$$\hat{f}\ (x) = \frac{1}{nh} v_k$$

e diz-se ter largura do rectângulo, h, fixa.

A escolha da origem e da largura do rectângulo não é indiferente. A primeira pode originar interpretações diferentes, pois o número de pontos da amostra que cai em cada rectângulo vai variar de acordo com a origem escolhida podendo-se assim esconder, por exemplo, a ocorrência de um segundo pico (moda) ou dar uma leitura diferente da separação destas, enquanto que a segunda vai determinar a quantidade de alisamento inerente ao procedimento.

[2] Devroye (1987); Devroye e Gyõrfi (1985)

Podemos generalizar o histograma permitindo que a largura do rectângulo varie:

$$\hat{f}\ (x) = \frac{1}{nh_i} v_k$$

onde h_i é a largura do rectângulo que contem x.

Uma das desvantagens do histograma, que ocorre tanto no caso univariado como no multivariado, é a descontinuidade na fronteira de cada rectângulo, que causa um salto do nível deste para o rectângulo vizinho. Isto causa extrema dificuldade no caso de serem necessárias derivadas das estimativas.

Em qualquer caso o histograma requer a escolha da quantidade de alisamento desejada. Se h for demasiado pequeno, o histograma será demasiado irregular e terá uma grande variância; pelo contrário, se h for demasiado grande o histograma estará demasiado alisado e teremos um grande enviesamento. A selecção de h deve ter em conta o equilíbrio que se pretende entre a variância e o enviesamento, optimizando um determinado critério.

Embora a opção pelo criério a optimizar seja subjectiva e se deva ter presente que um estimador pode ser óptimo para um dado objectivo e péssimo para outro, um critério frequentemente utilizado é o *erro quadrático médio integrado (MISE)* (2.4) que, sob certas condições de f, quando $h_n \to 0$ e $nh_n \to \infty$,[3] converge para zero como foi demonstrado por Scott (1979) e Freedman e Diaconis (1981). Assimptoticamente o *MISE* é minimizado se

$$h_{n,opt} = [6/R\left(f'\right)]^{1/3} n^{-1/3}, \text{ onde } R(g) = \int_{-\infty}^{\infty} [g\,(x)]^2 dx.$$

A escolha de $h_{n,opt}$ óptimo depende do conhecimento da densidade f, através de $R\,(f)$. Scott (1979) propôs que, como referência, a escolha da largura dos rectângulos fosse

$$h_n = 3.49 s n^{-1/3} \tag{2.9}$$

onde s é o desvio padrão amostral. Embora a densidade Normal esteja na base da escolha de (2.9), o autor considera não ser uma hipótese tão forte como seria no caso

[3] Com o subscripto n em h_n pretende-se enfatizar a dependência da largura do rectângulo do tamanho da amostra.

paramétrico e que é útil para várias classes de densidades, embora leve a h_n's que são em geral demasiado elevados para dados não Gaussianos. Sugere que não se utilize directamente (2.9), mas que se escolha um h_n ligeiramente maior ou menor, propondo as distribuições *lognormal* e t como densidades de referência para modificar a regra da Normal, quando a distribuição não é simétrica ou tem caudas pesadas.

Freedman e Diaconis (1981b) propuseram $h_n = 2(IQR)n^{-1/3}$, uma medida mais robusta, onde IQR é a amplitude interquartil.

Outras regras práticas de escolha da largura da janela como, por exemplo, a validação cruzada ou "oversmooth bandwidth" podem ser encontradas em Scott (1992).

Um estimador relacionado com o histograma e que surgiu pela necessidade de resolver o problema da escolha da origem do rectângulo é o *ASH (Average Shifted Histogram),* de Scott e Thompson (1983) e Scott (1985b). Este histograma é construído tomando a média de vários histogramas com largura dos rectângulos igual, mas com origem dos mesmos diferente. Considerando um conjunto de m histogramas, $\hat{f}_1, \hat{f}_2, ..., \hat{f}_m$, cada um com largura do rectângulo h, e com origem dos rectângulos nos pontos

$$t_0 = 0, \frac{h}{m}, \frac{2h}{m}, ..., \frac{(m-1)h}{m}$$

o histograma *ASH naive* é definido por

$$\hat{f}(\cdot) = \hat{f}_{ASH}(\cdot) = \frac{1}{m}\sum_{i=1}^{m}\hat{f}_i(\cdot).$$

Outros estimadores relacionados são, por exemplo, o *Polígono de Frequências(FP)* de Scott (1985a), que é construído ligando os valores centrais dos rectângulos com linhas rectas, o *"histospline"* de Boneva, Kendall e Stefanov (1971), o *histograma ponderado* também conhecido como *aproximação tipo-polinomial de Bernstein,* de Vitale (1975) e Gawronski e Stadtmuller (1980)

2.1.3 Estimadores do núcleo

Fix e Hodges (1951) foram quem primeiro propôs estimativas de densidade por procedimentos não paramétricos, com o intuito expresso de análise discriminante. São

eles os responsáveis pela introdução de dois métodos não-paramétricos populares para estimação de densidade: o estimador do núcleo e o do vizinho mais próximo.

O método do núcleo consiste em estimar uma função de densidade duma variável aleatória contínua, a partir da "densidade da amostra".

A função de densidade de probabilidade empírica é definida como sendo a derivada da função de distribuição cumulativa empírica: $f_n(x) = \frac{d}{d(x)} F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$, onde $\delta(t)$ é a função delta de Dirac, que não permite obter um estimador alisado da função de densidade.

Para densidades univariadas, a estimativa não paramétrica dada por este método é

$$\hat{f}\,(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2.10}$$

onde K é tal que:

$$\int_{-\infty}^{+\infty} K\,(x)\; dx = 1 \tag{2.11}$$

(habitualmente uma função simétrica como, por exemplo, a normal) e h é o parâmetro de alisamento, também chamado "largura da janela" ou "largura da banda" por alguns autores.

O estimador do núcleo pode ser considerado como a soma de "saltos" nas observações. A função núcleo, K, determina a forma dos saltos enquanto que o parâmetro de alisamento, h, determina a sua largura. No limite, quando $h \to 0$, a estimativa resulta numa colecção de "picos", $f_n(x)$, enquanto que se $h \to \infty$, todos os detalhes ficam ocultos. A escolha do parâmetro de alisamento é crucial, pelo que será abordada em maior detalhe mais tarde.

O estimador do núcleo originalmente sugerido por Fix e Hodges (1951), hoje conhecido como estimador "naive", tinha $h = 1$ e K era uma função de densidade uniforme numa vizinhança pré-estabelecida.

Propriedades das estimativas do núcleo:

Se $K\,(x) \geq 0$ e se verifica (2.11) então $\hat{f}\,(x)$ será também uma função de densidade e herdará todas as propriedades de continuidade e diferenciabilidade do núcleo K. Por

exemplo, se K é uma normal, $\hat{f}(x)$ será uma curva alisada com derivadas de todas as ordens.

Segundo alguns autores, a estimação da densidade pelo método do núcleo é resistente a valores extremos por definição, uma vez que $K\left(\frac{x-x_i}{h}\right)$ deve tornar-se mais pequena quando x_i se afasta de x.

As propriedades assimptóticas das estimativas do núcleo foram investigadas por diversos autores, nomeadamente Rosenblatt (1956, 1971), Parzen (1962), Wertz (1978), para o caso univariado e por Epanechnikov (1969) para o caso multivariado.

Considerando que as estimativas de densidade são construídas a partir de uma amostra $X_1, ..., X_n$ com distribuição f, assumindo que o núcleo K e as estimativas da densidade f satisfazem algumas condições de regularidade e que a largura da janela, h, depende de alguma forma do tamanho da amostra (h_n), Parzen estudou a consistência de f no ponto x, sob certas condições. Essas condições para o núcleo K, são

$$\int |K(t)|\, dt < \infty \tag{2.12}$$

$$\int K(t)\, dt = 1 \tag{2.13}$$

$$|tK(t)| \to 0 \text{ quando } |t| \to \infty \tag{2.14}$$

que são satisfeitas por quase todos os núcleos possíveis.

A largura da janela h_n deve satisfazer o seguinte:

$$h_n \to 0 \quad \text{e} \quad nh_n \to \infty \quad \text{quando} \quad n \to \infty$$

Nestas condições, se f for uma função contínua em x

$$\hat{f}(x) \to f(x) \quad \text{em probabilidade quando} \quad n \to \infty$$

As condições sobre h_n implicam que a janela se deve tornar mais pequena quando n aumenta e deve convergir para zero à taxa n^{-1}.

A consistência uniforme, em vez da pontual, foi também considerada por alguns autores, como Parzen (1962), Nadaraya (1965), Bertrand-Retali (1978) e Silverman (1978b).

Supondo que o núcleo K é finito, tem variância finita e satisfaz as condições (2.12) e (2.13), que f é uniformemente contínua em $(-\infty, \infty)$ e que $h_n \to 0$ e $nh_n (\log n)^{-1} \to \infty$ quando $n \to \infty$, Bertrand-Retali (1978) mostrou que $\sup_x \left|\hat{f}(x) - f(x)\right| \to 0$ quando $n \to \infty$.

Uma desvantagem que é apontada a este método é que há uma tendência para aparecer um ruído espúrio nas caudas devido à largura da janela ser fixa para toda a amostra, quando aplicado a dados cuja distribuição tenha caudas pesadas, existindo vários métodos adaptativos para lidar com esta situação.

Estimador "naïve"

O estimador "naïve" é dado pela escolha de um número pequeno h e fazendo

$$\hat{f}(x) = \frac{1}{2nh} [\text{n}^o \text{ de } X_1, ..., X_n \text{ que cai no intervalo } (x-h, x+h)].$$

Pode também escrever-se este estimador na mesma forma que o estimador do núcleo (2.10), onde K é uma função peso dada por

$$K(x) = \begin{cases} \frac{1}{2} & \text{se} \quad |x| < 1 \\ 0 & \text{outros casos} \end{cases}. \tag{2.15}$$

A estimativa é construída colocando uma "caixa" de largura $2h$ e altura $(2nh)^{-1}$ em cada observação e somando depois para obter a estimativa. É assim uma soma ponderada de pontos, onde o peso é $(2nh)^{-1}$ para um ponto no intervalo $(x-h, x+h)$ e 0 fora desse intervalo.

Com este estimador $\hat{f}(x)$ não é uma função contínua, apresentando saltos nos pontos $X_i \pm h$, tendo por isso uma aparência localmente irregular.

2.1.4 Estimador dos k-vizinhos mais próximos

Este método representa uma tentativa de adaptar a quantidade de alisamento à densidade "local" dos dados. A quantidade de alisamento é controlada por um inteiro k (habitualmente $k \approx n^{1/2}$).

Defina-se a distância $d(x,y)$ como $|x-y|$ e ordenem-se as distâncias de t aos pontos da amostra, $d_i(t)$, por ordem ascendente.

Supondo que a densidade em t é $f(t)$, numa amostra de tamanho n, espera-se que k observações caiam no intervalo $[t-d_k(t),\, t+d_k(t)]$ e pode-se obter uma estimativa da densidade em t a partir de $k = 2\,d_k(t)\,n\,\hat{f}\,(t)$.

A estimativa da densidade do método do k-ésimo vizinho mais próximo é então definida por:

$$\hat{f}\,(t) = \frac{k}{2nd_k\,(t)}.$$

Nas caudas da distribuição a distância $d_k(t)$ será mais larga que na parte principal reduzindo assim o subalisamento nas caudas.

O estimador dos vizinhos mais próximos não é uma curva alisada, pois embora a função $d_k(t)$ possa ser vista como contínua, a sua derivada terá uma descontinuidade em cada ponto da forma $\frac{1}{2}\left(X_{(j)} + X_{(j+k)}\right)$, onde $X_{(j)}$ são as estatísticas de ordem da amostra, donde $\hat{f}\,(t)$ terá derivadas descontínuas nos mesmos pontos que $d_k(t)$.

Para t menor que o menor ponto $d_k(t) = X_{(k)} - t$ e para $t > X_{(n)}$ tem-se $d_k(t) = t - X_{(n-k+1)}$ o que implica que $\int_{-\infty}^{+\infty} \hat{f}\,(t)\,dt$ é infinito e que as caudas caiem a uma taxa muito lenta. Por esse motivo o estimador do vizinho mais próximos não é aconselhável se necessitarmos estimar toda a densidade.

2.1.5 Estimador generalizado dos k-vizinhos mais próximos

O estimador generalizado dos k-vizinhos mais próximos é definido por :

$$\hat{f}\,(t) = \frac{1}{nd_k\,(t)} \sum_{i=1}^{n} K\left(\frac{t-X_i}{d_k\,(t)}\right). \tag{2.16}$$

Como se pode ver $\hat{f}\,(t)$ é o estimador do núcleo avaliado em t, em que o parâmetro de alisamento é $d_k(t)$. Embora o alisamento total dependa do inteiro k, o parâmetro de alisamento num ponto depende da densidade das observações perto desse ponto.

Tal como o estimador dos k-vizinhos mais próximos, este estimador será descontínuo em todos os pontos em que $d_k(t)$ tiver derivada descontínua. As propriedades da cauda e a integrabilidade vão depender da forma do núcleo K.

2.1.6 Estimador do núcleo adaptativo

O estimador do núcleo adaptativo ajusta a largura da janela de forma a que ela seja mais estreita nos locais em que a densidade é elevada e mais larga nos locais em que a densidade é baixa. Este procedimento é baseado na ideia de que, para funções de densidade com caudas longas se deve utilizar um núcleo mais largo nas regiões de baixa densidade. Como não se conhece a distribuição $f(x)$ é necessário começar por encontrar uma estimativa preliminar desta densidade, como por exemplo a que é obtida pelo método do núcleo fixo, $\tilde{f}(t)$ que satisfaça $\tilde{f}(X_i) > 0$ para todo o i. Depois avaliando esta estimativa inicial em cada observação X_i, calculam-se os parâmetros locais λ_i, por

$$\lambda_i = \left\{ \frac{\tilde{f}(X_i)}{g} \right\}^{\frac{1}{2}}$$

onde g é a média geométrica de $\tilde{f}(X_i)$:

$$\log g = n^{-1} \sum_{i=1}^{n} \log \tilde{f}(X_i)$$

Utilizando estes pesos, calcula-se o estimador adaptativo do núcleo $\hat{f}(t)$ por:

$$\hat{f}(t) = n^{-1} h^{-1} \sum_{i=1}^{n} \lambda_i^{-1} K\left\{ h^{-1} \lambda_i^{-1} (x - X_i) \right\}$$

onde K é a função do núcleo que, como no caso do núcleo fixo, é simétrica e $\int K(x)\, dx = 1$ e h é o parâmetro de alisamento. O factor λ_i é necessário para ajustar a largura da janela em cada observação de forma a assegurar que a área total da estimativa da densidade seja um.

Estimador do núcleo variável

Este estimador é um caso particular do anterior. O estimador do núcleo variável é definido por:

$$\hat{f}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h d_{j,k}} K\left(\frac{t - X_j}{h d_{j,k}} \right) \tag{2.17}$$

onde $d_{j,k}$ é a distância de X_j ao k-ésimo ponto mais próximo, k é um inteiro positivo, K é a função do núcleo e h é o parâmetro de alisamento.

Tal como nos casos anteriores este estimador também pretende adaptar a quantidade de alisamento à densidade local dos dados. Neste caso, em locais onde a densidade é pequena os núcleos serão mais alisados, uma vez que a largura da janela do núcleo no ponto X_j é proporcional a $d_{j,k}$.

A estimativa do núcleo variável é uma função de densidade de probabilidade desde que a função do núcleo o seja. A diferença entre este estimador e o estimador generalizado dos k-vizinhos mais próximos é que em (2.16) a largura da janela utilizada para construir a estimativa em t depende das distâncias de t aos dados, enquanto que em (2.17) a largura das janelas são independentes do ponto t em que a densidade está a ser estimada e depende apenas da distância entre os pontos da amostra.

2.1.7 Estimador das séries ortogonais

Este método foi introduzido por Whittle (1958), Čencov (1962), Schwartz (1967) e Kronmal e Tarter (1968). Ott e Kronmal (1976) sugeriram a utilização das séries de Walsh para estimar a densidade para dados binários multivariados e Hall (1983b) mostraram que o método das séries pode ser aplicado à estimação da densidade de dados discretos, que podem não ser binários e a dados mistos.

A abordagem consiste em considerar que a função de densidade de probabilidade pode ser vista como uma onda ("waveform") que pode ser aproximada por uma expansão em série de funções base ortonormais (por exemplo, séries de Fourier). Seja $\{\phi_i\}$ o conjunto de funções base ortonormais, tal que

$$\int \phi_i(x)\phi_j(x)\,dx = \delta_{ij} = \begin{cases} 1 & i = j \\ \\ 0 & i \neq j \end{cases} .$$

Seja $f^*(x) = \sum_{i=1}^{\infty} a_i\ \phi_i(x)$ a projecção de $f(x)$ no espaço de $\{\phi_i\}$, onde a_i são coeficientes.

Para estimar $f^*(x)$, tem que se truncar a série num número finito de termos, s, e estimar a_i.

Os coeficientes podem ser estimados, minimizando um critério pré-definido, como por exemplo o erro quadrático integrado *(ISE)*,

$$\int \left\{ f(x) - \sum_{i=1}^{\infty} a_i \phi_i(x) \right\}^2 dx. \tag{2.18}$$

Diferenciando (2.18) em ordem a a_j e igualando a zero, tem-se

$$a_j = \int f(x)\phi_j(x)\, dx$$

que é o valor esperado de $\phi_i(x)$. Assim,

$$\hat{a}_j = \frac{1}{n} \sum_{k=1}^{n} \phi_j(x_k) \tag{2.19}$$

onde x_k, $(k = 1, 2, ..., n)$, são os valores amostrais.

O estimador final de $f(x)$ é

$$\hat{f}(x) = \sum_{i=1}^{s} \left\{ \frac{1}{n} \sum_{k=1}^{n} \phi_i(x_k) \right\} \phi_i(x).$$

As propriedades das estimativas obtidas pelo método das séries ortogonais dependem da série utilizada.

A generalização ao caso multivariado é quase imediata. A forma mais comum de definir as funções base ortonormais multivariadas, é como produto de funções base univariadas, como por exemplo

$$\Phi_{ij}(x, y) = \phi_i(x)\phi_j(y)$$

mas o número de termos necessários na série aumenta exponencialmente com o número de dimensões, o que dificulta a aplicabilidade deste método.

Para o estimador das séries ortogonais a escolha do ponto em que se trunca a série, vai determinar a quantidade de alisamento. Quando a densidade é pequena numa determinada região é necessário que s, o número de termos da série considerados, seja maior para obter uma estimativa satisfatória. Uma forma de obviar este problema é minimizar o erro quadrático médio integrado *(MISE)*, mas a solução não é tão simples como (2.19).

2.1.8 Estimador de funções peso gerais

Pode-se definir uma classe de estimadores gerais que, por um lado engloba alguns dos estimadores já apresentados e por outro permite definir outos estimadores com propriedades idênticas a esses.

Se definirmos uma função $w(x,y)$ tal que

$$\int_{-\infty}^{\infty} w(x,y)\,dy = 1 \tag{2.20}$$

e

$$w(x,y) \geq 0 \quad \text{para qualquer } x \text{ e } y \tag{2.21}$$

pode-se obter uma estimativa da função de densidade da seguinte forma:

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} w(x_i,t)\,. \tag{2.22}$$

Esta estimativas são conhecidas por *estimativas de funções peso gerais.*

Para obter alguns dos estimadores anteriormente discutidos como caso especial de (2.22) basta definir $w(x,y)$. Assim, por exemplo, para o histograma teremos

$$w(x,y) = \begin{cases} \frac{1}{h(x)} & \text{se } x \text{ e } y \text{ cairem no mesmo rectângulo} \\ 0 & \text{outros casos} \end{cases}$$

onde $h(x)$ é a largura do rectângulo que contém x. Para a estimativa do núcleo

$$w(x,y) = \frac{1}{h}K\left(\frac{y-x}{h}\right).$$

2.2 Escolha do parâmetro de alisamento

2.2.1 Método do núcleo

A proximidade do estimador $\hat{f}$ à verdadeira função de densidade f depende em grande parte do parâmetro de alisamento h utilizado. A escolha de um valor para h demasiado pequeno leva a que $\hat{f}$ tenha demasiados picos nos pontos x_i da amostra e que tenha um aspecto irregular por toda a parte. Se h é demasiado grande todos os detalhes, espúrios

ou não, ficam ocultos e $\hat{f}$ ficará demasiado alisada do que resultará enviesamento. O valor de h deve ser escolhido de forma a obter um alisamento que seja um compromisso aceitável entre o enviesamento que pode ser tolerado e a flutuação aleatória. O h óptimo está também dependente do critério escolhido e do núcleo utilizado. A eficiência do estimador do núcleo está portanto muito dependente do parâmetro de alisamento, o que levou ao aparecimento de muitas propostas para a sua selecção a partir dos dados observados. Vai-se de seguida considerar algumas dessas propostas.

Escolha subjectiva

Uma forma de escolher o parâmetro de alisamento é fazer o gráfico das diversas curvas estimadas com diferentes parâmetros de alisamento e escolher a estimativa que estiver mais de acordo com a ideia que o investigador tem acerca dessa densidade.[4]

Esta forma de escolher o parâmetro de alisamento tem a vantagem de nos dar uma visão mais completa dos dados em análise, mas se for necessário estimar a densidade para um número elevado de conjuntos de dados é essencial utilizar um método automático.

Minimização do erro quadrático médio integrado (MISE)

Para medir a discrepância entre o estimador $\hat{f}$ e a verdadeira função de densidade f, podemos começar por considerar a estimação para um só ponto, sendo a medida mais óbvia dada pelo *erro quadrático médio* (2.1), existindo um compromisso entre os termos da variância e do enviesamento, como se pode verificar em (2.2): ajustando o parâmetro de alisamento, podemos reduzir (aumentar) o enviesamento se permitirmos que a variância aumente (diminua).

Se $f(x)$ fosse conhecida poderíamos escolher h de forma a minimizar o *erro quadrático integrado (ISE)* (2.3) ou, em alternativa, o *erro quadrático médio integrado (MISE)* (2.4).

No caso de amostras finitas, para obter as expressões exactas de (2.2) e (2.7) podemos utilizar

$$E\left[\hat{f}(x)\right] = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)\, dy \tag{2.23}$$

[4] Hand, D.J. (1981)

e

$$n \operatorname{var} \hat{f}(x) = \int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)\, dy - \left\{\frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)\, dy\right\}^2 \qquad (2.24)$$

mas os cálculos são árduos e de difícil interpretação, excepto no caso em que o núcleo é uma função de densidade normal estandardizada e a verdadeira densidade é também normal. [5] Neste caso

$$h_{opt} = (4\pi)^{-\frac{1}{10}} \left(\frac{3}{8}\pi^{-\frac{1}{2}}\right)^{-\frac{1}{5}} \sigma n^{-\frac{1}{5}} = 1.06\sigma n^{-\frac{1}{5}} \qquad (2.25)$$

minimizará o *MISE*.

Uma propriedade interessante de (2.23) é que o enviesamento depende do parâmetro de alisamento escolhido, mas não depende directamente do tamanho da amostra, excepto se h for escolhido como função de n.

Admitindo que o núcleo K é uma função simétrica que satisfaz

$$\int K(t)\, dt = 1 \qquad (2.26)$$

$$\int t\, K(t)\, dt = 0 \qquad (2.27)$$

$$\int t^2 K(t)\, dt = k_2 \neq 0 \qquad (2.28)$$

e que f tem derivadas contínuas de todas as ordens necessárias,[6] o erro quadrático médio integrado *(MISE)* admite a seguinte representação assimptótica:

$$AMISE = \frac{1}{4} h^4 k_2^2 \int f''(x)^2\, dx + n^{-1}h^{-1} \int K(t)^2\, dt \qquad (2.29)$$

Como foi mostrado por Parzen, (1962, lemma 4A),

$$h_{opt} = k_2^{-2/5} \left\{\int K(t)^2\, dt\right\}^{1/5} \left\{\int f''(x)^2\, dx\right\}^{-1/5} \qquad (2.30)$$

Substituindo (2.30) em (2.29) o seu valor aproximado será

$$\frac{5}{4} C(K) \left\{\int f''(x)^2\, dx\right\}^{\frac{1}{5}} n^{-\frac{4}{5}}$$

[5] Para detalhes sobre o cálculo ver Fryer (1976) ou Deheuvels (1977)

[6] Silverman (1986), cap.3, pag.38

onde $C(K) = k_2^{\frac{2}{5}} \left\{ \int K(t)^2 \, dt \right\}^{\frac{4}{5}}$. Hodges e Lehman (1956) mostraram que a minimização de $C(K)$, sob a condição de que (2.26) e (2.28) sejam iguais a um, se resolve se $K(t)$ for o núcleo de Epanechnikov

$$K_e(x) = \begin{cases} \frac{3}{4\sqrt{5}} \left[1 - \frac{x^2}{5} \right] & \text{se } |x| \leq \sqrt{5} \\ 0 & \text{outros casos} \end{cases} . \tag{2.31}$$

Foram comparados vários núcleos simétricos com o de Epanechnikov para testar as suas eficiências relativamente a este e a conclusão é que, considerando o *MISE,* as diferenças não são significativas. [7]

Validação cruzada dos mínimos quadrados (LSCV)

Este método foi sugerido por Rudemo (1982) e Bowman (1984). Ver também Bowman, Hall e Titterington (1984), Hall (1983a) e Stone (1984). O erro quadrático integrado *(ISE)* pode ser escrito na seguinte forma:

$$\int \left\{ \hat{f}(x) - f(x) \right\}^2 dx = \int \hat{f}(x)^2 \, dx - 2 \int \hat{f}(x) f(x) \, dx + \int f(x)^2 \, dx$$

Como $\int f(x)^2 \, dx$ não depende de h podemos ignorar este termo e a escolha da largura da janela corresponde então a escolher h que minimize

$$R\left(\hat{f}\right) = \int \hat{f}(x)^2 \, dx - 2 \int \hat{f}(x) f(x) \, dx.$$

Definindo $\hat{f}_{-i}$ como estimativa da densidade considerando todos os pontos excepto X_i, $\hat{f}_{-i}$ não é dependente de X_i e podemos considerar $\hat{f}_{-i}(X_i)$ como uma forma de medir se o valor considerado para h é apropriado como valor do parâmetro de alisamento. [8] Como i percorre toda a amostra obtêm-se n medidas de ajustamento e considera-se a média dessas medidas.

Se definirmos agora

$$M_0(h) = \int \hat{f}(x)^2 \, dx - 2n^{-1} \sum_i \hat{f}_{-i}(X_i) \tag{2.32}$$

[7] Silverman (1986), cap.3, pag.43

[8] Schilling (1983)

a ideia da validação cruzada dos mínimos quadrados *(LSCV)* é minimizar M_0 em ordem a h, porque $E[M_0(h)] = E\left[R\left(\hat{f}\right)\right]$. Stone (1984) mostrou que, sob condições fracas para o núcleo, este método é assimptoticamente óptimo, no sentido de que minimiza o *MISE,* para todas as funções limitadas.

Apesar deste método ser atractivo, Park e Marron (1990) demonstraram que o parâmetro de alisamento, h, que minimiza (2.32) sofre de demasiada variabilidade amostral, isto é, diferentes conjuntos de dados com a mesma distribuição darão, demasiadas vezes, resultados muito diferentes. Quando aplicado a dados discretizados tem resultados fracos, sendo aconselhável limitar a pesquisa de h a um intervalo restrito.

Validação cruzada de máxima verosimilhança

Este método consiste em omitir uma observação X_i da amostra utilizada para obter a estimativa da densidade $\hat{f}$ e utilizar essa observação como se fosse uma observação independente de f. Nesse caso o logaritmo da verosimilhança de $\hat{f}_{-i}(X_i)$ será $log\ \hat{f}_{-i}(X_i)$, onde $\hat{f}_{-i}$ tem a mesma definição que no caso anterior. A função "score" é dada por

$$CV(h) = n^{-1} \sum_{i=1}^{n} log\ \hat{f}_{-i}(X_i). \tag{2.33}$$

Escolhe-se para h o valor que maximiza esta função, para os dados em análise. A função *"score"* (2.33) foi sugerida por Habemma, Hermans e van der Broek (1974) e Duin (1976). Scott e Factor (1981) chamaram a atenção de que $CV(h)$ é muito sensível a valores extremos. Além disso, Schuster e Gregory (1981) mostraram que a utilização deste método pode levar a estimativas inconsistentes nos casos em que a cauda da distribuição de f seja monótona e caia a uma taxa lenta.

As dificuldades que ocorrem para a aplicação de *LSCV* a dados discretos também se verificam com este método, sendo também neste caso aconselhável fazer a pesquisa de h num intervalo limitado.

Teste gráfico

Este método é parcialmente subjectivo tendo como hipótese que as estimativas estão uniformemente próximas da verdadeira densidade e tem como base um teorema de

Silverman (1978a), que nos diz que, sob certas condições, se h for uma função de n e for escolhido de forma a minimizar o erro máximo na estimação da densidade, então quando $n \to \infty$,

$$\frac{\sup\left|\hat{f}'' - E\left(\hat{f}''\right)\right|}{\sup\left|E\left(\hat{f}''\right)\right|} \to k \tag{2.34}$$

onde $\hat{f}''$ representa as segundas derivadas da densidade estimada e onde k apenas depende do núcleo escolhido. O numerador de (2.34) representa o ruído aleatório na curva $\hat{f}''$, uma vez que $E\left(\hat{f}''\right)$ é uma versão alisada de uma $\hat{f}''$ e o denominador é a tendência dessa curva. O ruído aleatório aparecerá como flutuações rápidas na curva $\hat{f}''$ e, assimptoticamente terá amplitude máxima de $\pm k \sup\left|\hat{f}''\right|$.

O método proposto para a escolha do parâmetro é desenhar gráficos de $\hat{f}''$ para diferentes valores de h e escolher o parâmetro de alisamento de forma a que $\hat{f}''$ tenha flutuações rápidas bastante acentuadas mas que não escondam completamente a variação sistemática e com esse parâmetro construir a estimativa da densidade original.

Este método pode generalizar-se ao caso multivariado. Supondo que $X_1, X_2, ..., X_n$, são observações independentes e identicamente distribuídas de uma densidade f, d-dimensional, o teste gráfico será nesse caso:

$$\nabla^2 \hat{f}(x) = \sum_{j=1}^{n} n^{-1} h^{-d-2} \nabla^2 \delta\{h^{-1}(x - X_j)\}$$

onde δ é um núcleo d-dimensional e a melhor largura da janela é a que dá flutuações de dimensão $k \sup\left|\nabla^2 \hat{f}\right|$ no teste gráfico, que é uma superfície d-dimensional. Para dados multivariados, as principais dificuldades apontadas por Silverman para utilização deste método são: a obtenção de testes gráficos para mais de duas dimensões, a de que é necessário avaliar a ordenada do teste gráfico num grande número de pontos para ter uma ideia grosseira da sua forma e o tempo de computação necessário.

Devido ao grau de subjectividade da escolha, este método é geralmente utilizado para confirmação da escolha efectuada por outros métodos.

2.3 Estimadores não paramétricos para dados multivariados

2.3.1 Histogramas multivariados

A generalização do histograma univariado ao caso multivariado é imediata: em vez de dividir uma linha em intervalos disjuntos de tamanho igual, divide-se todo o espaço em células disjuntas de igual volume.

Dada uma amostra $f(\mathbf{x})$, onde $\mathbf{x}$ pertence a $\mathcal{R}^d$, faça-se uma partição do espaço em paralelipípedos de tamanho $h_1 \times h_2 \times ... \times h_d$. Considere-se B_k, o k-ésimo paralelípipedo que contém v_k pontos ($\sum_k v_k = n$). O histograma é então definido por:

$$\hat{f}(\mathbf{x}) = \frac{v_k}{n h_1 h_2 ... h_d}$$

Assimptoticamente o *MISE multivariado* é minimizado se

$$h_{k,opt} = R(f_k)^{-1/2} \left(6 \prod_{i=1}^{d} R(f_i)^{1/2} \right)^{1/(2+d)} n^{-1/(2+d)}$$

sendo a regra de referência da Normal

$$h_{k,opt} = 3.5\, s_k\, n^{-1/(2+d)}$$

A construção do histograma para dados multivariados traz também diversas dificuldades associadas à escolha da origem do sistema e das coordenadas das direcções das células da grelha. Mesmo para dados bivariados existem várias dificuldades associadas à representação e interpretação do histograma, devido à natureza descontínua dos "blocos" que torna difícil a percepção da estrutura dos dados. Ainda que se escolha uma largura do rectângulo suficientemente pequena para conseguir "apanhar" razoavelmente a informação local dos dados em qualquer das direcções das coordenadas, o número total de "caixas" torna-se tão grande que o efeito dos erros aleatórios pode tornar-se dominante.

No caso de dados multivariados, em que cada variável está dividida em M intervalos, se tivermos p variáveis, teremos M^p células. Uma vez que a função de densidade de probabilidade é estimada pela proporção de pontos da amostra que caiem em cada

célula, ou temos um número elevadíssimo de observações ou a estimativa será zero em quase toda a parte, o que é uma grande desvantagem.

Para obviar o problema da necessidade de muitas observações mesmo para problemas de baixa dimensionalidade, Sebestyen e Edie (1966) propuseram um algoritmo[9] que permite que a localização, forma e tamanho das classes sejam definidos iterativamente pelos próprios dados, de forma a poder representar os dados com muito menos células.

Para cada classe o algoritmo toma os pontos amostrais, um de cada vez, e começa por centrar a primeira célula do histograma no primeiro ponto amostral. Os pontos subsequentes podem cair nas células existentes ou numa zona próxima (nesse caso é armazenado temporariamente e mais tarde é utilizado para actualizar a célula mais próxima) ou ainda numa zona completamente diferente das células existentes, formando o centro de uma nova célula. Quando um ponto cai numa célula ela é actualizada recalculando o centro e a dimensão desta. O procedimento de actualização é apenas uma forma sequencial de calcular médias e variâncias dos pontos que definem cada célula.

O problema da queda abrupta para zero fora da fronteira de cada classe foi resolvido por estes autores utilizando um decrescimento normal local para a densidade na fronteira de cada célula.

Mucciardi e Gose (1972)[10] desenvolveram métodos mais formais, que incluem a estimação da densidade marginal de cada classe por uma mistura de duas distribuições normais.

Estes autores resolveram o problema da descontinuidade na fronteira de cada célula e o problema da queda abrupta para zero fora da fronteira, substituindo cada célula por uma distribuição normal com um peso proporcional ao número de pontos dessa célula.

Estes métodos são pouco conhecidos, havendo fortes argumentos para a utilização de outras estimativas de densidade em vez dos histogramas no caso de duas ou mais dimensões.

[9] Ver Hand (1981), cap.2, pag. 18, para uma descrição destes procedimentos.

[10] Hand (1981)

2.3.2 Estimadores do produto de núcleos

Dada uma matriz, $\mathbf{X}$, $n \times p$, com elemento genérico x_{ij}, de vectores aleatórios $\mathbf{x}_i=(x_{i1},...,x_{ip})$, onde $\mathbf{x}_1,...,\mathbf{x}_n$ são observações independentes provenientes de uma densidade multivariada $f(\mathbf{x})$ de dimensão p, o estimador multivariado do produto de núcleos, utiliza o mesmo núcleo univariado em cada dimensão, mas com parâmetros de alisamento diferentes:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1...h_p} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{p} K\left(\frac{\mathbf{x}_j - x_{ij}}{h_j} \right) \right\}.$$

Para o estimador multivariado do produto de núcleos, as componentes do erro quadrático médio assimptótico *(AMISE)*,[11] são:

$$\begin{aligned} AISB &= \frac{1}{4}\sigma_K^4 \left[\sum_{i=1}^{p} h_i^4 R(f_{ii}) + \sum_{i \neq j} h_i^2 h_j^2 \int_{\mathcal{R}^d} f_{ii} f_{jj}\, d\mathbf{x} \right] \\ AIV &= \frac{R(K)^p}{nh_1h_2...h_p} - \frac{R(f)}{n} + O\left(\frac{h}{n}\right). \end{aligned}$$

Para calcular os parâmetros de alisamento óptimos é necessário encontrar a solução de p equações não lineares.

Se os dados tiverem distribuição normal bivariada e o núcleo for também normal, o erro quadrático médio assimptótico *(AMISE)* é minimizado quando:

$$h_{i,opt} \approx \sigma_i(1 - \rho^2/2 - \rho^4/16 - ...)n^{-1/6} \qquad i = 1, 2.$$

Note-se contudo que, se os dados forem perfeitamente correlacionados, o *AMISE* diverge para infinito.

Se os dados tiverem distribuição normal multivariada, com todas as variáveis independentes e se for utilizado um núcleo normal, a regra de referência será:

$$h_{i,opt} = \left(\frac{4}{d+2} \right)^{1/(d+4)} \sigma_i n^{-1/(d+4)}.$$

Para outros núcleos, o parâmetro de alisamento correspondente pode-se obter dividindo $h_{i,opt}$ pelo desvio padrão desse núcleo.

Scott (1992) recomenda a utilização deste núcleo nos casos práticos.

[11] Scott (1992, pag. 150) - Teorema 6.4

2.3.3 Estimadores do núcleo

Para uma matriz de observações contínuas, $\mathbf{X}$, com p varáveis, a estimativa não paramétrica de densidade multivariada dada pelo estimador do núcleo K, com parâmetro de alisamento h, é definida por:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} \mathbf{K}_p \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right)$$

onde $\mathbf{K}_p$ é agora uma função não negativa definida para $\mathbf{x}$ p-dimensional, tal que

$$\int_{\mathcal{R}^p} \mathbf{K}_p(\mathbf{x})\, d\mathbf{x} = 1. \tag{2.35}$$

Muitas vezes exige-se que $\mathbf{K}_p$ seja simétrico, isto é:

$$\mathbf{K}_p(\mathbf{x}) = \mathbf{K}_p(-\mathbf{x}) \qquad \mathbf{x} \in \mathcal{R}^p.$$

Uma escolha possível é, por exemplo, a função de densidade normal multivariada estandardizada

$$\mathbf{K}_p(\mathbf{x}) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{x} \right).$$

Os núcleos normais são bastante populares por garantirem a continuidade e diferenciabilidade.

O facto destes núcleos terem uma região de suporte ilimitada ($\mathbf{K}(\mathbf{x}) > \mathbf{0}$ para todo o $\mathbf{x}$) significa que todos os pontos da amostra contribuem para estimar $\hat{f}(\mathbf{x})$.

Um outro núcleo possível é o núcleo multivariado de Epanechnikov

$$\mathbf{K}(\mathbf{x}) = \begin{cases} -\frac{1}{2} c_p^{-1} (p+2) \left(1 - \mathbf{x}^T \mathbf{x}\right) & \text{se } \mathbf{x}^T \mathbf{x} < \mathbf{1} \\ 0 & \text{outros casos} \end{cases}$$

onde c_p é o volume de uma esfera unitária p-dimensional: $c_1 = 2, c_2 = \pi, c_3 = 4\pi/3$, etc..

Este núcleo é considerado óptimo no sentido em que minimiza o menor erro quadrático médio integrado. Contudo outros núcleos podem conseguir um erro quadrático médio integrado semelhante.

A escolha da função $\mathbf{K}$ adequada ao problema em análise é bastante difícil e é importante ter algum cuidado na sua escolha para estimativas de densidade multivariadas,

principalmente se o tamanho da amostra não é grande. Embora os núcleos normais sejam os mais frequentemente considerados, existem outras formas para a função $\mathbf{K}$ de cálculo mais fácil e rápido, como por exemplo o caso em que temos um núcleo rectangular. Os núcleos lineares por partes podem ser uma alternativa aceitável. Quando os núcleos têm regiões de suporte finitas ($\mathbf{K}(\mathbf{x}) = \mathbf{0}$ para um grande número de $\mathbf{x}$) levam a uma computação mais rápida uma vez que $\mathbf{K}$ não necessita ser calculado para todos os pontos de $\mathbf{x}$, mas $\hat{f}(\mathbf{x})$ pode tomar valores zero, o que é uma desvantagem especialmente em espaços de dimensão elevada.

Para $p = 2$ podemos considerar dois núcleos que têm melhores propriedades de diferenciabilidade que o núcleo de Epanechnikov e que podem ser calculados mais rapidamente que o núcleo normal, utilizando o algoritmo apropriado:[12]

$$\mathbf{K}(\mathbf{x}) = \begin{cases} 3\pi^{-1}\left(1 - \mathbf{x}^T\mathbf{x}\right)^2 & \text{se} \quad \mathbf{x}^T\mathbf{x} < \mathbf{1} \\ 0 & \text{outros casos} \end{cases}$$

$$\mathbf{K}(\mathbf{x}) = \begin{cases} 4\pi^{-1}\left(1 - \mathbf{x}^T\mathbf{x}\right)^3 & \text{se} \quad \mathbf{x}^T\mathbf{x} < \mathbf{1} \\ 0 & \text{outros casos} \end{cases}$$

A utilização de um único parâmetro de alisamento, h, implica que o núcleo colocado em cada ponto da amostra tem a mesma escala em todas as direcções, o que pode não ser apropriado, por exemplo, no caso em que os dados tenham maior dispersão na direcção de uma das coordenadas do que nas outras. Nestes casos é preferível utilizar um vector de parâmetros ou então transformar os dados para evitar a diferença de dispersão ao longo das várias coordenadas.

Um estimador do núcleo geral multivariado pode ser definido da seguinte forma:[13]

$$\hat{f}(\mathbf{x}) = \frac{1}{n\,|\mathbf{H}|} \sum_{i=1}^{n} \mathbf{K}_p\left\{\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_j)\right\}$$

onde $\mathbf{H}$ é uma matriz não singular $p \times p$. A transformação linear $\mathbf{H}$ pode ser incorporada na definição do núcleo, permitindo escolher um núcleo multivariado com uma matriz de covariâncias simples. Por exemplo escolher $\mathbf{K}_p$ como $N(\mathbf{0}, \boldsymbol{\Sigma})$ e $\mathbf{H} = \mathbf{I}_p$ ou $\mathbf{K}_p$ como $N(\mathbf{0}, \mathbf{I_p})$ e $\mathbf{H} = \boldsymbol{\Sigma}^{1/2}$ é equivalente.

[12] Ver Silverman (1986) cap.4, pag.87

[13] Scott, D.W. (1992)

Se se definir um escalar $h > 0$ e uma matriz $\mathbf{A}$ que satisfaça

$$\mathbf{H} = h\mathbf{A} \text{ onde } |\mathbf{A}| = 1$$

e se o núcleo multivariado satisfizer (2.35) e

$$\begin{aligned} \int_{\mathcal{R}^p} \mathbf{x}\, \mathbf{K}_p(\mathbf{x})\, d\mathbf{x} &= \mathbf{0} \\ \int_{\mathcal{R}^p} \mathbf{x}\mathbf{x}^T \mathbf{K}_p(\mathbf{x})\, d\mathbf{x} &= I_p \end{aligned}$$

o *erro quadrático médio assimptótico (AMISE)* é dado por

$$AMISE = \frac{R(K)}{nh^d} + \frac{1}{4}h^4 \int_{\mathcal{R}^p} [tr\{\mathbf{A}\mathbf{A}^T \nabla^2 f(\mathbf{x})\}]^2\, d\mathbf{x}.$$

Note-se que esta parametrização permite parâmetros de alisamento diferentes para cada dimensão, sendo equivalente a utilizar um núcleo de forma elíptica com uma rotação arbitrária. Por exemplo, se

$$\mathbf{H} = \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_p \end{pmatrix} \quad \text{então} \quad \mathbf{H} = h \begin{pmatrix} h_1/h & & 0 \\ & \ddots & \\ 0 & & h_p/h \end{pmatrix}$$

onde $h = (h_1 h_2 ... h_p)^{1/p}$ é a média geométrica de p parâmetros de alisamento.

2.3.4 Estimadores do núcleo para dados binários

Aitchison e Aitken (1976) propuseram a utilização de um núcleo binomial para as situações em que cada variável do vector $\mathbf{x}$, é uma variável que só pode tomar os valores zero ou um, sendo estimado por:

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_p(\mathbf{x}; \mathbf{x}_i, h) \quad \text{onde} \quad \mathbf{K}_p(\mathbf{x}; \mathbf{x}_i, h) = h^{p-d_i^2}(1-h)^{d_i^2} \tag{2.36}$$

com $\frac{1}{2} \leq h \leq 1$ e $d_i^2 = \|\mathbf{x} - \mathbf{x}_i\|^2 = (\mathbf{x} - \mathbf{x}_i)^T (\mathbf{x} - \mathbf{x}_i)$ é o número de discordâncias dos elementos correspondentes.

2.3.5 Estimador do núcleo adaptativo

Este estimador é especialmente importante no caso dos dados multivariados dada a importância das caudas da distribuição neste tipo de dados, que são bastante mais graves que no caso unidimensional. Primeiro faz-se uma estimativa piloto, com parâmetro de alisamento, h, fixo. A partir desta estimativa são definidos os factores de alisamento locais, λ_i, dados por

$$\lambda_i = \left\{ \tilde{f}(\mathbf{x}_i) / g \right\}^{-\alpha}$$

onde α é o *parâmetro sensitivo,* que reflecte a sensibilidade da largura da janela a variações na estimativa piloto, satisfazendo $0 \leq \alpha \leq 1$ e g é um factor de escala [média geométrica de $\tilde{f}(\mathbf{x}_i)$]. Utilizando estes pesos, calcula-se o estimador do núcleo adaptativo $\tilde{f}(\mathbf{x})$ por:

$$\hat{f}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} h^{-p} \lambda_i^{-p} \mathbf{K} \left\{ h^{-1} \lambda_i^{-1} (\mathbf{x} - \mathbf{x}_i) \right\}$$

onde $\mathbf{K}$ é a função do núcleo e h é o parâmetro de alisamento. Como no caso do núcleo fixo, esta função é simétrica e $\int_{\mathcal{R}^p} \mathbf{K}_p(\mathbf{x})\, d\mathbf{x} = 1$.

Embora se possam levantar algumas questões em relação a este método, nomeadamente em relação à construção da estimativa inicial, por requerer que se utilize outro método para estimar a densidade, Breiman, Meisell e Purcell (1977), Abramson (1982) e Silverman (1986) consideram que o método é insensível à estimativa inicial, sendo a escolha natural para o caso multivariado o núcleo de Epanechnikov.

A escolha de $\alpha = \frac{1}{2}$ é sugerida por alguns autores, que demonstraram haver um melhor comportamento deste estimador para as caudas da distribuição do que o conseguido com o estimador do núcleo com parâmetro de alisamento fixo, e que o seu enviesamento se situa entre o do estimador do núcleo com parâmetro de alisamento fixo e o do estimador dos k-vizinhos mais próximos.

2.3.6 Estimador dos k-vizinhos mais próximos

Para um ponto fixo $\mathbf{x}$, e para um inteiro k também fixo, seja $d_k(\mathbf{x})$ a distância Eucledíana de $\mathbf{x}$ ao k-ésimo ponto mais próximo entre $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ e $C_k(\mathbf{x})$ o volume p-dimensional

de uma esfera p-dimensional de raio $d_k(\mathbf{x})$.

$C_k(\mathbf{x}) = v_p d_k(\mathbf{x})^p$ onde v_p é o volume da esfera unitária em p dimensões ($v_1 = 2, v_2 = \pi, v_3 = 4\pi/3$, etc.).

A densidade estimada pelo método do k-ésimo vizinho mais próximo é definida por:

$$\hat{f}(\mathbf{x}) = \frac{k/n}{C_k(\mathbf{x})} = \frac{k/n}{v_p d_k(\mathbf{x})^p}. \tag{2.37}$$

Abramson (1984) propôs que no caso p-dimensional k seja escolhido proporcional a $n^{4/(p+4)}$, sendo a constante de proporcionalidade dependente de $\mathbf{x}$.

Este estimador é considerado equivalente ao estimador do núcleo quando se estima a densidade num único ponto, mas não é muito útil quando se estima a densidade total f, dado que (2.37) é descontínua e tem integral infinito devido às caudas muito pesadas.

2.3.7 Estimador generalizado dos k-vizinhos mais próximos

O estimador generalizado dos k-vizinhos mais próximos para dados multivariados é definido por

$$\hat{f}(\mathbf{x}) = n^{-1}\{d_k(\mathbf{x})\}^{-p} \sum_{i=1}^{n} K_p\left\{[d_k(\mathbf{x})]^{-1}(\mathbf{x} - \mathbf{x}_i)\right\} \tag{2.38}$$

que se reduz à estimativa dos k-vizinhos mais próximos se:

$$K_p(\mathbf{x}) = \begin{cases} v_p^{-1} & \text{se} \quad \|\mathbf{x}\| \leq 1 \\ \\ 0 & \text{outros casos} \end{cases}$$

2.4 Análise discriminante não paramétrica

2.4.1 Regras de classificação

Para obter uma regra de classificação no contexto das estimativas de densidade é necessário não esquecer que esta está dependente de uma boa estimativa da densidade de cada um dos grupos. A escolha do parâmetro de alisamento pode ser feita por um dos métodos discutidos e deve ter em conta o caso concreto que se está a analisar, o critério

a optimizar e o núcleo utilizado. A utilização do mesmo parâmetro de alisamento para ambos os grupos pode não ser indicada e ser mais adequado utilizar parâmetros de alisamento diferentes para cada grupo.

O primeiro método não paramétrico de classificação foi a regra dos vizinhos mais próximos de Fix e Hodges (1951). Para classificar uma observação $\mathbf{x}$, suponha-se que se têm observações $\mathbf{x}_{ij}$ $(j = 1, 2, ..., n)$ do grupo G_i $(i = 1, 2)$.

Ordenem-se as n observações utilizando uma função distância $D(\mathbf{x}, \mathbf{x}_{ij})$, habitualmente a distância Euclediana, por ordem ascendente dessa distância, para obter n indíces $R_1, R_2, ..., R_n$ e escolha-se um inteiro k. Seja k_i o número de observações de G_i nas k observações mais próximas de $\mathbf{x}$, isto é, as observações com $R_j \leq k$.

A regra de classificação é a seguinte:

- Classifica-se $\mathbf{x}$ no grupo G_1 se

$$\frac{k_1}{n_1} \Big/ \frac{k_2}{n_2} > \frac{\pi_2}{\pi_1}. \tag{2.39}$$

Pressupondo que os grupos estão misturados e que a ordem das observações é aleatória, quando a distância entre as observações $\mathbf{x}_j$ é igual, a observação "mais próxima" pode ser escolhida por comparação dos indíces, isto é, se $\|\mathbf{x}_{j1} - \mathbf{x}\| = \|\mathbf{x}_{j2} - \mathbf{x}\|$ então $R_{j1} < R_{j2}$ se $j_1 < j_2$, caso contrário $R_{j1} > R_{j2}$. Se existirem pesos diferentes para estas observações, esse peso é dividido igualmente entre elas [Stone (1977)].

Alguns autores ignoram essa possibilidade, assumindo que $P\{f_i(\mathbf{x}) = kf_j(\mathbf{x})\} = 0$, sendo os casos fronteira decididos de forma arbitrária.

Se $k = 1$, a regra discriminante obtida em (2.39) é conhecida por regra do vizinho mais próximo de ordem 1 (1-NN) e classifica $\mathbf{x}$ no grupo do seu vizinho mais próximo na amostra. Para $k > 1$, a abordagem de Tomek (1976) incorpora uma variedade de regras discriminantes chamadas k-NN, que incluem maioria simples, isto é, classifica-se no grupo a que pertencem a maioria das k observações mais próximas de $\mathbf{x}$, sendo a decisão aleatória no caso em que existe o mesmo número de observações de ambos os grupos,[14] e algumas modificações que têm em conta a diferença de tamanho da amostra

[14] Tomek (1976) sugere que k seja ímpar para evitar esta situação.

para os dois grupos.[15]

As taxas de erro de má classificação para as regras $k-NN$ são habitualmente maiores que as taxas de erro de Bayes, o erro mínimo, mas têm a propriedade interessante de, assimptoticamente, o seu limite ser o dobro da taxa de erro de Bayes. Para amostras finitas, a escolha de k é importante e deve depender do número de variáveis e da rugosidade da função de densidade envolvida [Hand (1981)]. Loftsgaarden e Quesenberry (1965) sugeriram que se utilizasse, para o grupo G_i, um valor de k próximo de $\sqrt{n_i}$. Enas e Choi (1986) a partir de um estudo de simulação que efectuaram para o caso em que apenas existem dois grupos, sugerem que, para amostras de tamanho semelhante, se escolha k aproximadamente igual a $n^{3/8}$ ou $n^{2/8}$, consoante as diferenças entre as matrizes de covariâncias sejam pequenas ou grandes. Esta escolha será invertida se os tamanhos dos grupos forem muito diferentes.

No caso do *método do núcleo* a regra de classificação obtém-se, substituindo as densidades condicionais ao grupo pelas estimativas de densidade pelo método do núcleo, para cada grupo. A regra $\frac{\hat{f}_i(\mathbf{x})}{\hat{f}_j(\mathbf{x})} > k_{ij}$ assim obtida é consistente, desde que o núcleo escolhido satisfaça as condições de regularidade já apontadas.

A escolha do parâmetro de alisamento, segundo a abordagem tradicional à análise discriminante, é efectuada separadamente para cada grupo de acordo com o critério a optimizar.

Van Ness e Simpson (1976), Van Ness (1979), Tutz (1986,1988,1989), Hall e Wand (1988) consideraram uma abordagem diferente: a escolha conjunta dos parâmetros de alisamento. Uma vez que o essencial da análise discriminante é obter uma regra de classificação, justifica-se a utilização de funções perca directamente relacionadas com o problema de discriminação em análise. Os parâmetros de alisamento são seleccionados simultaneamente por minimização da estimativa da taxa de erro de Bayes total, por validação cruzada ou *"leaving-one-out"*, obtida com as estimativas do núcleo das densidades condicionais aos grupos. Tutz (1986) mostrou que esta abordagem leva a uma regra de classificação consistente. Contudo, esta estimativa é uma função descontínua

[15] Silverman e Jones (1989)

dos parâmetros de alisamento, necessitando de ser alisada para poder ser implementada.

Hall e Wand (1988) propuseram que, no caso de existirem apenas dois grupos, os parâmetros de alisamento fossem escolhidos por aplicação da validação cruzada dos mínimos quadrados *(LSCV)*, em relação à *diferença entre as densidades condicionais dos dois grupos.*

Para *variáveis binárias*, a estimativa de densidade do núcleo de Aitchison e Aitken (1976), para o grupo G_i, é

$$\hat{f}_i(\mathbf{x}) = \frac{1}{n_i} \sum_{i=1}^{n_i} h_i^{p-d_{ij}^2} (1-h_i)^{d_{ij}^2} \quad (i=1,2) \tag{2.40}$$

onde $d_{ij}^2 = \|\mathbf{x} - \mathbf{x}_{ij}\|^2$ para $(i=1,2)$ e $(j=1,2,...,n_i)$ e a regra de classificação pode-se escrever da seguinte forma:

- $\mathbf{x}$ pertence ao grupo G_1 se $\pi_1 f_1(\mathbf{x}) - \pi_2 f_2(\mathbf{x}) \geq 0$.

Os autores estimaram esta diferença, considerando π_1 e π_2 fixos e utilizando a estimativa do núcleo (2.40) e uma variante da validação cruzada para escolher h_1 e h_2 que conjuntamente minimizem

$$MISE(h_1,h_2) = E \sum_{\mathbf{x}} \{\hat{v}(\mathbf{x}) - v(\mathbf{x})\} \tag{2.41}$$

onde $v(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) - \pi_2 f_2(\mathbf{x})$ e $\hat{v}(\mathbf{x})$ é a estimativa de $v(\mathbf{x})$. Como os valores assimptóticos óptimos de h_1 e h_2 que minimizam (2.41) podem ser negativos, sugeriram que na prática se minimizasse

$$\sum_{\mathbf{x}} \{\hat{v}(\mathbf{x})\}^2 - 2 \left[\sum_{i=1}^{2} \pi_i^2 n_i^{-1} \sum_{j=1}^{n_i} \hat{f}_{i(ij)}(\mathbf{x}_{ij}) - \pi_1 \pi_2 \sum_{i=1}^{2} n_i^{-1} \sum_{j=1}^{n_i} \hat{f}_i(\mathbf{x}_{ij}) \right] \tag{2.42}$$

onde $\hat{f}_{i(ij)}$ é a estimativa de densidade do núcleo $\hat{f}_i(\mathbf{x})$ para $\mathbf{x}_{ik}$ $(k=1,...,n;\ k \neq j)$. Retirando o termo que não envolve h_1 e h_2, (2.42) é um estimador não enviesado de (2.41).

Para dados contínuos, os autores basearam a sua abordagem no estimador do produto do núcleo pressupondo que os dados estão estandardizados. A expressão a minimizar é (2.42), substituindo o primeiro termo por $\int \{\hat{v}(\mathbf{x})\}^2 dv$.

A utilização das regras de classificação obtidas a partir de estimativas pelo *método do produto do núcleo* está limitada a casos em que as variáveis são independentes ou

têm uma correlação moderada. Murphy e Moran (1986), num estudo efectuado para variáveis com função de distribuição condicional ao grupo normal multivariada, em que as variáveis não eram sempre independentes, concluiram que as regras de classificação baseadas neste método, quando as variáveis eram independentes ou tinham uma correlação positiva moderada eram superiores às obtidas pelas regras em que se assume a distribuição normal. Quando existiam outros padrões de correlação, o desempenho das regras discriminantes obtidos por este método dariam muitas vezes resultados fracos em termos de classificação.

Capítulo 3

Exemplos ilustrativos

Dado que uma importante motivação para o estudo da análise discriminante é a sua potencialidade para aplicação a diversos tipos de dados, neste capítulo serão apresentados alguns resultados da aplicação das principais metodologias apresentadas, para dois grupos. Concretamente as metodologias serão aplicadas a três conjuntos de dados reais e alguns dados simulados.

A comparação do desempenho das metodologias será feita tendo em conta as taxas de erro aparentes, as obtidas por validação cruzada e as *"bootstrap"*, quando possível, das regras de classificação obtidas pelos vários métodos.

Serão ainda descritos, duma forma sucinta, os programas que implementam os vários métodos. Estes foram desenvolvidos em linguagem S, utilizando o "software" S-Plus, desenvolvido pela StatSci da MathSoft, que tem um ambiente flexível para análise de dados e incorpora algumas ferramentas que facilitam a programação a efectuar pelo utilizador.

3.1 Descrição dos exemplos

Para obter as regras de classificação para estes conjunto de dados, consideraram-se as seguintes hipóteses:

1 - Qualquer que seja o método, consideram-se os custos de má classificação constantes e iguais probabilidades *a priori* para ambos os grupos.

2 - Para obter a regra *discriminante linear (LDA) óptima*, tem-se como hipótese subjacente que as matrizes de covariâncias dos grupos são idênticas e a distribuição condicional aos grupos é normal multivariada. Como não se conhecem os parâmetros da população utilizam-se as médias e variâncias empíricas dos grupos como estimativas desses parâmetros para poder calcular a recta de separação dos dois grupos.

3 - Para obter a regra da *discriminante quadrática (QDA)*, está-se a considerar que as matrizes de covariâncias dos grupos são diferentes e a sua distribuição condicional é normal multivariada. Tal como na hipótese anterior utilizam-se as médias e variâncias empíricas dos grupos como estimativas dos parâmetros da população.

4 - A regra obtida a partir do *estimador do produto do núcleo (Nucleo)*, não põe qualquer hipótese sobre a distribuição condicional aos grupos, que é necessário estimar. Considera-se que a correlação entre as variáveis é moderada. Embora habitualmente se pressuponha a estandardização dos dados para a implementação deste método, como se utilizaram diferentes parâmetros de alisamento, para cada variável, em cada um dos grupos, não se considerou necessário efectuar essa transformação nas variáveis. Scott (1992), cap. 6, pag.180, diz ser importante ter um parâmetro de alisamento para cada direcção, mesmo para dados estandardizados.[1]

O núcleo utilizado em todos os casos é o Normal estandardizado e o parâmetro de alisamento utilizado em cada variável é escolhido por aplicação de (2.25).

5 - Para obter a regra a partir do *estimador dos k-vizinhos mais próximos (k-NN)*, também não se supõe conhecida a distribuição condional aos grupos e utiliza-se a regra de classificação dada por (2.39).

A escolha de k é a sugerida por Enas e Choi (1986) e descrita na secção 2.4.1. Quando a distância entre as observações é igual, a escolha da observação "mais próxima" é feita de forma aleatória.

[1]Em qualquer dos exemplos que se apresentam as taxas de erro obtidas com ou sem transformação das variáveis eram iguais.

3.1.1 Mulheres portadoras do gene da hemofilia

A distinção entre mulheres portadoras ou não de hemofilia é de grande importância, uma vez que a hemofilia é uma doença hereditária ligada aos cromossomas sexuais, que se transmite aos descendentes, manifestando-se principalmente nos de sexo masculino, sendo a probabilidade de atingir descendentes do sexo feminino muito pequena e geralmente mortal. Em diversos estudos efectuados [Zimmerman *et al.*,1971, Bouma *et al.*,1975] verificou-se que existem duas variáveis (FactorVIII-actividade e FactorVIII-antigene), obtidas em análises ao sangue, que são significativamente diferentes para as mulheres portadoras ou não de hemofilia. Poderá ainda existir informação de tipo probabilístico, obtida através das leis da hereditariedade, que indique qual a probabilidade de uma mulher ser portadora do gene que provoca a hemofilia. Nesse caso essa informação (probabilidades *a priori*) deve ser tida em atenção.

Os dados em análise dizem respeito a 22 mulheres portadoras e 30 mulheres não portadoras do gene que provoca a hemofilia, tendo sido tratados por Hermans e Habbema (1975) e encontram-se no Anexo A. A figura (3.1) representa os dois grupos no plano das variáveis transformadas log-FactorVIII-actividade e log-FactorVIII-antigene, que se apresentam razoavelmente separados.

As figuras (3.2) e (3.3) representam a estimativa da densidade de cada grupo, pelo método do produto do núcleo num conjunto de pontos equidistantes.

No quadro (3.1) apresentam-se as taxas de erro aparentes (*ea*), as de validação cruzada (*ec*) e as *"bootstrap"* (*eb*), obtidas por aplicação dos diferentes métodos a este conjunto de dados, em percentagem. Os parâmetros de alisamento utilizados para o método *Nucleo* foram os seguintes:

- Grupo das mulheres não portadoras do gene da hemofilia (A): $h = (0.283, 0.281)$;
- Grupo das mulheres portadoras do gene da hemofilia (B): $h = (0.287, 0.341)$.

Para o método *k-NN* utilizou-se $k = 3$.

Para obtenção das taxas de erro *"bootstrap"* foram utilizadas 100 réplicas, tanto neste exemplo como nos que se seguem.

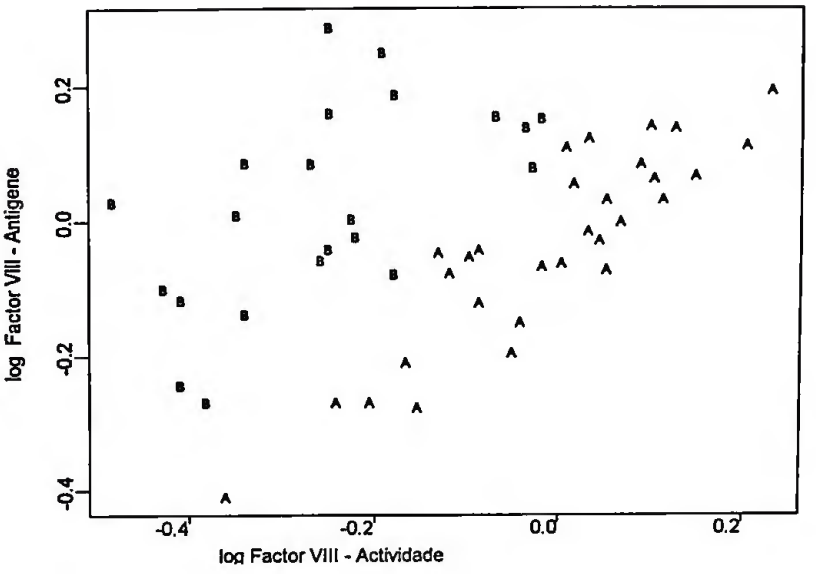


Figura 3.1: Dados bidimensionais do exemplo das mulheres portadoras do gene da hemofilia

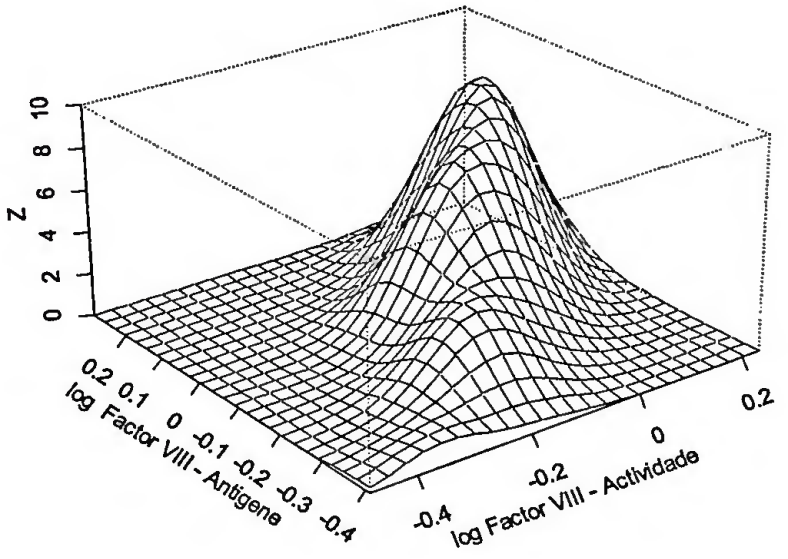


Figura 3.2: Função de densidade estimada do grupo A - Mulheres não portadoras do gene da hemofilia

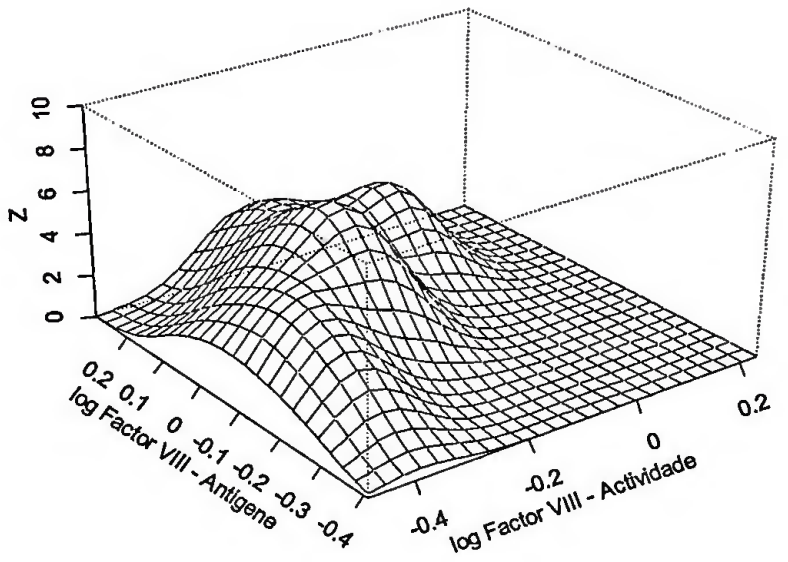


Figura 3.3: Função de densidade estimada do grupo B - Mulheres portadoras do gene da hemofilia

	Taxas de erro de má classificação								
Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ec_{(1)}$	$ec_{(2)}$	ec	$eb_{(1)}$	$eb_{(2)}$	eb
LDA	0.0%	9.1%	4.5%	0.0%	9.1%	4.5%	0.0%	10.6%	5.3%
QDA	3.3%	0.0%	1.7%	3.3%	4.5%	3.9%	5.3%	1.3%	3.3%
Nucleo	0.0%	13.6%	6.8%	3.3%	18.2%	10.8%	12.5%	22.1%	17.5%
k-NN	3.3%	9.1%	6.2%	6.7%	9.1%	7.9%	4.1%	13.9%	9.0%

Tabela 3.1: Hemofilia - Resultados considerando iguais probabilidades a priori

	Taxas de erro de má classificação								
Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ec_{(1)}$	$ec_{(2)}$	ec	$eb_{(1)}$	$eb_{(2)}$	eb
LDA	0.0%	9.1%	3.8%	0.0%	13.6%	5.8%	0.1%	10.1%	4.3%
QDA	0.0%	4.5%	1.9%	3.3%	4.5%	3.8%	1.6%	6.2%	3.6%
Nucleo	0.0%	13.6%	5.8%	3.3%	18.2%	10.8%	12.5%	22.1%	17.3%

Tabela 3.2: Hemofilia - Resultados após introdução das probabilidades a priori estimadas

Como se pode verificar as taxas de erro obtidas por validação cruzada nunca são inferiores às taxas de erro aparentes. As taxas de erro *"bootstrap"* são superiores às de validação cruzada, para todos os métodos, excepto para a discriminante quadrática. Os resultados apontam para uma vantagem na utilização da discriminante quadrática, cuja taxa de erro é bastante inferior à obtida pelos outros métodos. Dos métodos não paramétricos é o *k-NN* que apresenta taxas de erro mais baixas, sendo a diferença bastante elevada quando analisada em termos das taxas *"bootstrap"*.

Dado que os grupos têm dimensão diferente, considerou-se também a hipótese de estimar as probabilidades *a priori* por (1.3) e tomá-las como informação adicional. Neste caso estamos a considerar que os dados foram obtidos por amostragem mista. As taxas de erro obtidas são as que se apresentam no quadro (3.2). Como o programa utilizado para estimar as taxas de erro pelo método *k-NN* não permite a introdução de probabilidades *a priori* e não tendo sido possível alterá-lo, não se apresentam resultados para este método.

Comparando as taxas de erro assim obtidas com as anteriores verifica-se um agravamento das taxas de erro de validação cruzada obtidas a partir da discriminante linear, não havendo grande alteração nos outros dois métodos. Apesar das taxas de erro *"bootstrap"* serem ligeiramente superiores às obtidas anteriormente quando se utiliza a discriminante quadrática, os resultados continuam a evidenciar que existe vantagem na utilização deste método, quaisquer que sejam as taxas de erro em análise.

3.1.2 Dados Iris

Estes dados foram utilizados pela primeira vez por Fisher (1936) num artigo intitulado "The use of multiple measurements in taxonomic problems", tendo sido posteriormente utlizados por vários autores.

Foram considerados apenas os dois grupos inicialmente utilizados por Fisher, correspondentes às duas espécies observadas: *Iris setosa e Iris versicolor.* A amostra é composta por 50 observações de cada espécie, relativas a quatro variáveis: comprimento da sépala, largura da sépala, comprimento da pétala, largura da pétala.

Apesar dos dados serem quadridimensionais apresenta-se o gráfico (3.4) no plano constituído pelas variáveis comprimento da pétala, largura da pétala por estas serem suficientes para separarem completamente os dois grupos.[2]

Dada a boa separação dos dois grupos com base nas características observadas, não é de estranhar que a amostra tenha sido correctamente classificada por qualquer dos métodos utilizados, não se apresentando por essa razão os quadros resumo.

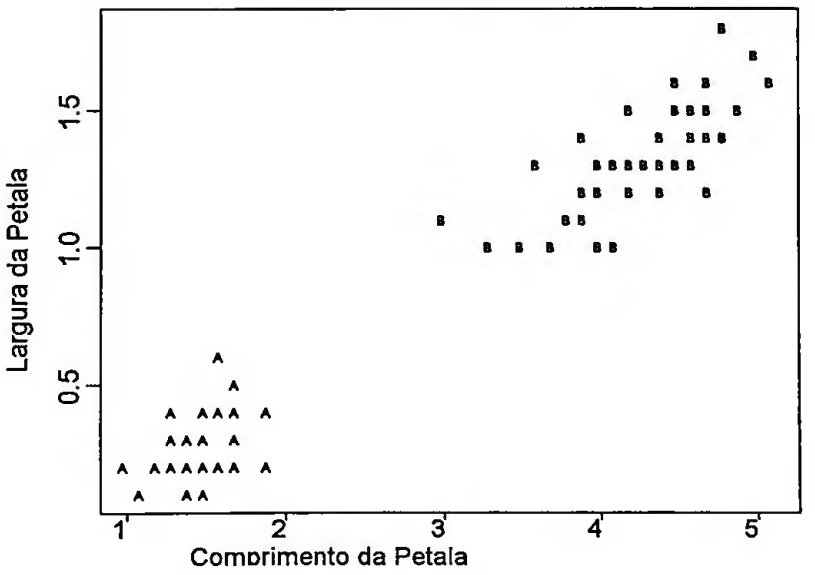


Figura 3.4: Dados do exemplo Iris

[2]Pires, A.M. (1995), cap.1, pag.6

3.1.3 Dados "Grupo"

Estes dados foram utilizados no estudo efectuado por Pires, A. M. (1995), para o qual foram geradas $n_1 = n_2 = 30$ observações com distribuição normal simetricamente 0.1 contaminada *(NSC)*

$$\mathbf{X}|G_i \sim 0.9N_2(\boldsymbol{\mu}_i, \mathbf{I}) + 0.1N_2(\boldsymbol{\mu}_i, 9\mathbf{I})$$

onde $\boldsymbol{\mu}_1 = (0,0)^T$ e $\boldsymbol{\mu}_2 = (4,0)^T$. A escolha deste conjunto de dados permite analisar o comportamento dos vários métodos numa situação de contaminação simétrica leve. O gráfico (3.5) permite-nos ver a maior dispersão existente no grupo B e que existem algumas observações do grupo A que estão mais próximas do grupo B que do seu grupo de origem.

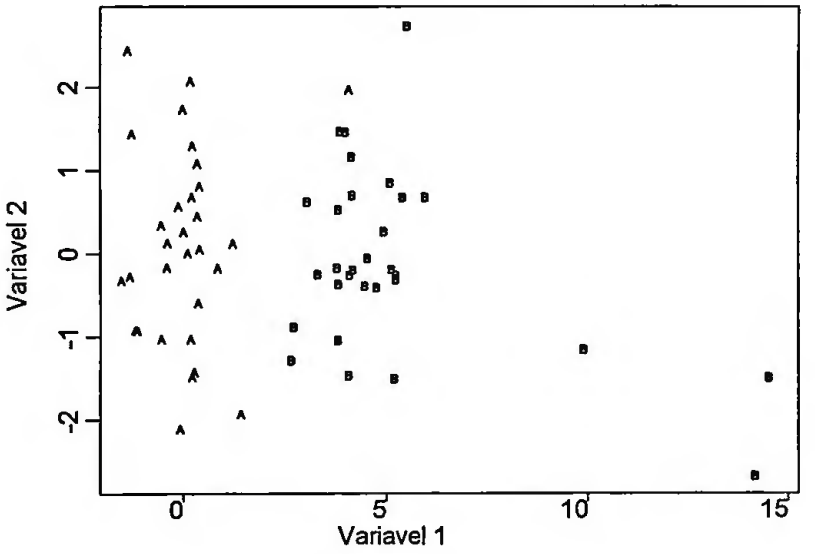


Figura 3.5: Observações bidimensionais do exemplo da NSC

As figuras (3.6) e (3.7) representam a estimativa da densidade de cada grupo, pelo método do produto do núcleo num conjunto de pontos equidistantes

Comparando os dois gráficos podemos ver que a distribuição das observações está mais dispersa no grupo B e que os dois grupos se encontram bem separados.

Dada a boa separação dos grupos, as taxas de erro das regras de classificação obtidas pelos diferentes métodos, que se encontram no quadro (3.3), são bastante baixas.

Os parâmetros de alisamento utilizados para o método *Nucleo* foram os seguintes:

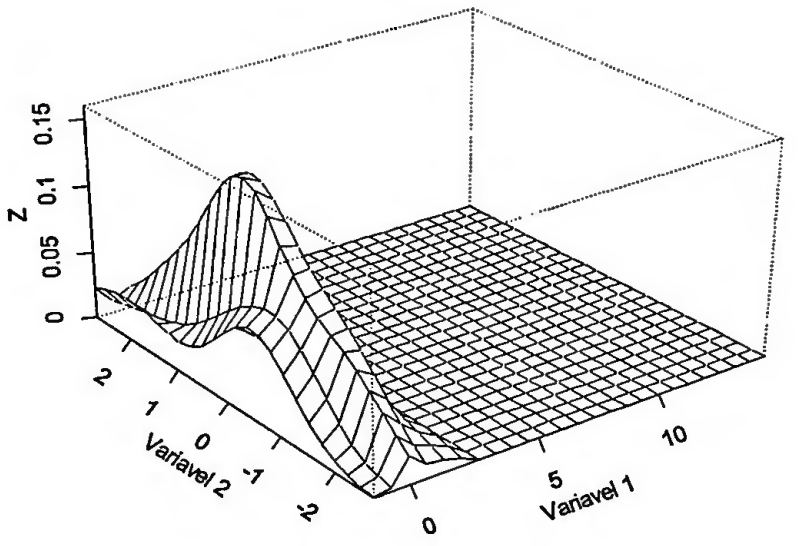


Figura 3.6: Função de densidade estimada para o grupo A, do exemplo da NSC

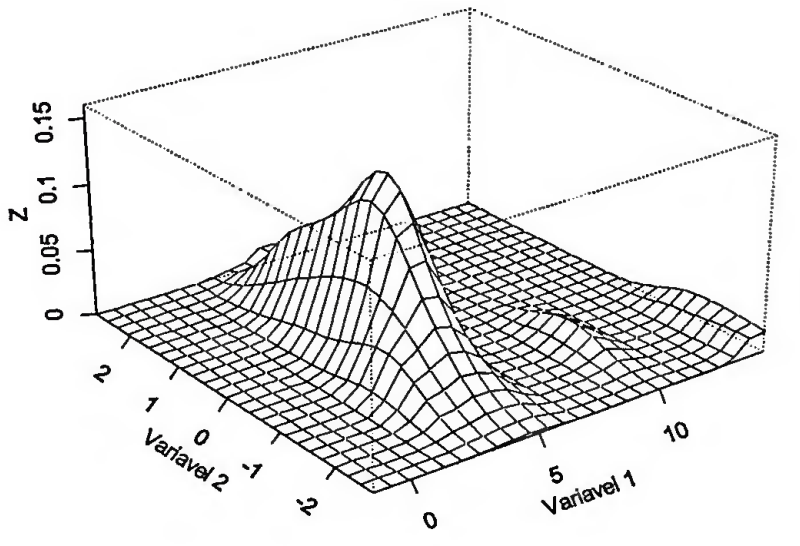


Figura 3.7: Função de densidade estimada para o grupo B, do exemplo da NSC

	Taxas de erro de má classificação								
Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ec_{(1)}$	$ec_{(2)}$	ec	$eb_{(1)}$	$eb_{(2)}$	eb
LDA	3.3%	6.7%	5.0%	3.3%	6.7%	5.0%	2.7%	7.4%	5.0%
QDA	3.3%	0.0%	1.7%	6.7%	0.0%	3.3%	3.9%	0.7%	2.3%
Nucleo	3.3%	0.0%	1.7%	6.7%	0.0%	3.3%	5.2%	0.6%	2.9%
k-NN	3.3%	0.0%	1.7%	3.3%	0.0%	1.7%	4.4%	-0.5%	2.0%

Tabela 3.3: Resultados do exemplo da NSC

- Grupo A: $h = (1.37, 2.54)$
- Grupo B: $h = (2.22, 2.30)$.

Para o método *k-NN* utilizou-se $k = 3$.

Neste exemplo é com o método dos k-vizinhos mais próximos, que adapta a quantidade de alisamento à densidade local dos dados, que se obtêm as taxas de erro mais baixas. No entanto os métodos *QDA* e *Nucleo* apresentam também um bom desempenho. Note-se que, o valor das taxas de erro *"bootstrap"* para estes dois métodos se encontra entre o das taxas de erro aparentes e o das taxas de erro de validação cruzada e que todas as taxas de erro são sensivelmente iguais para os outros dois métodos. O melhor desempenho do método *QDA* em relação ao *LDA* deve-se ao facto dos dois grupos apresentarem matrizes de covariâncias empíricas bastante diferentes.

3.1.4 Dados "Outro"

Tal como no exemplo anterior os dados foram utilizados no estudo efectuado por Pires, A.M. (1995). Para este exemplo e para o primeiro grupo, foram geradas 90 observações com distribuição $N_2(\mathbf{0}, \mathbf{I})$ a que se juntaram 10 observações no ponto (10,10).

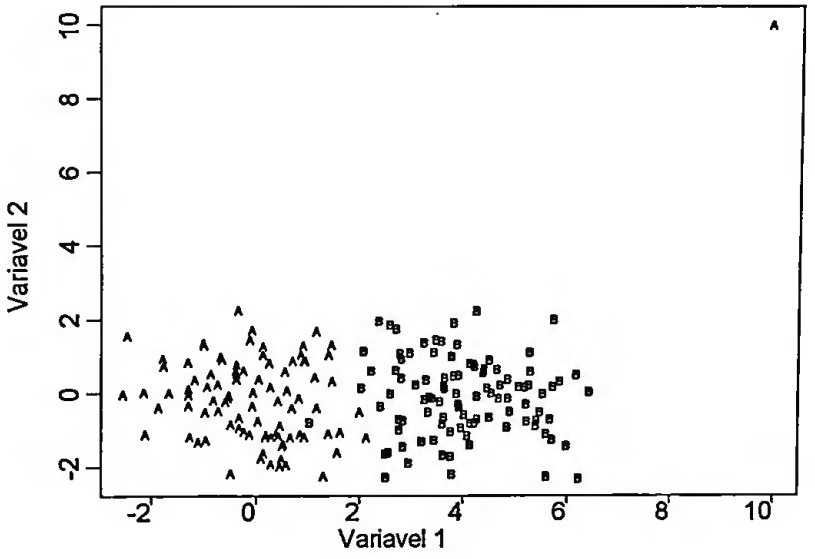


Figura 3.8: Dados bidimensionais do exemplo da Normal com "outliers"

O objectivo de juntar estas últimas observações, que são evidentes no gráfico (3.8), é simular a ocorrência de erros ou a observação involuntária de elementos de outra população. Para o segundo grupo geraram-se 100 observações com distribuição $N_2(\mu_i, \mathbf{I})$, onde $\mu_2 = (4, 0)^T$.

No gráfico (3.9) também é clara a presença dos *"outliers"* no grupo A. O gráfico (3.10) apresenta a estimativa da densidade para o grupo B.

Neste exemplo, cujos resultados se encontram no quadro (3.4), é significativa a diferença existente entre as taxas de erro obtidas pelos métodos paramétricos e não paramétricos.

Os parâmetros de alisamento utilizados para o método *Nucleo* foram os seguintes:

- Grupo A: $h = (1.97, 2.54)$
- Grupo B: $h = (1.89, 1.75)$.

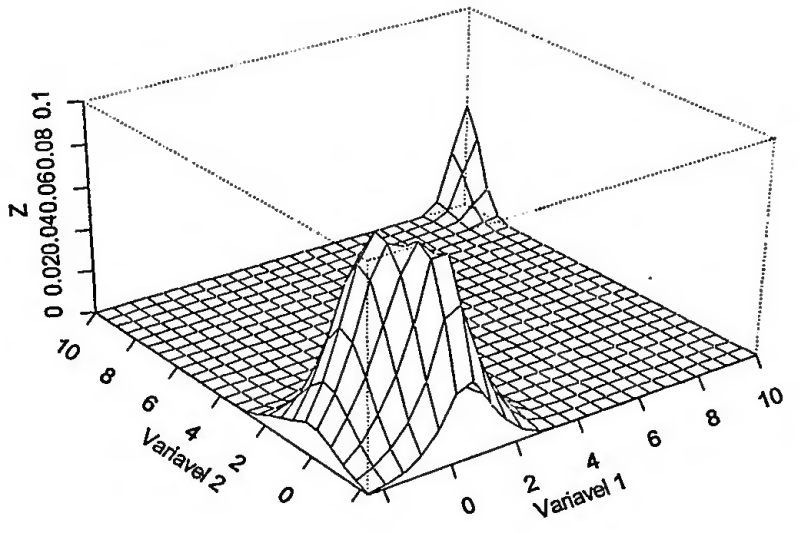


Figura 3.9: Função de densidade estimada para o grupo A, do exemplo com *"outliers"*

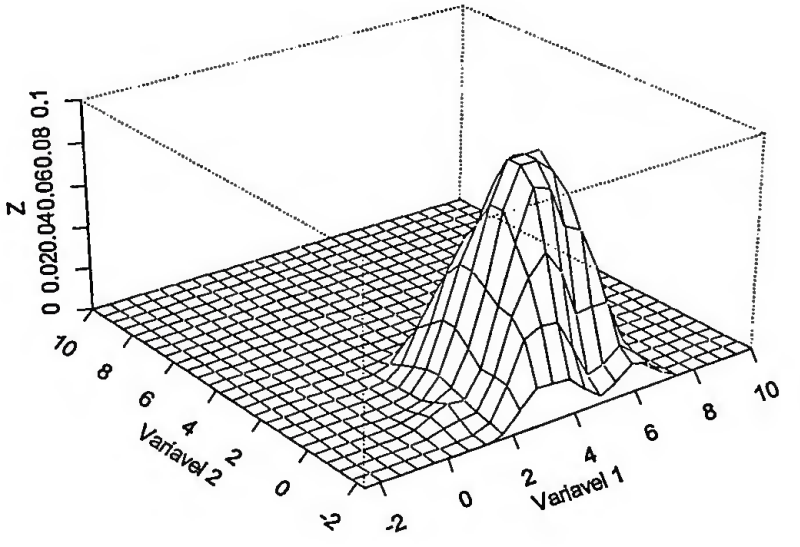


Figura 3.10: Função de densidade estimada para o grupo B, do exemplo com *"outliers"*

	Taxas de erro de má classificação								
Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ec_{(1)}$	$ec_{(2)}$	ec	$eb_{(1)}$	$eb_{(2)}$	eb
LDA	8.0%	12.0%	10.0%	8.0%	13.0%	10.5%	9.6%	12.1%	10.9%
QDA	2.0%	9.0%	5.5%	2.0%	9.0%	5.5%	2.3%	9.2%	5.7%
Nucleo	0.0%	1.0%	0.5%	1.0%	2.0%	1.5%	2.1%	1.7%	1.9%
k-NN	1.0%	1.0%	1.0%	2.0%	1.0%	1.5%	1.3%	0.8%	1.1%

Tabela 3.4: Resultados do exemplo com 'outliers'

Para o método *k-NN* utilizou-se $k = 7$.

Dado que os dados têm distribuição normal, situação em que a discriminante linear é óptima, o fraco desempenho desta relativamente às obtidas pelos outros métodos deve-se sobretudo à presença dos *"outliers"*. Com o método do núcleo obtêm-se as taxas de erro mais baixas, mas também o método dos k-vizinhos mais próximos tem um bom desempenho.

3.2 Aspectos computacionais

Os programas foram escritos em linguagem S, tendo sido utilizado nalguns casos partes de programas já existentes na biblioteca do S-Plus, encontrando-se no Anexo B os mais representativos.

Métodos paramétricos

Para o método da discriminante linear *(LDA)* são utilizados os programas: *lda, predict.lda e jackk.lda e tboot.lda.* Com estes programas obtem-se a regra discrimininante, as taxas de erro aparentes, as taxas de erro por validação cruzada (*"leaving-one-out"*) e as taxas de erro *"bootstrap"*.

O funcionamento deste conjunto de programas é o seguinte:

Com o programa *lda,* onde x=matriz da amostra e x.lab=vector que indica a dimensão dos grupos, obtêm-se as estimativas dos coeficientes:

o <-lda(x, x.lab).

Com o programa *predict.lda,* obtem-se a discriminante ou discriminantes (*dimen=nº de grupos G_i menos* 1*),* entrando ou não com informação adicional (probabilidades *a priori*), e utilizando essa discriminante classifica-se de novo a amostra para obter as taxas de erro aparentes:

- predict.lda(o, x, x.lab, dimen=2,prior)

O programa *jackk.lda* junta os dois anteriores, fazendo a separação da amostra em duas subamostras, uma com $n-1$ observações da amostra com a qual se obtem a regra discriminante e outra com 1 observação que vai ser classificada utilizando a regra anteriormente obtida, calculando depois as taxas de erro de validação cruzada:

- jackk.lda(x. x.lab, dimen, prior, dim1, dim2) onde dim1=n1, dim2=n2.

O programa *tboot.lda* calcula as taxas aparentes e extrai uma nova amostra, com reposição, a partir da amostra inicial, para amostragem mista (*s=1*) ou separada, com a qual é obtida uma nova regra discriminante e são calculadas as taxas de erro desta

nova amostra e da amostra inicial, sendo depois efectuados os cálculos das estimativas *"bootstrap"*:

- tboot.lda(x.x.lab,prior,dim1,dim2).

O conjunto de programas para obter a regra da discriminante quadrática *(QDA)* e as respectivas taxas de erro funcionam de forma semelhante ao anteriormente descrito:

- o<-qda(x, x.lab)
- predict.lda(o, x, x.lab)
- jackk.qda(x.x.lab,prior,dim1,dim2) onde dim1=n1, dim2=n2
- tboot.qda(x.x.lab,prior,dim1,dim2).

Métodos não paramétricos

Os programas utilizados para dados bidimensionais e para dois grupos, são os seguintes:

Para o método do produto núcleo é estimada a densidade de cada observação no espaço de cada um dos grupos e calculado o rácio das densidades em cada grupo para obter a respectiva regra descriminante.

- tkde2r(x,y,x1,y1,xr,yr,group,g=2) sendo necessário especificar x=variável 1 do grupo 1, y=variável 2 do grupo 1, x1=variável 1 do grupo 2, y=variável 2 do grupo 2, xr e yr conjunto de pontos do espaço em que se quer estimar a densidade. Se g=2, são calculadas as taxas de erro aparentes.

Para o cálculo das taxas de erro de validação cruzada*("leaving-one-out")*, é estimada a função de densidade da i-ésima observação no espaço de todas as obsevações, excepto a i-ésima, para cada um dos grupos e calculado o rácio das densidades da observação nos espaços definidos.

- tkde2r.cv(x,y,x1,y1,xr,yr,group,g=2).

O cálculo das taxas de erro *"bootstrap"* é feito de forma semelhante ao que foi descrito para o caso paramétrico. Para o cálculo das observações iniciais mal classifcadas pela regra discriminante obtida após reamostragem é estimada a densidade das observações iniciais no espaço das observações após reamostragem, para cada um dos grupos.

- tboot.kde2r(x,y,x1,y1,xr,yr,group,g=2).

De forma semelhante foram feitos os programas para o caso de dados quadridimensionais (tkde4r e tkde4r.cv e tboot.kde4r).

Para o método dos k-vizinhos mais próximos foram utilizados dois programas disponíveis na Internet,[3] um contributo de Brian Ripley que contem várias funções em linguagem S para classificação.[4] O algoritmo utilizado é o descrito no capítulo Análise discriminante não paramétrica - Regras de classsificação. Os programas são:

- knna(train=x,test=x,x.lab,k)
- knna.cv(train=x,x.lab,k)
- tboot.knn(train=x,test=x,x.lab,k).

3.3 Experiências de simulação

Não sendo possível tirar conclusões acerca do desempenho dos vários métodos considerados apenas com os exemplos anteriores e pretendendo-se ter uma indicação sobre as situações em que há, ou não, vantagem na utilização dos métodos não paramétricos para posterior investigação, recorreu-se à simulação para avaliar e comparar cada um desses métodos e assim obter essa indicação.

Os métodos que se pretende avaliar e comparar são assim, os mesmos que já foram utilizados para obter uma regra discriminante nos exemplos ilustrativos, estando as hipóteses para obter essa regra descritas na secção 3.1.

Este estudo é necessariamente restritivo dado que se consideram apenas duas variáveis, dois grupos, iguais probabilidades *a priori* e iguais custos e que existem muitas outras possibilidades para obter uma regra discriminante. Espera-se, no entanto, que seja útil para avaliar o desempenho dos principais métodos não paramétricos considerados no capítulo 2.

Para poder verificar um número razoável de situações no espaço de tempo que se dispõe, tomaram-se em conta alguns factores que, entre outros, podem influenciar o desempenho dos vários métodos, nomeadamente a distribuição de $\mathbf{X}$ condicional aos

[3] A partir do endereço statlib@lib.stat.cmu.edu

[4] O endereço do autor destes programas é ripley@stats.ox.ac.uk

grupos (G_i, $i = 1, 2$), e respectiva localização ($\boldsymbol{\mu}_i$) e dispersão ($\boldsymbol{\Sigma}_i$) e ainda o número de observações de cada grupo na amostra de treino, n_1 e n_2.

Em relação à dimensão da amostra optou-se por considerar três situações: $n_1 = n_2 = 50$, $n_1 = n_2 = 100$ e $n_1 = 25$ e $n_2 = 75$, para poder observar o efeito da dimensão total da amostra de treino e o efeito da diferença de dimensão dos dois grupos.

Para gerar as observações das amostras de treino consideraram-se as seguintes distribuições:

(a) Normal (NOR), $\mathbf{X}\,|G_i \sim N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, em que a separação dos dois grupos seja de nível médio ($\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{2}$) e com $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$;

(b) Normal simetricamente contaminada *(NSC)*

$$\mathbf{X}|G_i \sim 0.9N_2(\boldsymbol{\mu}_i, \mathbf{I}) + 0.1N_2(\boldsymbol{\mu}_i, 9\mathbf{I})$$

onde ($\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{2}$), uma situação de contaminação simétrica leve.

(c) t- Student, ($\mathbf{T}(3)$), com 3 graus de liberdade, uma distribuição simétrica mas com abas mais pesadas;

(d) Normal assimetricamente contaminada *(NAC)*, em que 90% das n_i ($i = 1, 2$), observações são normais, $\mathbf{X}\,|G_i \sim N_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, com ($\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{2}$) e os restantes 10%, $\mathbf{x}_{ij}$, estão concentradas no mesmo ponto, sendo $\mathbf{x}_{1j} = (9, 10)$ e $\mathbf{x}_{2j} = (10, 9)$.

Estas distribuições são as mesmas que foram utilizadas num estudo de robustez em análise discriminante desenvolvido por Pires, A. M. (1995).

Com estas quatro opções pretende-se, por um lado, testar os métodos com situações mais ou menos representativas da realidade e, por outro, cobrir situações em que os vários métodos apresentam pontos mais fracos ou mais fortes. O número de réplicas utilizado em cada caso foi de 500.

A comparação dos vários métodos vai ser efectuado através de algumas taxas de erro de má classificação. Dado que as taxas de erro aparentes tendem a produzir valores optimistas, uma vez que se estão a classificar as observações com base na regra discriminante

que foi obtida a partir dessas mesmas observações, estimaram-se também as taxas de validação cruzada que, segundo Efron(1983), reduzem substancialmente o enviesamento em relação às taxas aparentes, apesar de apresentarem uma grande variabilidade.

Considerou-se também a hipótese de estimar as taxas *"bootstrap"*, uma vez que estas, além de reduzirem o enviesamento, têm também uma menor dispersão, mas como o procedimento era muito dispendioso em termos computacionais e era impossível estimá-las em tempo útil, não foi possível apresentá-las.

Assim, para cada réplica, registou-se o valor das taxas de erro aparentes e de validação cruzada, para ambos os grupos, sendo a estimativa da taxa de erro total $e(\cdot) = (e(1) + e(2))/2$.

A informação resultante encontra-se resumida através de duas estatísticas habitualmente utilizadas: a média e o desvio padrão.

Uma vez que as distribuições subjacentes aos dados são conhecidas, foram calculados os valores de e_{opt} para cada uma delas, que se apresentam em (3.1) e calculados os rácios $ea_{(\cdot)}/e_{opt(\cdot)}$ e $ec_{(\cdot)}/e_{opt(\cdot)}$, de forma a tornar os valores obtidos mais comparáveis, pois não é possível fazer comparações directas entre resultados para distribuições com taxas óptimas diferentes.

Distribuição	e_{opt}
NOR	0.0786
T(3)	0.1261
NSC	0.1026
NAC	0.0786

(3.1)

3.3.1 Análise dos resultados

Nesta secção faz-se a análise dos resultados obtidos, procurando detectar quais as situações em que os diferentes métodos têm vantagens. Nas páginas seguintes apresentam-se os quadros que resumem os resultados das simulações e também os gráficos dos rácios mencionados na secção anterior.

1. **Taxas de erro aparentes *versus* taxas de erro por validação cruzada:**

   As médias das taxas de erro aparentes, independentemente do método utilizado, da dimensão da amostra e da distribuição, são sempre inferiores às médias das taxas de validação cruzada. É de realçar que essa situação é mais evidente no caso dos métodos não paramétricos, o que não é de estranhar dado que, como já se disse, as taxas de erro aparentes tendem a ser optimistas e é natural que o sejam mais quando se utilizam métodos que estão completamente dependentes dos dados.

   O mesmo acontece com o desvio padrão, excepto em quatro situações:

a) para a distribuição NAC e método QDA, quando $n_1 = n_2 = 50$ e $n_1 = 25$, $n_2 = 75$;

b) para a distribuição $\mathbf{T}(3)$, e método *Núcleo*, quando $n_1 = 25$, $n_2 = 75$;

c) para a distribuição NAC e método LDA, quando $n_1 = n_2 = 100$.

   Se, em vez do desvio padrão, olharmos para o coeficiente de variação verifica-se que, na maioria dos casos, é idêntico para as médias das taxas de erro aparentes e das taxas de erro de validação cruzada

2. **Efeito de diferença de dimensão dos grupos:**

   Para o mesmo valor de $n_1 + n_2 = 100$, uma situação que parece afectar muito o desempenho dos métodos não paramétricos, sobretudo o método do produto do núcleo, é a diferença de dimensão dos grupos. Nestes casos a percentagem de más classificações do grupo de menor dimensão é muito elevada, sendo bastante baixa no outro grupo. A variabilidade das taxas de erro é também muito maior, tanto para os grupos, como para o total.

   Também o desempenho dos métodos paramétricos é afectado por esta situação, embora muito menos que os não paramétricos. O único caso em que se verifica a situação inversa é com o método LDA em NAC.

3. **Efeito do aumento de dimensão da amostra:**

   Quando a dimensão da amostra aumenta as médias de ea aproximam-se de e_{opt}, excepto quando se aplicam os métodos clássicos à distribuição *NAC.*

   As taxas de erro de validação cruzada mostram-nos que todos os métodos melhoram o seu desempenho, como se pode ver pela sua média e pela redução da variabilidade. Esta situação é mais evidente nos métodos não paramétricos, que se aproximam mais da taxa óptima, sugerindo uma maior adequabilidade dos mesmos para amostras de grande dimensão.

   Quando se utilizam os métodos clássicos na distribuição *NAC,* apesar das médias das taxas de erro baixarem, o mesmo não acontece ao coeficiente de variação que não se altera no método *LDA* e aumenta no método *QDA.*

4. **Efeito da distribuição de X condicional aos grupos:**

   Para a distribuição *NOR* são os estimadores clássicos que conduzem a melhores resultados, notando-se uma grande diferença entre estes e os não paramétricos, que se agrava quando a dimensão dos grupos não é igual. Entre os não paramétricos, o método do núcleo apresenta taxas de erro mais baixas, excepto quando a dimensão dos grupos é diferente.

   Quando a distribuição é *NSC,* considerando os rácios antes mencionados, todos os métodos são menos eficientes que no caso anterior, como seria de esperar, continuando em vantagem os estimadores clássicos. Destes é o método *LDA* que apresenta os melhores resultados, enquanto que, para os não paramétricos, é o método dos k-vizinhos mais próximos que tem as taxas de erro mais baixas.

   Na presença de *"outliers"* (distribuição *NAC*) são os estimadores não paramétricos que têm um melhor desempenho, sendo a diferença entre os estimadores clássicos e não paramétricos maior que nos casos anteriores. Para as amostras de maior dimensão e para amostras cujos grupos têm dimensão diferente os métodos *k-NN* e do *Núcleo* estão bastante próximos, mas há uma vantagem nítida na utilização

do *Núcleo* quando a amostra é mais pequena e tem $n_1 = n_2$. Dos métodos clássicos o *QDA* apenas tem vantagem quando $n_1 \neq n_2$.

Em presença de uma distribuição de caudas pesadas, $\mathbf{T}(3)$, o método *LDA* é, entre todos os estimadores utilizados, o que tem um rácio mais baixo. Os métodos *k-NN* e do *Núcleo* aproximam-se bastante quando a dimensão da amostra é maior, mas há vantagem na utilização do método *k-NN*, como se esperava.

- **Em resumo:**

  Existe vantagem em utilizar as taxas de erro de validação cruzada em vez das aparentes, principalmente nos métodos não paramétricos, para avaliar o desempenho das regras de classificação.

  Nas condições destas experiências, em que os grupos têm uma separação média e que as matrizes de covariâncias não são, em geral, muito diferentes, os métodos paramétricos têm um melhor desempenho, apesar de se notar um agravamento das respectivas taxas de erro para as distribuições não normais. A única excepção ocorre quando em presença de *"outliers"*, situação em que é mais indicado utilizar o método do *Núcleo*.

  Entre os dois métodos não paramétricos é clara a vantagem da utilização do método *k-NN* quando a dimensão dos dois grupos é muito diferente, qualquer que seja a distribuição subjacente aos dados. Quando a distribuição é *NOR* ou *NAC* e a dimensão dos grupos é igual, o método do *Núcleo* tem um melhor desempenho. Se a distribuição for *NSC* ou $\mathbf{T}(3)$ o método *k-NN* obtem melhores resultados, principalmente se a dimensão da amostra não for muito grande.

$n_1 = 50$	$n_2 = 50$	Média taxas de erro aparentes (ea)			Desvio padrão das taxas (ea)			Rácio ea/e_{opt}		
Distribuições	Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ea_{(1)}$	$ea_{(2)}$	ea	$ea_{(1)}$	$ea_{(2)}$	ea
	LDA	7.5	7.5	7.5	3.4	3.6	3.5	0.96	0.95	0.95
NOR	QDA	7.5	7.5	7.5	3.1	3.3	3.2	0.95	0.95	0.95
	Núcleo	6.0	5.9	5.9	3.5	3.5	3.5	0.76	0.75	0.76
	k-NN	6.8	6.7	6.8	3.5	3.6	3.5	0.87	0.85	0.86
	LDA	10.3	10.0	10.2	3.9	3.7	3.8	1.00	0.97	0.99
NSC	QDA	10.1	10.0	10.0	3.8	3.7	3.8	0.98	0.97	0.99
	Núcleo	7.0	7.2	7.1	3.6	3.8	3.7	0.69	0.70	0.69
	k-NN	8.9	8.9	8.9	4.0	3.8	3.9	0.87	0.86	0.86
	LDA	14.5	37.8	26.1	5.1	5.2	12.7	1.84	4.81	3.32
NAC	QDA	6.7	28.7	17.7	5.5	7.1	12.7	0.85	3.65	2.25
	Núcleo	5.9	6.1	6.0	3.4	3.7	3.6	0.75	0.78	0.76
	k-NN	6.3	6.2	6.2	3.4	3.0	3.2	0.81	0.78	0.79
	LDA	13.2	12.9	13.0	4.4	4.6	4.5	1.05	1.02	1.03
$T(3)$	QDA	12.9	12.5	12.7	4.6	4.3	4.4	1.03	1.02	1.03
	Núcleo	8.3	8.6	8.5	4.1	4.2	4.1	0.66	0.68	0.67
	k-NN	10.8	11.0	10.9	4.4	4.3	4.4	0.86	0.87	0.87

Tabela 3.5: Taxas de erro aparentes, por grupo e totais, para n_1=n_2=50

$n_1 = 50$	$n_2 = 50$	Média das taxas de erro de validação cruzada (ec)			Desvio padrão das taxas (ec)			Rácio ec/e_{opt}		
Distribuições	Métodos	$ec_{(1)}$	$ec_{(2)}$	ec	$ec_{(1)}$	$ec_{(2)}$	ec	$ec_{(1)}$	$ec_{(2)}$	ec
	LDA	8.1	8.0	8.0	3.5	3.6	3.6	1.02	1.01	1.02
NOR	QDA	8.1	8.3	8.2	3.4	3.4	3.4	1.03	1.05	1.04
	Núcleo	8.9	8.8	8.8	4.1	4.2	4.2	1.13	1.12	1.12
	k-NN	9.1	9.4	9.3	4.1	4.3	4.2	1.16	1.20	1.18
	LDA	10.9	10.6	10.8	4.0	3.9	4.0	1.07	1.03	1.05
NSC	QDA	11.4	11.2	11.3	3.9	4.0	4.0	1.11	1.10	1.10
	Núcleo	12.1	11.6	12.0	5.2	4.9	5.0	1.21	1.13	1.17
	k-NN	11.5	11.4	11.5	4.4	4.6	4.5	1.12	1.11	1.12
	LDA	15.7	40.3	28.0	5.3	5.5	13.4	2.00	5.13	3.66
NAC	QDA	15.4	33.0	24.2	7.7	7.4	11.6	1.96	4.20	3.08
	Núcleo	8.1	8.2	8.1	4.2	4.3	4.2	1.03	1.04	1.03
	k-NN	9.0	9.1	9.1	4.1	3.9	4.0	1.15	1.16	1.16
	LDA	13.8	13.5	13.7	4.4	4.6	4.6	1.09	1.07	1.09
$T(3)$	QDA	14.5	14.1	14.3	4.8	4.6	4.7	1.15	1.12	1.13
	Núcleo	14.7	15.1	14.9	5.8	5.9	5.9	1.17	1.20	1.18
	k-NN	14.2	14.2	14.2	5.4	4.9	5.1	1.12	1.13	1.13

Tabela 3.6: Taxas de erro de validação cruzada, por grupo e totais, para n_1=n_2=50

$n_1 = 25$	$n_2 = 75$	Média taxas de erro aparentes (ea)			Desvio padrão das taxas (ea)			Rácio ea/e_{opt}		
Distribuições	Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ea_{(1)}$	$ea_{(2)}$	ea	$ea_{(1)}$	$ea_{(2)}$	ea
	LDA	7.0	7.4	7.2	4.5	3.0	3.9	0.89	0.94	0.92
NOR	QDA	7.1	7.2	7.2	4.2	3.0	3.7	0.90	0.92	0.91
	Núcleo	23.0	0.4	11.7	10.6	0.7	13.6	2.93	0.04	1.48
	k-NN	11.6	2.5	7.1	6.3	1.8	6.5	1.48	0.32	0.90
	LDA	10.2	10.4	10.3	5.4	3.8	4.7	0.99	1.01	1.00
NSC	QDA	10.2	9.9	10.1	5.3	3.9	4.6	0.99	0.97	0.98
	Núcleo	27.5	0.6	14.1	12.3	0.9	16.0	2.68	0.06	1.37
	k-NN	14.5	4.1	9.3	7.2	2.1	7.5	1.41	0.40	0.91
	LDA	10.9	32.5	21.7	4.1	4.6	11.6	1.39	4.13	2.76
NAC	QDA	7.7	28.6	18.2	6.2	6.9	12.3	0.98	3.64	2.32
	Núcleo	23.7	0.5	12.1	10.5	0.7	13.8	3.02	0.06	1.54
	k-NN	18.9	2.4	10.6	5.9	1.7	9.3	2.40	0.31	1.36
	LDA	12.8	12.9	12.9	6.2	4.2	5.3	1.02	1.02	1.02
$T(3)$	QDA	13.0	11.9	12.4	5.7	4.3	5.1	1.03	0.94	0.99
	Núcleo	32.1	0.8	16.5	13.6	1.0	18.4	2.55	0.06	1.31
	k-NN	20.7	5.6	13.2	8.5	2.5	9.8	1.64	0.45	1.04

Tabela 3.7: Taxas de erro aparentes, por grupo e totais, para n_1=25, n_2=75

$n_1 = 25$	$n_2 = 75$	Média das taxas de erro de validação cruzada (ec)			Desvio padrão das taxas (ec)			Rácio ec/e_{opt}		
Distribuições	Métodos	$ec_{(1)}$	$ec_{(2)}$	ec	$ec_{(1)}$	$ec_{(2)}$	ec	$ec_{(1)}$	$ec_{(2)}$	ec
	LDA	7.9	7.8	7.9	4.7	3.2	4.0	1.01	0.99	1.01
NOR	QDA	8.9	7.7	8.3	4.6	3.1	4.0	1.13	0.98	1.06
	Núcleo	27.6	1.9	14.8	11.6	1.6	15.3	3.54	0.24	1.88
	k-NN	19.1	4.7	11.9	8.1	2.4	9.4	2.43	0.59	1.51
	LDA	11.2	11.0	11.1	5.6	3.8	5.6	1.09	1.07	1.08
NSC	QDA	13.0	10.8	11.9	5.5	4.1	5.0	1.27	1.05	1.16
	Núcleo	35.8	4.1	19.9	14.0	2.3	18.8	3.49	0.40	1.94
	k-NN	22.5	6.4	14.5	8.8	2.9	10.4	2.20	0.62	1.41
	LDA	12.9	34.2	23.6	4.9	4.9	11.7	1.64	4.35	3.00
NAC	QDA	18.0	30.9	24.5	6.3	6.9	9.2	2.29	3.93	3.12
	Núcleo	28.9	1.7	15.3	11.9	1.5	16.1	3.68	0.22	1.95
	k-NN	25.8	4.4	15.1	8.1	2.5	12.3	3.28	0.56	1.92
	LDA	14.0	13.4	13.7	6.4	4.2	5.4	1.11	1.06	1.09
$T(3)$	QDA	16.6	12.8	14.7	6.3	4.5	5.8	1.32	1.02	1.17
	Núcleo	28.8	9.6	19.2	9.8	4.3	10.4	2.27	0.76	1.52
	k-NN	26.1	7.4	16.7	9.9	2.9	11.8	2.07	0.59	1.33

Tabela 3.8: Taxas de erro de validação cruzada, por grupo e totais, para n_1=25, n_2=75

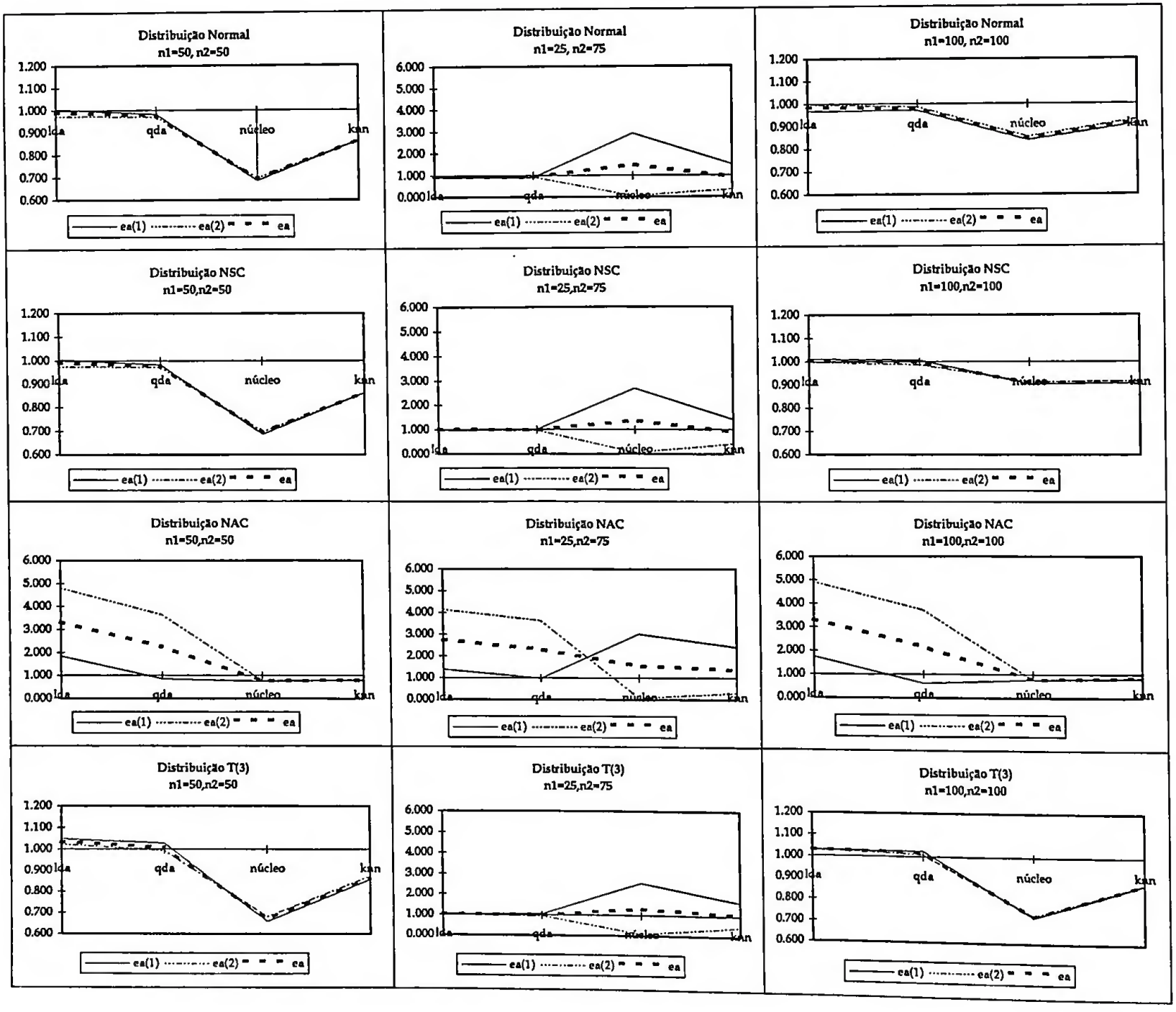
$n_1 = 100$	$n_2 = 100$	Média das taxas de erro aparentes (ea)			Desvio padrão das taxas (ea)			Rácio ea/e_{opt}		
Distribuições	Métodos	$ea_{(1)}$	$ea_{(2)}$	ea	$ea_{(1)}$	$ea_{(2)}$	ea	$ea_{(1)}$	$ea_{(2)}$	ea
	LDA	7.6	7.9	7.7	2.5	2.5	2.5	0.97	1.00	0.98
NOR	QDA	7.6	7.8	7.7	2.3	2.3	2.3	0.97	0.99	0.98
	Núcleo	6.6	6.7	6.6	2.5	2.7	2.6	0.84	0.85	0.85
	k-NN	7.2	7.3	7.2	2.6	2.5	2.6	0.91	0.93	0.92
	LDA	10.4	10.2	10.3	2.8	2.8	2.8	1.01	1.00	1.00
NSC	QDA	10.3	10.1	10.2	2.9	2.8	2.8	1.01	0.99	1.00
	Núcleo	9.3	9.4	9.3	2.9	2.9	2.9	0.90	0.91	0.91
	k-NN	9.3	9.4	9.3	2.9	2.9	2.9	0.91	0.92	0.91
	LDA	13.7	38.6	26.1	3.7	3.7	13.6	1.74	4.91	3.32
NAC	QDA	4.8	29.3	17.0	3.8	6.9	13.4	0.61	3.73	2.16
	Núcleo	6.1	6.1	6.1	2.6	2.7	2.6	0.78	0.78	0.78
	k-NN	6.4	6.5	6.5	2.5	2.5	2.5	0.81	0.83	0.82
	LDA	12.9	13.0	13.0	3.3	3.3	3.3	1.03	1.03	1.03
$T(3)$	QDA	12.9	12.7	12.8	3.4	3.4	3.4	1.03	1.01	1.02
	Núcleo	9.1	9.2	9.2	3.2	3.2	3.2	0.72	0.73	0.73
	k-NN	11.0	11.1	11.0	3.1	2.8	3.0	0.87	0.88	0.88

Tabela 3.9: Taxas de erro aparentes, por grupo e totais, para n_1=100, n_2=100

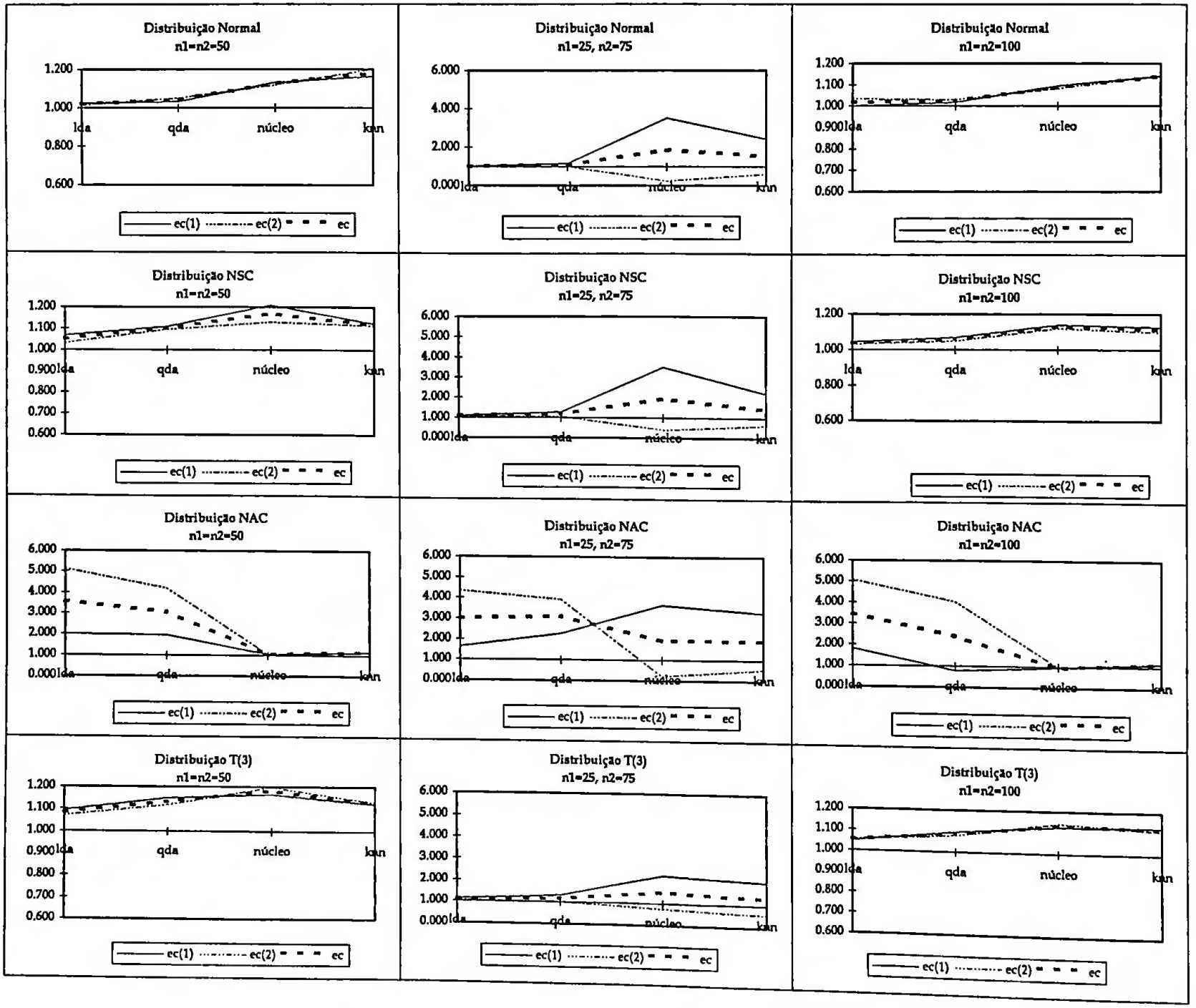
$n_1 = 100$	$n_2 = 100$	Média das taxas de erro de validação cruzada (ec)			Desvio padrão das taxas (ec)			Rácio ec/e_{opt}		
Distribuições	Métodos	$ec_{(1)}$	$ec_{(2)}$	ec	$ec_{(1)}$	$ec_{(2)}$	ec	$ec_{(1)}$	$ec_{(2)}$	ec
	LDA	7.9	8.1	8.0	2.6	2.5	2.5	1.00	1.03	1.02
NOR	QDA	8.0	8.1	8.1	2.4	2.4	2.4	1.02	1.03	1.03
	Núcleo	8.6	8.5	8.5	3.0	3.0	3.0	1.09	1.08	1.08
	k-NN	9.0	9.0	9.0	3.0	2.9	3.0	1.14	1.14	1.14
	LDA	10.7	10.6	10.7	2.9	2.8	2.8	1.04	1.03	1.04
NSC	QDA	11.0	10.8	10.9	3.0	2.9	2.9	1.07	1.05	1.06
	Núcleo	11.7	11.5	11.6	3.6	3.6	3.6	1.14	1.12	1.13
	k-NN	11.6	11.3	11.4	3.3	3.3	3.3	1.13	1.10	1.11
	LDA	14.2	39.9	27.0	3.9	3.7	13.4	1.81	5.08	3.44
NAC	QDA	6.1	31.8	18.9	4.2	6.7	14.0	0.78	4.05	2.40
	Núcleo	7.4	7.6	7.5	2.7	3.1	2.9	0.94	0.97	0.95
	k-NN	8.9	9.1	9.0	2.9	2.9	2.9	1.13	1.16	1.15
	LDA	13.2	13.3	13.3	3.4	3.3	3.4	1.05	1.05	1.05
$T(3)$	QDA	13.8	13.6	13.7	3.5	3.4	3.5	1.09	1.08	1.09
	Núcleo	14.2	14.4	14.3	4.1	4.1	4.1	1.13	1.14	1.13
	k-NN	14.2	14.0	14.1	3.6	3.6	3.6	1.13	1.11	1.12

Tabela 3.10: Taxas de erro de validação cruzada, por grupo e totais, para n_1=100, n_2=100

Gráfico do rácio das taxas de erro: ea/e_{opt}



Gráficos do rácio das taxas de erro: ec/e_{opt}



Capítulo 4

Conclusões

Neste trabalho apresentaram-se vários métodos de estimação de funções de densidade, para o caso univariado e multivariado, com o objectivo de apresentar a abordagem não paramétrica na análise discriminante e compará-la, em termos de desempenho, com a abordagem clássica.

Para fazer essa comparação foram aplicados dois métodos paramétricos (discriminante linear e discriminante quadrática) e dois não paramétricos (método dos k-vizinhos mais próximos e método do produto do núcleo) a um conjunto de exemplos ilustrativos, para os quais foram estimadas as taxas de erro aparentes, *"bootstrap"* e de validação cruzada.

Com base num estudo de simulação pelo método de Monte Carlo, procedeu-se ainda à comparação dos métodos acima mencionados, utilizando as taxas de erro de má classificação aparentes e de validação cruzada.

Os resultados obtidos sugerem os seguintes comentários:

A discriminante linear apresenta o melhor desempenho tanto sob o modelo normal, como sob modelos com contaminação leve (normal simetricamente contaminada) e mesmo em modelos com contaminação simétrica de causas pesadas ($t-$Student, com 3 graus de liberdade).

Para as mesmas distribuições, a discriminante quadrática tem também um bom desempenho, globalmente superior ao dos métodos não paramétricos.

Os métodos não paramétricos são mais eficientes que os paramétricos na presença de *"outliers"*, destacando-se, nesta situação, o método do produto do núcleo. O bom comportamento deste método sugere que seja, entre os métodos apresentados, o mais indicado para lidar com situações de multimodalidade.

Para distribuições com caudas longas o método do produto do núcleo é pouco eficiente, sendo inferior ao do método dos k-vizinhos mais próximos.

Os métodos não paramétricos não se revelam adequados para situações em que haja uma grande diferença de dimensão dos grupos.

Devem ser preferidas as taxas de erro de validação cruzada para avaliar o desempenho das regras discriminantes, que parecem mais fiáveis que as taxas de erro aparentes. A comparação com as taxas de erro *"bootstrap"* não foi incluída na parte de simulação, devido ao procedimento ser muito dispendioso em termos computacionais.

Existem outras possibilidades de continuação do trabalho iniciado, nomeadamente utilizando outros métodos, que adaptem a quantidade de alisamento à densidade local dos dados.

Outra área de interesse para futuro desenvolvimento, tem a ver com o facto de se terem considerado fixos alguns aspectos das distribuições e que se podem fazer variar (localização, dispersão, etc.).

Salienta-se também a necessidade de aplicação destes métodos a problemas com mais de duas variáveis e mais de dois grupos.

Apêndice A

Dados dos exemplos ilustrativos

Iris

Grupo A

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
5.1	3.5	1.4	0.2	4.9	3.0	1.4	0.2	4.7	3.2	1.3	0.2	4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2	5.4	3.9	1.7	0.4	4.6	3.4	1.4	0.3	5.0	3.4	1.5	0.2
4.4	2.9	1.4	0.2	4.9	3.1	1.5	0.1	5.4	3.7	1.5	0.2	4.8	3.4	1.6	0.2
4.8	3.0	1.4	0.1	4.3	3.0	1.1	0.1	5.8	4.0	1.2	0.2	5.7	4.4	1.5	0.4
5.4	3.9	1.3	0.4	5.1	3.5	1.4	0.3	5.7	3.8	1.7	0.3	5.1	3.8	1.5	0.3
5.4	3.4	1.7	0.2	5.1	3.7	1.5	0.4	4.6	3.6	1.0	0.2	5.1	3.3	1.7	0.5
4.8	3.4	1.9	0.2	5.0	3.0	1.6	0.2	5.0	3.4	1.6	0.4	5.2	3.5	1.5	0.2
5.2	3.4	1.4	0.2	4.7	3.2	1.6	0.2	4.8	3.1	1.6	0.2	5.4	3.4	1.5	0.4
5.2	4.1	1.5	0.1	5.5	4.2	1.4	0.2	4.9	3.1	1.5	0.2	5.0	3.2	1.2	0.2
5.5	3.5	1.3	0.2	4.9	3.6	1.4	0.1	4.4	3.0	1.3	0.2	5.1	3.4	1.5	0.2
5.0	3.5	1.3	0.3	4.5	2.3	1.3	0.3	4.4	3.2	1.3	0.2	5.0	3.5	1.6	0.6
5.1	3.8	1.9	0.4	4.8	3.0	1.4	0.3	5.1	3.8	1.6	0.2	4.6	3.2	1.4	0.2
5.3	3.7	1.5	0.2	5.0	3.3	1.4	0.2	-	-	-	-	-	-	-	-

Grupo B

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
7.0	3.2	4.7	1.4	6.4	3.2	4.5	1.5	6.9	3.1	4.9	1.5	5.5	2.3	4.0	1.3
6.5	2.8	4.6	1.5	5.7	2.8	4.5	1.3	6.3	3.3	4.7	1.6	4.9	2.4	3.3	1.0
6.6	2.9	4.6	1.3	5.2	2.7	3.9	1.4	5.0	2.0	3.5	1.0	5.9	3.0	4.2	1.5
6.0	2.2	4.0	1.0	6.1	2.9	4.7	1.4	5.6	2.9	3.6	1.3	6.7	3.1	4.4	1.4
5.6	3.0	4.5	1.5	5.8	2.7	4.1	1.0	6.2	2.2	4.5	1.5	5.6	2.5	3.9	1.1
5.9	3.2	4.8	1.8	6.1	2.8	4.0	1.3	6.3	2.5	4.9	1.5	6.1	2.8	4.7	1.2
6.4	2.9	4.3	1.3	6.6	3.0	4.4	1.4	6.8	2.8	4.8	1.4	6.7	3.0	5.0	1.7
6.0	2.9	4.5	1.5	5.7	2.6	3.5	1.0	5.5	2.4	3.8	1.1	5.5	2.4	3.7	1.0
5.8	2.7	3.9	1.2	6.0	2.7	5.1	1.6	5.4	3.0	4.5	1.5	6.0	3.4	4.5	1.6
6.7	3.1	4.7	1.5	6.3	2.3	4.4	1.3	5.6	3.0	4.1	1.3	5.5	2.5	4.0	1.3
5.5	2.6	4.4	1.2	6.1	3.0	4.6	1.4	5.8	2.6	4.0	1.2	5.0	2.3	3.3	1.0
5.6	2.7	4.2	1.3	5.7	3.0	4.2	1.2	5.7	2.9	4.2	1.3	6.2	2.9	4.3	1.3
5.1	2.5	3.0	1.1	5.7	2.8	4.1	1.3	-	-	-	-	-	-	-	-

Mulheres portadoras do gene da hemoflia

Grupo A

x_1	x_2	x_1	x_2	x_1	x_2
0.049218	-0.026872	0.056905	0.033424	-0.013228	-0.065502
-0.113509	-0.075721	0.130334	0.139879	0.103804	0.143015
0.071882	0.000000	-0.148742	-0.275724	0.056905	-0.070581
0.107210	0.064458	0.206826	0.113943	0.117271	0.033424
-0.124939	-0.045758	-0.045758	-0.193820	-0.091515	-0.050610
-0.080922	-0.119186	0.021189	0.056905	-0.080922	-0.040959
0.152288	0.068186	0.008600	-0.060481	-0.236572	-0.267606
-0.036212	-0.148742	-0.356547	-0.408935	-0.200659	-0.267606
-0.161151	-0.207608	0.037427	-0.013228	0.093422	0.086360
0.232996	0.195900	0.037427	0.123852	0.012837	0.110590

Grupo B

x_1	x_2	x_1	x_2	x_1	x_2
-0.24466	-0.04067	-0.42318	-0.09981	-0.24529	0.28764
-0.22047	0.00455	-0.21539	-0.02191	-0.34470	0.00969
-0.25404	-0.05729	-0.37780	-0.26816	-0.40465	-0.11618
-0.06391	0.15694	-0.33510	-0.13676	-0.01493	0.15392
-0.03124	0.14001	-0.17402	-0.07764	-0.26421	0.08669
-0.02344	0.08038	-0.33525	0.08753	-0.18782	0.25096
-0.17443	0.18924	-0.40546	-0.24184	-0.24443	0.16137
-0.47837	0.02822	-	-	-	-

Grupo – Normal Simetricamente Contaminada

Grupo A

x_1	x_2	x_1	x_2	x_1	x_2
-1.06507070	-0.91224458	0.00128922	0.57538114	-0.29174288	-0.16107278
-1.20537017	-0.26954624	0.11169533	1.74517987	0.37453288	-1.42234286
0.11997240	0.27552460	0.22626246	0.01052956	-0.42747438	-1.01983173
0.32967152	0.68099947	-1.16038738	1.44612636	1.52352762	-1.92195218
-1.39983772	-0.32006314	-0.27768543	0.13866398	0.50032952	0.06072202
1.33139509	0.13134255	0.02827484	-2.10388071	0.28960316	2.08217597
0.48052076	-0.59133345	0.31736637	-1.47782583	-1.00518666	-0.91564256
0.34004722	1.30323053	0.45644865	1.09115469	-0.43137206	0.34836926
-1.25163817	2.45085516	4.18019781	1.97630722	0.29734149	-1.02185518
0.46175279	0.45771702	0.95481837	-0.16700218	0.51108838	0.82023368

Grupo B

x_1	x_2	x_1	x_2	x_1	x_2
2.74311858	-1.27788996	5.59344288	2.74968341	5.21015251	0.85921092
3.91987254	0.53711072	3.90211866	-1.03923407	5.33890805	-0.25229887
2.81530337	-0.88447675	4.85468529	-0.40648448	4.15927600	-1.46289260
5.25040263	-0.18144830	14.61798946	-1.49473997	5.51413657	0.68931047
4.27248444	0.70621692	4.25484204	1.17158480	10.01281721	-1.15478656
3.92963967	-0.36569773	4.08840082	1.47290765	3.16010474	0.62931126
4.19000812	-0.25344922	5.33645783	-0.30913791	3.89635131	-0.16768080
5.06251762	0.26950248	4.27895517	-0.19092486	5.29058289	-1.50038428
4.64216630	-0.04934975	14.28704893	-2.67567806	4.57964013	-0.38999725
3.96544224	1.47659393	6.06788729	0.68218321	3.41508737	-0.24965449

Outro – Normal Com "Outliers"

Grupo A

x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
0.0299	-0.3276	-0.1409	-1.0204	0.1462	0.4034	0.3665	0.1972
1.0151	-1.1550	0.5091	-0.1854	-0.0253	-1.0870	1.4867	1.0639
0.3765	-1.9078	-0.4534	-0.1203	-1.2157	0.8619	-0.2313	-0.9147
1.6514	-1.5878	0.9928	1.3079	1.4726	-1.0806	-1.5701	0.0228
0.6266	-1.3565	-1.1949	0.1436	-2.0154	-1.0870	-1.0724	0.3919
-1.6848	0.7641	0.1880	-1.7422	-1.2035	-0.3051	0.5566	-0.8546
0.5822	-1.7768	-0.2558	2.2634	0.6806	-1.9318	0.2206	1.0611
0.6875	0.0953	2.0798	-0.4917	1.2582	-0.3835	0.5542	-1.9704
0.9385	-1.0961	-2.0517	0.0474	0.2803	-1.1126	-0.2757	0.5546
-0.5907	0.9956	-0.5009	-0.2141	-0.1470	0.6367	0.5930	-1.4168
1.2402	1.7055	-1.1988	-0.0215	0.9108	-0.1245	-0.6349	0.2530
0.7508	-1.1711	0.0038	1.7360	-2.4544	-0.0113	0.3839	-1.1596
0.0343	0.0492	-0.8658	-1.2481	0.2383	-1.6053	-0.2999	0.6383
1.7039	-1.0327	-1.6900	0.9442	0.5211	-1.1488	0.1211	-0.7446
1.5438	0.3501	-0.5632	0.9285	-0.3849	-0.8176	-0.2816	0.4020
-1.0023	-1.3134	0.6488	0.6131	0.7782	-0.3796	-0.3876	-2.1651
1.5287	1.3313	-0.8825	-0.4865	0.3364	0.8438	1.0291	0.9001
1.3935	-2.2366	-0.9190	1.3826	0.9462	1.0651	-0.8451	0.2157
0.2129	1.3013	0.2751	-1.1631	-0.0329	1.4755	1.2119	0.4567
-0.7795	0.5539	-0.2939	0.7979	-0.7213	-0.1600	0.4738	-1.0873
2.2168	-1.1760	-0.4244	-0.0360	-0.8979	1.3230	-2.3676	1.5859
0.7870	0.8983	0.9983	0.8940	-1.1787	-1.1654	-0.2287	-0.6534
-1.7729	-0.3616	-0.6275	-0.4524	10.0000	10.0000	10.0000	10.0000
10.0000	10.0000	10.0000	10.0000	10.0000	10.0000	10.0000	10.0000
10.0000	10.0000	10.0000	10.0000	10.0000	10.0000	10.0000	10.0000

Outro – Normal Com "Outliers"

Grupo B

x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
2.65036	1.87616	5.61617	0.01615	2.14398	1.16973	2.87220	0.44186
3.96152	1.34356	4.57989	-0.62462	5.27098	0.18541	2.67082	0.01726
3.88800	1.93122	4.54787	0.16010	5.79843	-1.23360	3.30557	1.40362
5.68458	-1.07358	4.78935	0.23875	2.91572	-0.71270	3.32350	-0.15676
3.39464	-0.49455	5.69162	-2.25658	2.59421	-2.26023	5.94499	0.34173
3.71893	0.15253	4.29273	0.74181	2.84743	1.09632	2.87721	0.95293
2.76200	0.64794	6.26401	0.53210	2.91728	-1.43005	5.37030	0.60645
3.41441	-0.07481	4.57441	0.92615	3.01320	-1.86910	2.84218	-0.97250
4.33666	-0.68466	2.76244	1.75951	3.53162	1.47564	2.63769	-1.58535
4.20938	-0.78803	4.97905	-0.47258	6.51031	0.06403	4.27293	0.80736
5.48199	-0.85341	5.29412	-0.27478	5.56644	-0.48682	4.46493	0.57464
3.65193	1.43493	5.80343	0.20262	3.03622	1.11159	3.15044	0.25262
4.03901	-0.91598	5.48909	-0.71578	3.84259	1.01681	2.56335	-1.62560
4.09150	-0.55104	4.14847	-1.12959	4.73427	0.66669	6.07922	-1.41838
4.76251	-0.10744	4.19573	-1.37874	2.10189	0.16362	1.11276	-0.76992
5.31374	-0.73267	4.91786	-0.90129	3.83296	-1.02004	5.33901	0.24522
5.36508	1.13564	2.83846	-0.67685	2.43318	1.98704	3.67368	-0.81370
4.93229	-0.11676	3.89539	0.48409	3.26907	-1.28342	3.70270	-0.62471
3.93951	0.01843	5.76709	-0.67727	3.72089	0.43061	3.49357	1.13621
4.32810	2.25107	3.62395	-0.19513	3.67632	-1.66667	4.62179	0.03249
4.34049	-0.06651	3.83375	-1.69634	3.99877	-0.33483	3.69702	0.21649
3.46907	-0.11985	6.32260	-2.30001	3.97912	0.51617	4.20436	0.83026
5.14961	0.20551	2.29345	0.62947	2.47213	-0.31817	3.34614	0.38252
5.83059	2.01062	4.28761	-0.78773	3.51306	-1.25815	4.47150	0.64180
4.47719	0.67997	3.98653	-0.28149	4.93107	0.40285	3.85998	-2.18601

Apêndice B

Programas

LDA - Estimativa dos coeficientes discriminantes

```
function(x, grouping, prior, tol = 0.0001)
{
    x <- as.matrix(x)
    n <- dim(x)[1]
    p <- dim(x)[2]
    if(n != length(grouping))
        stop("x e vector de identificação são de dimensão diferente")
    g <- as.factor(grouping)
    counts <- tapply(rep(1, length(g)), g, sum)
    prop <- counts/sum(counts)    # permite prior diferente
    if(missing(prior))
        prior <- prop
    else if(any(prior < 0) || round(sum(prior), 5) != 1)
        stop(" prior invalida")
    if(length(prior) != length(prop)) stop(
            "prior de dimensão incorrecta")
    # média por grupo (linhas) e variaveis (colunas)
    group.means <- tapply(x, list(rep(g, ncol(x)), col(x)), mean)
    f1 <- sqrt(diag(var(x - group.means[g,  ])))
    if(any(f1 < tol))
        stop(paste(" variaveis", paste(format((1:p)[f1 < tol]),
            collapse = " "),
            "parecem constantes dentro dos grupos"))
    scaling <- diag(1/f1)
    X <- sqrt(1/(n - length(prop))) * (x - group.means[g,  ]) %*%
        scaling
    X.s <- svd(X)
    rank <- sum(X.s$d > tol)
    if(rank < p)
        warning("variaveis são colineares")
    scaling <- scaling %*% X.s$v[, 1:rank] %*% diag(1/X.s$d[1:rank])
    X <- scale(group.means[g,  ], T, F) %*% scaling
```

```
    xbar <- apply(prior %*% group.means, 2, sum)
    X <- sqrt((n * prior)/(length(prop) - 1)) * scale(group.means, xbar,
        F) %*% scaling
    X.s <- svd(X)
    rank <- sum(X.s$d > tol * X.s$d[1])
    rank <- rank + 1
    scaling <- scaling %*% X.s$v[, 1:rank]
    structure(list(prior = prior, means = group.means, scaling = scaling,
        lev = levels(g), svd = X.s$d[1:rank]), class = "lda")
}
```

PREDICT.LDA - Cálculo das taxas de erro aparentes

```
function(object, x, group, prior = object$prior, dimen)
{
    if(!inherits(object, "lda"))
        stop("object not of class lda")
    which.is.min <- function(x)
    {
        d <- (1:length(x))[x == min(x)]
        if(length(d) > 1)
            d <- sample(d, 1)
        d
    }
    if(is.null(dim(x)))
        dim(x) <- c(1, length(x))
    x <- as.matrix(x)
    if(dim(x)[2] != dim(object$means)[2])
        stop("wrong number of variables")
    if(missing(dimen))
        dimen <- length(object$svd)
    else dimen <- min(dimen, length(object$svd))
    scaling <- object$scaling[, 1:dimen]
    means <- apply(object$means, 2, mean)
    object$means <- scale(object$means, means, F)
    x <- scale(x, means, F)
    dm <- object$means %*% scaling
    dist <- matrix(0.5 * apply(dm^2, 1, sum) - log(prior), nrow(x),
        length(prior), byrow = T) - (x %*% scaling) %*% t(dm)
    cl <- apply(dist, 1, which.is.min)
    levels(cl) <- object$lev
    cl <- factor(cl)    # converte para probabilidades a posterior
    dist <- exp( - (dist))
    posterior <- dist/drop(dist %*% rep(1, length(prior)))
    dimnames(posterior) <- list(dimnames(x)[[1]], object$lev)
    list(class = cl, posterior = posterior, x = x %*% scaling)
    table(group, cl)
```

```
}
```

JACKK.LDA - Cálculo das taxas de erro de validação cruzada

```
function(x, group, dimen, prior, dim1, dim2, prior)
{
    p <- prior
    x <- as.matrix(x)
    n <- dim(x)[1]
    dim1 <- dim1
    quadro1 <- matrix(0, nrow = n, ncol = 1)
    for(i in 1:n) {
        p <- prior
        x <- as.matrix(x)
        tr <- seq(1, n)
        x1 <- x[ - i,  ]
        dtr <- tr[ - i][(tr[ - i] <= dim1)]
        y <- (length(dtr))
        dtr1 <- tr[ - i][(tr[ - i] > dim1)]
        y1 <- (length(dtr1))
        y2 <- (dim1 - (length(dtr)))
        y3 <- (dim2 - (length(dtr1)))
        train <- x1
        test <- x[i,  ]
        g1 <- factor(c(rep("A", y), rep("B", y1)))
        g2 <- factor(c(rep("A", y2), rep("B", y3)))
        a1 <- lda(train, g1, prior = p)
        test <- as.matrix(test)
        which.is.min <- function(test)
        {
            test <- (1:length(test))[test == min(test)]
            if(length(test) > 1)
                test <- sample(test, 1)
            test
        }
        if(length(test) != dim(a1$means)[2])
            stop("wrong number of variables")
        if(missing(dimen))
            dimen <- length(a1$svd)
        else dimen <- min(dimen, length(a1$svd))
        scaling <- a1$scaling[, 1:dimen]
        means <- apply(a1$means, 2, mean)
        a1$means <- scale(a1$means, means, F)
        test <- t(test)
        test <- scale(test, means, F)
        dm <- a1$means %*% scaling
        dist <- matrix(0.5 * apply(dm^2, 1, sum) - log(prior), nrow(
```

```
            test), length(prior), byrow = T) - (test %*% scaling
            ) %*% t(dm)
        cl <- apply(dist, 1, which.is.min)
        levels(cl) <- a1$lev
        cl <- factor(cl)
        quadro <- table(g2, cl)
        quadro1[i] <- levels(cl)
    }
    quadro1 <- as.factor(quadro1)
    table(group, quadro1)
}
```

Cálculo das taxas de erro bootstrap(QDA)

```
function(x, group, nboot, dimen, prior = c(0.5, 0.5), dim1, dim2, s = 1)
{
    a <- qda(x, group = group, prior = prior)
    #   "Taxas aparentes - mi"
    quadro <- predict.qda(a, x, group = group)
    aux5 <- matrix(0, nrow = 2, ncol = 2)
    aux5[1, 1] <- quadro[1, 1]/dim1 * 100
    aux5[2, 1] <- quadro[2, 1]/dim2 * 100
    aux5[1, 2] <- quadro[1, 2]/dim1 * 100
    aux5[2, 2] <- quadro[2, 2]/dim2 * 100
    dim11 <- dim1 + 1
    n <- dim(x)[1]
    quadro1 <- matrix(0, nrow = 2, ncol = 2)
    quadro2 <- matrix(0, nrow = 2, ncol = 2)
    quadro3 <- matrix(0, nrow = 2, ncol = 2)
    quadro31 <- matrix(0, nrow = 2, ncol = 2)
    quadro41 <- matrix(0, nrow = 2, ncol = 2)
    quadro4 <- matrix(0, nrow = 2, ncol = 2)
    quadro5 <- matrix(0, nrow = 2, ncol = 2)
    aux3 <- matrix(0, nrow = 2, ncol = 2)
    aux4 <- matrix(0, nrow = 2, ncol = 2)
    for(i in 1:nboot) {
    #amostragem mista(s=1)
        if(s == 1) {
            tr <- sample(dim(x)[1], replace = T)
        }
        else {
            tr1 <- sample(dim1, replace = T)
            tr2 <- sample(dim2, replace = T)
            tr2 <- dim1 + tr2
            tr <- c(tr1, tr2)
        }
        tr <- sort(tr)
```

```
        dtr <- tr[(tr <= dim1)]
        y <- (length(dtr))
        dtr1 <- tr[(tr > dim1)]
        y1 <- (length(dtr1))
        g1 <- c(rep("A", y), rep("B", y1))
        train <- (x[tr,  ])
        a1 <- qda(train, g1, prior = prior)
        dist1 <- predict.qda(a1, group = g1, train)
        dist2 <- predict.qda(a1, group = group, x)
        # "mi*"
        quadro1 <- dist1
        # "mi**"
        quadro2 <- dist2
        aux3[1, 1] <- (quadro2[1, 1]/dim1) * 100
        aux3[2, 1] <- (quadro2[2, 1]/dim2) * 100
        aux3[1, 2] <- (quadro2[1, 2]/dim1) * 100
        aux3[2, 2] <- (quadro2[2, 2]/dim2) * 100
        # "mi**/ni"
        # aux3
        aux4[1, 1] <- (quadro1[1, 1]/y) * 100
        aux4[2, 1] <- (quadro1[2, 1]/y1) * 100
        aux4[1, 2] <- (quadro1[1, 2]/y) * 100
        aux4[2, 2] <- (quadro1[2, 2]/y1) * 100
        #"mi*/ni*"
        # aux4
        quadro3 <- (aux3 - aux4)
        #"di"
        #quadro3
        quadro31 <- quadro31 + quadro3
        #"di-mean"
        quadro4 <- quadro31/(i)
        #"mi/ni"
        # aux5
        quadro5 <- (aux5 + quadro4)
    }
    print("Erro Total")
    a <- c(quadro5[1, 2], quadro5[2, 1])
    a
}
```

Nucleo - Gráfico da função de densidade estimada, para dados bivariados (kde2d)

```
function(x, y, n = 25, h, x1, y1, lims = c(range(x1), range(y1)), zlim, xlab,
    ylab)
{
    nx <- length(x)
```

```
    if(length(y) != nx)
        stop("Os dois vectores devem ter a mesma dimensão")
    gx <- seq(lims[1], lims[2], length = n)
    gy <- seq(lims[3], lims[4], length = n)
    if(missing(h))
        h <- c(bandwidth.nrd(x), bandwidth.nrd(y))
    h <- h/4
    ax <- outer(gx, x, "-")/h[1]
    ay <- outer(gy, y, "-")/h[2]
    z <- matrix(dnorm(ax), n, nx) %*% t(matrix(dnorm(ay), n, nx))/(nx *
        h[1] * h[2])    #    mini <- min(range(x1), range(y1))
    zzz <- (list(x = gx, y = gy, z = z))
    persp(zzz, xlab = xlab, ylab = ylab, box = T, ar = 1, axes = T, zlim
         = zlim)
    contour(zzz)
    a <- range(z)
    a
}
```

Apêndice C

Referências Bibliográficas

Abramson, I. S. (1982). On bandwidth variation in kernel estimates- a square root law. *The Annals of Statistics,* **10,** 1217-1223.

Abramson, I. S. (1984). Adaptive density flattening - A metric distortion principle for combating bias in nearest neighbor methods. *The Annals of Statistics,* **12,** 880–886.

Aitchison, J. e Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika,* **63** (3), 413-420.

Albert, A., e Anderson, J. A. (1981). Probit and logistic discriminant functions. *Communications in Statistics - Theory and Methods,* **A10**, 641-657.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* First edition. New York: Wiley.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika,* **59,** 19-35.

Anderson, J. A. (1975). Quadratic logistic discrimination. *Biometrika,* **62,** 149-154.

Anderson, J. A. (1982). Logistic discrimination. Em *Handbook of Statistics* (vol 2), P. R. Krishnaiah e L. Kanal (Editores). Amsterdam: North-Holland, pp. 169-191.

Anderson, J. A. e Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika,* **69,** 123-136..

Bertrand-Retali, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode de noyau. *Rev. Roumaine Math. Pures. Appl.*, **23**, 361-385.

Boneva, L. I., Kendall, D. G. e Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the statistical data-analyst (with Discussion). *Journal of the Royal Statistical Society* (series B), **33**, 1-70.

Bouma, B. N., van der Klaauw, M. M., Veltkamp, J. J., Starkenburg, A. E., van Tilburg, N. H. e Hermans, J. (1975). Evaluation of the detection of hemophilia carriers. *Thrombosis Research,* **7,** 339-350.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika,* **71 (2),** 353-360.

Bowman, A. W., Hall, P. e Titterington, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika.* **71 (2),** 341-351.

Breiman, L., Meisel, W., e Purcell, E.. (1977). Variable kernel estimates of multivariate densities. *Technometrics,* **19,** 135-144.

Čencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.,* **3,** 1559-1562.

Day, N. E. e Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics,* **23,** 313-324.

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés (II). *Publication de l'Institute de Statistique de l'Université de Paris,* **XXII,** 1-23.

Devroye, L. P. (1987). *A course in density estimation.* Boston: Birkhauser.

Devroye, L. P. e Gyõrfi, L. (1985). *Nonparametric density estimation: The L_1 View.* New York: Wiley..

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probabilities density functions. *IEEE Transactions on Computers,* **C-25,** 1175-1179.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics,* **7,** 1-26.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika,* **68,** 589-599.

Enas, G. G. e Choi, S.C. (1986). Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. *Comput. Math. Applic.* **12A,** 235-244.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theor. Prob. Appl.,* **14,** 153-158.

Fix , E. e Hodges, J. L. (1951). Discriminatory analysis - Nonparametric discrimination: Consistency properties. *em (1989) International Statistical Review,* **57,** (3), 238-247.

Fisher, R. A. (1986). The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* **7,** 179-188.

Freedman D. e Diaconis, P. (1981). On the histogram as a density estimator L_2 Theory. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebeite,* **57,** 453-476.

Fryer, M. J. (1976). Some errors associated with the nonparametric estimation of density functions. *J. Inst. Maths. Applics,* **18,** 371-380.

Gawronski, W. e Stadtmuller, U. (1980). On density estimation by means of Poisson's distribution. *Scandinavian Journal of Statistics,* **7,** 90-94.

Glick, N. (1973). Separation and probability of correct classification among two or more distributions. *Ann. Inst. Statist. Math.,* **25,** 373-382.

Goldstein, M. e Dillon, W. R. (1978). *Discrete discriminant analysis.* New York: Wiley.

Habbema, J. D. F., Hermans, J. e van der Broek, A. T. (1974). A stepwise discriminant analysis program using density estimation. *Compstat 1974, Proc. Computational Statistics,* Vienna: Phisica-Verlag, pp. 101-110.

Hall, P. (1983a). Large-sample optimality of least-squares cross-validation in density estimation. *The Annals of Statistics,* **11,** 1156-1174.

Hall, P. (1983b). Orthogonal series methods for both qualitative and quantitative data. *The Annals of Statistics,* **11,** 1004-1007.

Hall, P. e Wand, Mettil P. (1988). On nonparametric discrimination using density differences. *Biometrika,* **75 (3),** 541-547.

Hand, D. J. (1981). *Discrimination and Classification.* Wiley, New York.

Hastie T. e Tibshirani, R. J. (1986). Generalized additive models (with discussion). *Statist. Science,* **1,** 297-318.

Hastie T. e Tibshirani, R. J. (1990). *Generalized additive models.* London: Chapman and Hall.

Hermans, J. e Habbema, J. D. F. (1975). Comparision of five methods to estimate posterior probabilities.*EDV in Medizine und Biologie,* **6,** 14-19.

Hills, M. (1966). Allocation rules and their errors rates (with discussion). *Journal of the Royal Statistical Society* (series B) **28,** 1-31.

Hodges, J. L. e Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics,* **27**, 324-335.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association,* **86,** (413), 205-224.

Kronmal, R. A. e Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association,* **63,** 925-952.

Lachenbruch, P. A. (1975). *Discriminant Analysis.* Hafner: New York.

Lachenbruch, P. A. e Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics,* **10,** 1-11.

Loftsgaarden, D. O. e Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics,* **36,** 1049-1051.

McLachlan, G. J. (1980). A note on bias correction in maximum likelihood estimation with logistic discrimination. *Technometrics,* **22,** 621-627.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Mucciardi, A. N. e Gose, E. E. (1972). An automatic clustering algorithm and its properties in high-dimensional spaces. *IEEE Transactions on Systems, Man and Cybernetics,* **SMC-2**, 247-254.

Murphy, B. J. e Moran, M. A. (1986). Parametric and kernel density methods in discriminant analysis: another comparison.. *Comput. Math. Applic.,* **12A,** 197-207.

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theor. probab. Appl.*, **10**, 186-190.

Ott, J. e Kronmal, R. A. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *Journal of the American Statistical Association,* **71,** 391-399.

Park, B. U. e Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association,* **85**, 66-72.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics,* **33,** 1065-1076.

Pires, A. M. (1995). *Análise Discriminante - Novos Métodos Robustos de Estimação.* Tese de Doutoramento. I. S. T., Universidade Técnica de Lisboa, Lisboa.

Prentice, R. L. e Pyke, R. (1979). Logistic disease incidence models and case control studies. *Biometrika,* **66,** 403-411.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics,* **27,** 832-837.

Rosenblatt, M. (1971). Curve estimates. *Annals of Mathematical Statistics,* **42,** 1815-1842.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.,* **9,** 65-78.

S-PLUS (1990). *"User's Manual", "Programming Manual", "Guide to Statistical and Mathematical Analysis".* StatSci, Inc., Seattle, WA..

Schuster, E. F. e Gregory, G.G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. Em *15^{th} Annuals Symp. on the Interface of Computer Science and Statistics,* W. F. Eddy (editores). New York: Springer-Verlag, pp. 295-298.

Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *Annals of Mathematical Statistics,* **38,** 1261-1265.

Scott, D. W. (1979). On optimal and data based histograms. *Biometrika,* **66** (3), 605-610.

Scott, D. W. (1985a). Frequency Polygons. *Journal of the American Statistical Association,* **80,** 348-354.

Scott, D. W. (1985b). Averaged Shifted Histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics,* **13,** 1024–1040.

Scott, D. W. (1992). *Multivariate Density Estimation - Theory, Practice and Visualization.* New York: Wiley.

Scott, David W. e Factor, Lynette E. (1981). Montecarlo study of three data based nonparametric probability density estimators. *Journal of the American Statistical Association,* **76,** (373), 9-15.

Scott, D. W. e Thompson, J. R. (1983). Probability density estimation in higher dimensions. Em Gentle, J. E. (editores), *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface,* Amsterdam: North-Holand, pp. 173-179.

Seber, G. A. F. (1984). *Multivariate Observations.* New York: Wiley.

Sebestyen, G. e Edie, J. (1966). An algorithm for nonparametric pattern recognition. *IEEE Transactions on Electronic Computers,* **EC-15**, 908-915.

Silverman, B. W. (1978a). Choosing the window width when estimating a density. *Biometrika,* **65,** 1-11.

Silverman, B. W. (1978b). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *The Annals of Statistics,* **6,** 177-184.(Addendum 1980, *The Annals of Statistics,* **8,** 1175-1176).

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall (editores).

Silverman, B. W. e Jones, M.C. (1989). Commentary on a paper by Fix, E. e Hodges, J. L. (1951): An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review,* **57,** (3), 233-238.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *The Annals of Statisics,* **5,** 595-645.

Stone, C. J. (1984). An assimptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics,* **12,** 1285-1297.

Tomek, I. (1976). An experiment with the edited nearest neighbour rule. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-6,** 448-452.

Tutz, G. E. (1986). An alternative choice of smoothing for kernel based density estimates in discrete discriminant analysis. *Biometrika,* **73,** 405-411.

Tutz, G. E. (1988). Smoothing for discrete kernels in discrimination. *Biom. J.,* **6,** 729-739.

Tutz, G. E. (1989). On cross-validation for discrete kernel estimates in discrimination. *Communications in Statistics - Theory and Methods,* **18,** 4145-4162.

Van Ness, J. W. (1979). On the effects of dimension in discriminant analysis for unequal covariance populations. *Technometrics,* **21,** 119-127.

Van Ness, J. W. e Simpson, C. (1976). On the effects of dimension in discriminant analysis. *Technometrics,* **18,** 175-187.

Vitale, R. A. (1975). A Bernestein polynomial approach to density estimation. Em *Statistical Inference and Related Topics* (vol. 2), (editor. M. L. Puri), San Francisco: Academic Press, pp. 87-99.

Welch, B. L. (1939). Note on discriminant functions. *Biometrika,* **31,** 218-220.

Wertz, W. (1978). *Statistical Density Estimation: A Survey.* Göttingen: Vandenhoeck and Ruprecht.

Whittle, P. (1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society* (series B), **20,** 334-343.

Zimmerman, T. S., Ratnoff, O. D. e Littel, A. S. (1971). Detection of carriers of classic hemophilia using an imunologic assay for antihemophilia factor (factor VIII). *The Journal of Clinical Investigation,* **50,** 255-258.