REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 1999

ANÁLISE DISCRIMINANTE COM SELECÇÃO DE VARIÁVEIS
1ª PARTE: DESCRIÇÃO

DESCRIPTIVE DISCRIMINANT ANALYSIS WITH VARIABLE SELECTION

Autor: António Pedro Duarte Silva•
Professor Auxiliar – Faculdade de Ciências Económicas e Empresariais
Universidade Católica Portuguesa – Centro Regional do Porto

Resumo:

- Neste trabalho discute-se o problema de selecção de variáveis em Análise Discriminante entendida numa perspectiva descritiva. É feita uma revisão de várias técnicas onde se incluem: métodos informais de selecção implícita, métodos de selecção passo a passo, métodos de comparação entre todos os subconjuntos possíveis e testes estatísticos de adequação. São propostos métodos de comparação entre todos os subconjuntos possíveis baseados em vários índices alternativos, a escolher consoante o ênfase que se pretende dar a diferentes dimensões de separação. As técnicas apresentadas serão ilustradas por um exemplo relativo à descrição das diferenças entre três grupos de bancos a operar em Portugal em 1993.

PALAVRAS-CHAVE:

- *Análise Discriminante, Selecção de Variáveis, Índices Multivariados.*

ABSTRACT:

This paper discusses several issues concerning the problem of variable selection in Descriptive Discriminant Analysis. The topics covered include informal methods for discarding variables, stepwise and all-subsets methods for variable selection, statistical tests of subset adequacy and choice of criteria for variable selection. Methods for all-subsets comparisons are proposed. It is shown how the choice of criteria for comparing subsets is related to the importance given to different dimensions of group separation.

KEY-WORDS:

- *Discriminant Analysis, Variable Selection, Multivariate indexes.*

• O autor agradece a Mário Coutinho dos Santos toda a ajuda prestada com a recolha de dados, interpretação de resultados em termos financeiros, e várias discussões que contribuiram para melhorar a sua qualidade deste trabalho. Os erros e omissões que subsistem são da exclusiva responsabilidade do autor.

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 1999

1. INTRODUÇÃO

A expressão *Análise Discriminante* (AD) é utilizada para designar técnicas estatísticas que têm como objectivo o estudo das diferenças entre grupos bem definidos à partida com base num conjunto relevante de características dos seus elementos. Dentro desta designação genérica encontram-se duas grandes subdivisões: a das técnicas que procuram identificar e interpretar as diferenças existentes entre os grupos e a das técnicas que estudam regras que permitem classificar indivíduos de origem desconhecida num dos grupos existentes.

Na prática, é comum que no mesmo estudo se tenham que interpretar diferenças entre grupos e simultaneamente estabelecer e estudar propriedades de regras de classificação. No entanto, estes dois problemas, embora relacionados, são conceptualmente diferentes e requerem métodos de abordagem distintos. Neste artigo, o estudo das diferenças entre grupos tendo em vista a sua interpretação será designado por *Análise Discriminante Descritiva*[1] (ADD) enquanto que o estudo de regras de classificação será designado por *Análise Discriminante Classificatória* (ADC)[2].

As técnicas clássicas de AD assumem que as características consideradas em cada indivíduo são representadas por um conjunto de variáveis escolhido à partida. Na prática, é comum recolher inicialmente um elevado numero de variáveis, efectuando uma selecção no decorrer da análise, ou simplesmente ignorando para fins de interpretação aquelas variáveis que se revelarem menos importantes ou interessantes. Este tipo de abordagem é muitas vezes baseado em procedimentos ad-hoc de propriedades mal conhecidas. Além disso, não é de todo incomum que depois de se proceder a uma selecção de variáveis, se prossiga a análise ignorando os enviesamentos decorrentes do processo de selecção.

Neste artigo far-se-á uma revisão dos principais métodos de selecção de variáveis em ADD. Inicialmente, discutir-se-ão algumas formas habituais de lidar com este problema, nomeadamente métodos informais de análise e métodos de selecção passo a passo. Em seguida, discutir-se-á o problema de identificar os subconjuntos de

[1] O termo "descritiva" não é aqui utilizado em oposição a "inferencial" como frequentemente acontece em outras técnicas estatísticas. Com efeito, embora técnicas puramente descritivas (nomeadamente técnicas factoriais) sejam empregues em ADD, há também métodos inferenciais que podem ser usados para ajudar a compreender diferenças entre grupos. Nomeadamente, questões do tipo, "Quantas dimensões são necessárias para explicar as diferenças entre os grupos ?" ou "Qual o subconjunto minimo de variàveis que explica todas as diferenças observadas ?", são tipicamente abordadas com a ajuda de testes de hipóteses baseados em modelos probabilisticos.

[2] Desconhecemos a existência de alguma terminologia portuguesa já estabelecida para distinguir estas duas vertentes de *Análise Discriminante*. Daí a necessidade de definir uma terminologia própria. Aliás, tanto quanto sabemos, são raros os textos em português que fazem uma distinção clara das técnicas de *Análise Discriminante* quanto aos objectivos visados. Tal não é o caso por exemplo da literatura anglo-saxónica, onde técnicas de ADD são geralmente designadas por *Descriptive Discriminant Analysis* enquanto técnicas de ADC tem sido designadas por *Allocation, Classification in Discriminant Analysis* ou *Predictive Discriminant Analysis* (esta última designação é um pouco infeliz, na medida em que pode criar confusão com abordagens Bayesianas). Na literatura de expressão francesa as técnicas de ADD tem sido designadas por *Analyse Discriminante au but Descriptive* ou *Analyse Factoriel Discrimante* enquanto a ADC é designada por *Classement.*

variáveis que incluem toda a informação relevante para explicar as diferenças entre os grupos. Por último, propor-se-ão métodos de comparação entre todos os subconjuntos possíveis, sugerindo-se vários índices que poderão ser utilizados para esse efeito. A maioria das técnicas discutidas neste artigo são conhecidas, se bem que algumas estejam ainda pouco divulgadas. A discussão dos índices para a comparação entre todos os subconjuntos possíveis é original. O problema de selecção de variáveis em ADC será discutido num próximo artigo.

As técnicas e problemas discutidos neste artigo serão ilustradas com um exemplo relativo ao estudo das diferenças entre bancos portugueses criados depois de 1984, existentes antes de 1984, e bancos estrangeiros a operar em Portugal. O ano de 1984 foi escolhido devido a corresponder à data de aprovação da denominada "Lei de Delimitação dos Sectores" a qual terminou com as restrições ao acesso dos investidores privados à propriedade empresarial no sector bancário. As variáveis utilizadas serão indicadores de estrutura patrimonial, funcionamento e rendibilidade, construídos a partir de informação contabilistica extraída dos balanços e demonstrações de resultados de 1993 publicados no Boletim Informativo da Associação Portuguesa de Bancos. Os dados foram recolhidos e gentilmente cedidos pelo Dr. Mário Coutinho dos Santos, da FCEE da Universidade Católica Portuguesa, Centro Regional do Porto.

2. ABORDAGENS TRADICIONAIS

2.1 NOTAÇÃO E CONCEITOS FUNDAMENTAIS

Considere-se um conjunto de N indivíduos divididos em k grupos e descritos por vectores $\mathbf{x_{gi}} = [x_{gi1}, x_{gi2}, ..., x_{gip}]^T$ ($i = 1,2,...,n_g$; $g = 1,2,..k$), em que x_{gij} representa o valor que a variável X_j assume para o i-ésimo indivíduo do grupo g. Designe-se o número total de indivíduos (observações) por $N = \sum_{g=1}^{k} n_g$, os centroides de cada grupo por $\bar{\mathbf{x}}_\mathbf{g}$ e global por $\bar{\bar{\mathbf{x}}}$

$$\bar{\mathbf{x}}_\mathbf{g} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n_g}} \mathbf{x_{gi}}}{n_g} \qquad \bar{\bar{\mathbf{x}}} = \frac{\sum_{\mathbf{g=1}}^{\mathbf{k}} n_g \bar{\mathbf{x}}_\mathbf{g}}{N}$$

e as matrizes das somas dos desvios quadráticos e cruzados intra-grupos por **W** e entre-grupos por **B**

$$\boldsymbol{W} = \sum_{g=1}^{k} \sum_{i=1}^{n_g} (\boldsymbol{x}_{gi} - \bar{\boldsymbol{x}}_g)(\boldsymbol{x}_{gi} - \bar{\boldsymbol{x}}_g)^T \qquad \boldsymbol{B} = \sum_{g=1}^{k} \boldsymbol{n}_g (\bar{\boldsymbol{x}}_g - \bar{\bar{\boldsymbol{x}}})(\bar{\boldsymbol{x}}_g - \bar{\bar{\boldsymbol{x}}})^T$$

Designe-se ainda por

$$\boldsymbol{T} = \sum_{g=1}^{k}\sum_{i=1}^{n_g}\left(\boldsymbol{x}_{gi} - \bar{\bar{\boldsymbol{x}}}\right)\left(\boldsymbol{x}_{gi} - \bar{\bar{\boldsymbol{x}}}\right)^{T} = \mathbf{W} + \mathbf{B}$$

a matriz das soma dos desvios quadráticos e cruzados totais. Admita-se que **W** e **T** são matrizes não singulares e que **B** tem característica $r = \min(p, k\text{-}1)$[3].

As técnicas clássicas de ADD são baseadas na análise dos vectores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ [4]. Com efeito, é bem sabido que o vectores próprios associados ao primeiro valor próprio de $\mathbf{B}\,\mathbf{W}^{-1}$ (λ_1) definem combinações lineares que maximizam o rácio entre a inércia entre-grupos e a inércia intra-grupos. Mais concretamente, pretendendo-se maximizar o rácio

$$\frac{\sum_{g=1}^{k} n_g\left(\bar{z}_g - \bar{\bar{z}}\right)^2}{\sum_{g=1}^{k}\sum_{i=1}^{n_g}\left(z_{gi} - \bar{z}_g\right)^2} \quad \left(\text{com} \quad \bar{z}_g = \frac{\sum_{i=1}^{n_g} z_{gi}}{n_g} \text{ e } \quad \bar{\bar{z}} = \frac{\sum_{g=1}^{k} n_g\,\bar{z}_g}{N}\right)$$

entre todas as combinações lineares, $Z = b_1 X_1 + b_2 X_2 + ... + b_p X_p$, então $b_1, b_2, ..., b_p$ deverão ser escolhidos como as coordenadas de um dos vectores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ associados a λ_1. À função definida por este vector chama-se primeira Função Discriminante Linear (FDL_1) e a combinação linear resultante designa-se habitualmente por Z1. Dado que Z1 só está definida a menos de uma constante de proporcionalidade, é comum normalizá-la de forma a que tenha uma variância amostral (intra-grupos) unitária, ou seja, o vector próprio escolhido deverá garantir que $\sum_{g=1}^{k}\sum_{i=1}^{n_g}\left(z1_{gi} - \bar{z}\bar{1}_g\right)^2 / (N - k) = 1$. Aos valores assumidos por Z1 chamam-se "scores" na primeira FDL. A b_j chama-se coeficiente não padronizado de X_j na primeira FDL. Como b_j depende das unidades de medida de X_j, para fins de interpretação é conveniente definir também os coeficiente padronizados ($b_j^{\star}$) que se obtém multiplicando b_j por $s_w(X_j)$, o desvio padrão intra-grupos de X_j.

$$b_j^{\star} = b_j * s_w(X_j)\ ; \qquad s_w(X_j) = \sqrt{\mathbf{W}_{jj} / (N - k)}$$

---

[3] Note-se que se $k\text{-}1 < p$, como é habitualmente o caso, **B** será uma matriz singular. O mesmo acontece para os produtos matriciais $\mathbf{B}\,\mathbf{W}^{-1}$ e $\mathbf{B}\,\mathbf{T}^{-1}$ que tem a mesma caracteristica que **B**.

[4] É bem sabido que $\mathbf{B}\,\mathbf{W}^{-1}$ e $\mathbf{B}\,\mathbf{T}^{-1}$ têm os mesmos vectores próprios e que os valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ (λ_i) e $\mathbf{B}\,\mathbf{T}^{-1}$ (l_i) estão relacionados pela expressão $l_i = \lambda_i / (1+\lambda_i)$. Por conseguinte as técnicas de ADD podem ser apresentadas em termos dos vectores e valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ ou de $\mathbf{B}\,\mathbf{T}^{-1}$. Nós utilizaremos indistintamente qualquer um destes produtos, consoante o que a cada momento seja mais conveniente para a exposição.

Diz-se então que a primeira FDL define a primeira dimensão de separação entre os grupos, procedendo-se em seguida à tentativa da sua interpretação em termos de algum conceito teórico subjacente que lhe esteja associado. De igual modo, o vector próprio normalizado associado ao segundo valor próprio de $\mathbf{B}\,\mathbf{W}^{-1}$, define uma combinação linear, não correlacionada intra-grupos[5] com Z1, que maximiza o rácio entre a inércia entre-grupos não explicada pela primeira FDL, e a inércia intra-grupos. Diz-se então que este vector define a segunda dimensão de separação. Prosseguindo deste modo é possível determinar um máximo de *r* FDLs não correlacionadas[6] (intra-grupos)[7] entre si, que maximizam sucessivamente a "separação" ainda não explicada previamente, e a que se poderão eventualmente associar conceitos teóricos identificados com dimensões de separação. Os valores próprios de $B\ W^{-1}$ são habitualmente interpretados como sendo proporcionais à inércia entre-grupos explicada pela respectiva dimensão de separação (Huberty 1994, 214).

As FDLs são tipicamente interpretadas com base ou nos seus coeficientes padronizados, ou nas suas correlações intra-grupos com as variáveis originais, correlações essas que também se designam por correlações estruturais. Quando o número de variáveis é elevado, é comum ignorar para efeitos de interpretação, aquelas cujos coeficientes padronizados ou correlações estruturais são menores em valor absoluto. A escolha entre coeficientes padronizados e correlações estruturais não é indiferente nem pacífica. Com efeito, é frequente que coeficientes e correlações sugiram interpretações divergentes, não estando completamente resolvido o problema de saber como as conciliar (ver, por exemplo, McKay e Campbell, 1982, 9; Huberty, 1994, 262-264). A utilização dos coeficientes padronizados surge em parte por analogia com procedimentos semelhantes em análise de regressão, com base na ideia intuitiva de que as variáveis que mais fortemente contribuem para a definição de uma FDL são aquelas que mais facilmente a ajudarão a interpretar. A utilização das correlações estruturais surge em consonância com práticas habituais em outras técnicas factoriais, com base na ideia de que variáveis fortemente correlacionadas com uma FDL tenderão a partilhar aquilo que ela representa (ou pelo menos a ser influenciadas por causas comuns). No entanto, tendo em atenção a sua natureza e características próprias, coeficientes e correlações poderão ser utilizados quer de forma alternativa quer de forma complementar. Nomeadamente, FDLs que fundamentalmente representam contrastes, poderão estar fracamente correlacionadas com todas as variáveis originais. Nesse caso os coeficientes padronizados serão mais

---

[5] A correlação intra-grupos entre as variáveis Z1 e Z2 define-se pela fórmula:

$$r_{w(Z1,Z2)} = \frac{\sum_{g=1}^{k}\sum_{i=1}^{n_g}(Z1_{gi} - \overline{Z1}_g)(Z2_{gi} - \overline{Z2}_g)}{\sqrt{\sum_{g=1}^{k}\sum_{i=1}^{n_g}(Z1_{gi} - \overline{Z1}_g)^2 \sum_{g=1}^{k}\sum_{i=1}^{n_g}(Z2_{gi} - \overline{Z2}_g)^2}}$$

De forma semelhante é possivel definir correlações entre-grupos (r_B) e correlações totais (r_T) substituindo desvios intra-grupos por desvios entre-grupos ou desvios totais.

[6] Mais própriamente, as variáveis definidas pelas FDLs não estão correlacionadas entre si. A fim de simplificar a exposição, recorreremos frequentemente ao abuso de linguagem que consiste em chamar "correlações com FDLs" às correlações com as variáveis definidas por essas FDLs.

[7] Pode-se demonstrar que as FDLs tem correlações entre-grupos e totais igualmente nulas (Kobilinski 1990). No entanto, iremos focar a exposição nas correlações intra-grupos, uma vez que elas são as correlações mais importante em problemas de ADD.

úteis para a sua interpretação. Por outro lado, uma variável relativamente mal representada numa FDL pode estar fortemente correlacionada com ela, facilitando a identificação de um conceito que lhe esteja associado. Finalmente, a própria posição relativa dos "scores" de cada indivíduo nas FDLs ou das suas médias por grupo pode auxiliar a sua interpretação.

2.2 EXEMPLO: CARACTERIZAÇÃO DE DIFERENÇAS ENTRE INSTITUIÇÕES BANCÁRIAS A OPERAR EM PORTUGAL

A fim de tentar caracterizar as principais diferenças, em Dezembro de 1993, entre as instituições bancárias nacionais criadas antes e depois da aprovação da denominada "lei de Delimitação de Sectores" (1984) e as instituições bancárias estrangeiras a operar em Portugal, construíram-se vários indicadores de desempenho económico-financeiro a partir dos respectivos balanços e demonstrações de resultados. Incluíram-se na análise todas as instituições bancárias a operar em Portugal em 1993, com excepção do Banco Comercial de Macau e de instituições que, ainda que legalmente instituidas como Bancos, não exerciam à data em Portugal, uma actividade bancária relevante. O Banco Comercial de Macau foi excluído porque estando em fase de desagregação após a sua aquisição pelo BCP nesse mesmo ano, apresentava valores atípicos para vários indicadores. Ao todo foram analisadas 33 instituições: 14 bancos nacionais anteriores a 1984, 7 bancos nacionais posteriores a 1984 e 12 bancos estrangeiros. A lista das instituições analisadas pode ser consultada no quadro 1.

QUADRO 1

INSTITUIÇÕES BANCÁRIAS E SCORES CENTRADOS DAS FDLs

VARIÁVEIS: LR, CCG, ln TRCC, GE, ln SB, TMA, TMR, MF, MN, RCPE, RCPD, ln EB, ALE, RBA, RBCP, RA, RCP

INSTITUÍÇÃO	TIPO	Z1	Z2
ABN	ESTRANGEIRO	5.075	0.933
BANIF	NOVO	-3.163	-2.064
BARCLAYS	ESTRANGEIRO	3.607	-0.926
BANCO DO BRASIL	ESTRANGEIRO	4.078	-0.190
BBI	ANTIGO	-3.244	4.397
BBV	ESTRANGEIRO	2.678	-0.365
BCA	ANTIGO	-2.495	2.060
BCI	NOVO	-1.478	-0.941
BCP	NOVO	-2.160	-3.359
BESCL	ANTIGO	-2.093	1.767
BEX	ESTRANGEIRO	2.837	-0.300
BFB	ANTIGO	-2.414	0.195
BFE	ANTIGO	-1.307	-0.007
BIC	NOVO	-2.416	-3.057
BNC	NOVO	-0.139	-3.100
BNP	ESTRANGEIRO	4.947	0.776
BNU	ANTIGO	-0.882	1.780
BPA	ANTIGO	-2.520	2.347
BPI	NOVO	-3.175	-3.344
BPSM	ANTIGO	-3.380	1.105
BTA	ANTIGO	-2.255	2.184
BTQ	ESTRANGEIRO	5.623	-0.047
CGD	ANTIGO	-3.989	-0.008
CHEMICAL	NOVO	-1.645	-4.185
CITI	ESTRANGEIRO	2.786	0.546
CL	ESTRANGEIRO	3.754	1.406
CPP	ANTIGO	-1.458	0.963
DBI	ESTRANGEIRO	3.761	-0.411
GENERALE	ESTRANGEIRO	2.249	-0.204
HISPANO	ESTRANGEIRO	4.527	-0.422
UBP	ANTIGO	-2.807	0.635
MELLO	ANTIGO	-1.101	0.601
MG	ANTIGO	-1.800	1.234

Cada banco foi inicialmente descrito por um conjunto de 17 indicadores que caracterizam diversos aspectos da sua estrutura patrimonial, funcionamento, e rendibilidade. Esses indicadores constam do quadro 2. O quadro 3 apresenta várias estatísticas descritivas das distribuições desses indicadores em cada um dos grupos considerados, bem como o valor das estatísticas F* resultantes de Análises de Variância (ANOVA) efectuadas para cada um deles.

Relativamente às estatísticas F*, convém notar que, para vários indicadores, a distribuição nula habitual (F de Snedecor) não deve ser válida, uma vez que os

pressupostos da normalidade e igualdade de variâncias não parecem razoáveis. Como se pode verificar pelos resultados de testes de normalidade de Kolmogorov-Smirnov (com a correcção de Lilliefors) e de homogeneidade de variâncias de Levene, sumarizados no quadro 2, estas hipóteses foram frequentemente rejeitadas ao nível de significância de 5%. Por essa razão, incluiu-se como equivalente não-paramétrico de F*, o valor das estatísticas χ^{2*} resultantes da realização de testes de Kruskall-Wills. Manteve-se, apesar disso, a apresentação das estatísticas F* uma vez que elas podem ser simplesmente interpretadas como medidas descritivas univariadas da capacidade de cada variável em separar os grupos.

Importa aqui referir que grande parte das técnicas de ADD que vão ser discutidas neste artigo podem ser entendidas como técnicas essencialmente exploratórias, justificadas com base em argumentos não-paramétricos. No entanto estas técnicas, por um lado podem ser negativamente afectadas pela existência de "outliers", e por outro lado assumem implicitamente variâncias e covariâncias idênticas para todos os grupos. Por estas razões, é discutível a utilização das técnicas clássicas de ADD na presença de distribuições com caudas demasiado pesadas, e/ou matrizes de covariância substancialmente diferentes de grupo para grupo. Por vezes em ADD, é conveniente recorrer a testes de hipóteses que admitem explicitamente os pressupostos de normalidade multivariada e igualdade das matrizes de covariância. Para reduzir desvios em relação a estes pressupostos ou simplesmente para evitar condições adversas, é comum recorrer a transformações de variáveis. A utilização de transformações tem no entanto a desvantagem de tornar a interpretação dos resultados menos directa, sendo recomendável apenas na presença de condições particularmente adversas ou de desvios substanciais em relação aos pressupostos clássicos. Na presença de desvios moderados, é geralmente preferível trabalhar com as variáveis originais nomeadamente porque muitos dos testes usados em ADD[8] são relativamente robustos (ver Seber 1984, 440-442). Nesta aplicação, embora a hipótese de nomalidade seja frequentemente violada, uma análise dos coeficientes de assimetria (que variam entre –1.495 e 2.639) e achatamento (variando entre -2.607 e 7.158) não revela distribuições com caudas suficientemente pesadas para que, no nosso entender, se justifique a utilização de medidas correctoras. No entanto, para alguns indicadores existem diferenças marcadas quanto à dispersão por grupo, tendo-se procedido a transformações logarítmicas em três variáveis (TRCC, RCPD e EB) em que este problema era mais grave. Pelas razões apontadas, sempre que recorrermos a testes de hipóteses, deveremos interpretar os resultados obtidos apenas como indicações, não sendo os valores de prova ("p-values") a referir nem a distribuições nulas das estatísticas utilizadas, estritamente válidos.

A fim de tentar encontrar dimensões que explicassem as principais diferenças entre estes três grupos de bancos, calcularam-se os coeficientes das duas FDLs. Os valores destes coeficientes encontram-se no quadro 4, juntamente com as correlações estruturais.

[8] A maioria dos testes de hipóteses utilizados em ADD são testes de Análise Multivariada de Variância (MANOVA) e Análise Multivariada de Covariância (MANCOVA). Como se tornará evidente ao longo do texto, existe uma ligação estreita entre ADD e MANOVA, abordando estas duas metodologias, ainda que sob prespectivas diferentes, problemas fortemente relacionados.

QUADRO 2

INDICADORES DE ESTRUTURA PATRIMONIAL,

FUNCIONAMENTO E RENDIBILIDADE

INDICADOR	ABREV.	DEFINIÇÃO*
Liquidez Reduzida	**LR**	L/PF
Capacidade Creditícia Geral	**CCG**	A/PF
Transformação dos Recursos de Clientes em Crédito	**TRCC**	A/RC
Grau de Endividamento	**GE**	D/FP
Solvabilidade Bruta	**SB**	FP/AL
Taxa Média das Aplicações	**TMA**	JA/AF
Taxa Média dos Recursos	**TMR**	JP/PF
Margem Financeira	**MF**	RF/AF
Margem de Negócio	**MN**	PB/AF
Relevância dos Custos Pessoal	**RCPE**	CP/CA
Relevância Custos no Produto	**RCPD**	CO/PB
Número de Empregados por Balcão	**EB**	NP/NB
Activo Líquido por Empregado	**ALE**	AL/NP
Rendibilidade Bruta do Activo	**RBA**	RBT/AL
Rendibilidade Bruta Capitais Próprios	**RBCP**	RBT/KP
Rendibilidade do Activo	**RA**	RL/AL
Rendibilidade dos Capitais Próprios	**RCP**	RL/KP

***VARIÁVEIS DE GESTÃO BANCÁRIA**

1. DO BALANÇO (Valores finais)		**2. DA CONTA DE EXPLORAÇÃO**	
	SIMB		**SIMB**
1.0- Cx. Dep. Bancos Centrais	L	2.1- Juros e Proveitos Equiparados	JA
1.1- Crédito s/ Inst. Crédito		2.2- Juros e Custos Equiparados	JP
1.2- Crédito s/ Clientes (Bruto)	A	2.3- Resultado Financeiro	RF
1.3- Títulos Rendimento Fixo (Bruto)		2.4- Outros Resultados Correntes	ORC
1.4- Activo Financeiro (Bruto)	AF	2.5- Produto Bancário	PB
1.5- Activo Bruto	AB	2.6- Custos com Pessoal	CP
1.6- Activo Líquido	AL	2.7- Outros Gastos Administrativos	OGA
1.7- Débitos à Vista	DV	2.8- Custos Administrativos	CA
1.8- Débitos a Prazo	DP	2.9- Resultado Bruto Exploração	RBE
1.9- Débitos Repres. Títulos	DT	2.10- Resultados Extraordinários	RX
1.10- Passivos Subordinados	PS	2.11- Resultado Bruto Total	RBT
1.11- Passivo Financeiro	PF	2.12- Amortizações e Provisões	DAP
1.12- Fundos Próprios	FP	2.13- Resultados antes Impostos	RAI
1.13- Capital, Reservas e Res.Transitados	KP	2.13- Impostos sobre Lucros	I
1.14- Recursos de Clientes e Títulos	RC	2.15- Resultado Líquido	RL
3. OUTROS DADOS			
	SIMB		
3.1- Número de Balcões Domésticos	NB		
3.2- Número de Empregados Domésticos	NP		

QUADRO 3

ANÁLISE UNIVARIADA

		NOVOS	ANTIGOS	ESTRANG.	**F*** (p-value)	**K-W χ^2*** (p-value)
LR	MÉDIA	12.666	15.261	9.934	**2.717** (0.082) (IVR)	**3.945** (0.139)
	D. P.	6.372	3.891	7.214		
	C. ASSIM.	-0.462	-0.233	0.263		
	C. ACHAT.	1.428	0.739	-1.380		
CCG	MÉDIA	75.691	68.191	94.341	**6.774** (0.004) (NR)	**12.955** (0.002)
	D. P.	12.106	10.055	26.706		
	C. ASSIM.	-1.418	0.989	2.283		
	C. ACHAT.	1.780	1.489	6.491		
TRCC	MÉDIA	146.266	86.575	263.507	**9.163** (0.001) (NR ; IVR)	**21.145** (< 0.001)
	D. P.	104.544	21.024	154.966		
	C. ASSIM.	2.505	2.168	0.795		
	C. ACHAT.	6.395	6.611	-0.769		
ln TRCC	MÉDIA	4.846	4.438	5.420	**16.294** (< 0.001) (NR ; IVR)	**21.145** (< 0.001)
	D. P.	0.505	0.212	0.574		
	C. ASSIM.	2.200	1.277	0.379		
	C. ACHAT.	5.071	3.528	-1.645		
GE	MÉDIA	1.065	0.569	0.951	**1.024** (0.371) (NR ; IVR)	**0.830** (0.660)
	D. P.	0.943	0.456	1.130		
	C. ASSIM.	0.443	0.799	1.428		
	C. ACHAT.	-2.414	0.159	1.694		
SB	MÉDIA	11.455	5.642	11.692	**3.809** (0.034) (NR)	**9.274** (0.010)
	D. P.	5.207	1.990	9.122		
	C. ASSIM.	0.589	1.273	2.271		
	C. ACHAT.	0.187	1.657	6.771		
TMA	MÉDIA	12.765	12.493	11.638	**1.031** (0.369) (IVR)	**1.791** (0.408)
	D. P.	0.908	0.930	2.832		
	C. ASSIM.	-1.495	0.266	0.577		
	C. ACHAT.	2.466	-1.197	0.035		
TMR	MÉDIA	9.767	8.752	9.884	**1.497** (0.240) (IVR)	**2.292** (0.318)
	D. P.	1.451	0.867	2.594		
	C. ASSIM.	0.048	0.501	0.952		
	C. ACHAT.	0.612	-0.388	0.834		
MF	MÉDIA	3.697	4.019	2.694	**2.945** (0.068) (NR)	**13.337** (0.001)
	D. P.	0.780	0.753	2.113		
	C. ASSIM.	-0.359	2.381	2.328		
	C. ACHAT.	-2.607	7.158	7.119		
MN	MÉDIA	5.091	5.042	3.592	**3.329** (0.049) (NR)	**12.449** (0.002)
	D. P.	0.854	1.178	2.166		
	C. ASSIM.	0.277	2.176	2.290		
	C. ACHAT.	-2.086	5.213	6.956		
RCPE	MÉDIA	61.316	68.999	54.551	**13.656** (< 0.001)	**15.663** (< 0.001)
	D. P.	4.727	5.843	9.094		
	C. ASSIM.	-0.026	-0.160	-0.531		
	C. ACHAT.	-2.205	0.177	-0.854		

RCPD	MÉDIA	60.840	65.336	74.760	**0.502** (0.611) (IVR)	**0.261** (0.877)
	D. P.	23.593	9.478	48.196		
	C. ASSIM.	-0.944	-0.462	1.996		
	C. ACHAT.	2.655	-0.309	5.081		
ln RCPD	MÉDIA	4.002	4.169	4.164	**0.398** (0.675) (NR ; IVR)	**0.261** (0.877)
	D. P.	0.574	0.152	0.552		
	C. ASSIM.	-2.117	-0.778	0.614		
	C. ACHAT.	5.171	0.297	0.001		
EB	MÉDIA	28.839	22.290	21.697	**0.418** (0.662) (NR ; IVR)	**3.588** (0.166)
	D. P.	36.251	5.781	9.432		
	C. ASSIM.	2.639	1.508	0.053		
	C. ACHAT.	6.973	2.587	-0.804		
ln EB	MÉDIA	2.999	3.077	2.974	**0.160** (0.853) (NR)	**3.588** (0.166)
	D. P.	0.759	0.235	0.501		
	C. ASSIM.	2.583	0.961	-0.608		
	C. ACHAT.	6.742	0.731	-0.896		
ALE	MÉDIA	1021.630	446.299	1280.675	**2.593** (0.092) (NR ; IVR)	**7.107** (0.029)
	D. P.	1109.128	350.341	1286.471		
	C. ASSIM.	1.758	2.322	1.585		
	C. ACHAT.	2.494	6.273	1.302		
RBA	MÉDIA	0.0300	0.0281	0.0209	**1.397** (0.263)	**7.599** (0.022)
	D. P.	0.0086	0.0104	0.0174		
	C. ASSIM.	0.510	2.207	1.972		
	C. ACHAT.	-1.395	6.201	5.779		
RBCP	MÉDIA	0.298	0.549	0.214	**9.841** (0.001)	**16.194** (< 0.001)
	D. P.	0.119	0.265	0.129		
	C. ASSIM.	0.574	1.589	0.314		
	C. ACHAT.	-0.976	3.244	0.295		
RA	MÉDIA	0.0083	0.0048	0.0015	**1.861** (0.173) (NR)	**2.956** (0.228)
	D. P.	0.0087	0.0030	0.0100		
	C. ASSIM.	1.653	0.234	-1.688		
	C. ACHAT.	3.104	-1.330	4.941		
RCP	MÉDIA	0.0796	0.0877	0.0202	**3.152** (0.057) (NR)	**5.047** (0.080)
	D. P.	0.0644	0.0560	0.0897		
	C. ASSIM.	0.102	0.620	-2.374		
	C. ACHAT.	-1.835	-0.942	6.969		

Legenda:

ln -- Logaritmo Neperiano.

NR -- Hipótese de normalidade rejeitada (para $\alpha = 0.05$) por um teste de Kolmogorov-Sminorv com a correção de Lilliefors.

IVR -- Hipótese de igualdade de variâncias rejeitada (para $\alpha = 0.05$) por um teste de Levene.

QUADRO 4

FUNÇÕES DISCRIMINANTES LINEARES

VARIÁVEIS: LR, CCG, ln TRCC, GE, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RA, RCP

		FDL1			FDL2	
VAR	Coef. não Padron.	Coef. Padron.	Corr.	Coef. não Padron.	Coef. Padron.	Corr.
LR	0.29077	1.690	-0.129	0.01388	0.081	0.098
CCG	0.02791	0.511	0.216	0.02859	0.523	-0.084
ln TRCC	6.37803	2.791	0.325	1.30648	0.572	-0.201
GE	0.90519	0.777	0.044	-0.92802	-0.796	-0.134
SB	-0.51589	-3.165	0.115	-0.58152	-3.568	-0.217
TMA	-4.43787	-8.278	-0.084	-3.72430	-6.947	-0.039
TMR	4.15851	7.455	0.076	3.52955	6.328	-0.129
MF	6.06368	8.582	-0.144	5.57241	7.887	0.045
MN	-1.74939	-2.748	-0.155	-1.12865	-1.773	-0.016
RCPE	-0.07429	-0.523	-0.285	-0.21417	-1.508	0.241
ln RCPD	5.22460	2.263	0.018	-0.31996	-0.139	0.092
ln EB	-1.45883	-0.701	-0.027	1.34778	0.648	0.037
ALE	0.00070	0.669	0.115	-0.00169	-1.605	-0.136
RBA	55.83055	0.733	-0.098	-73.93866	-0.971	-0.039
RBCP	3.89523	0.774	-0.215	9.64063	1.916	0.286
RA	98.57053	0.735	-0.098	16.32608	0.122	-0.114
RCP	-9.94804	-0.713	-0.151	-5.54440	-0.397	0.018

QUADRO 5

SCORES MÉDIOS DAS FDLs EM CADA GRUPO
(SCORES CENTRADOS NA ORIGEM)

VARIÁVEIS: LR, CCG, ln TRCC, GE, ln SB, TMA, TMR, MF, MN, RCPE, RCPD, ln EB, ALE, RBA, RBCP, RA, RCP

	NOVOS	ANTIGOS	ESTRANGEIROS
Z1	-2.025	-2.268	3.827
Z2	-2.864	1.375	0.067

Os scores das FDLs e as sua médias em cada grupo encontram-se representados nos quadros 1 e 5. A primeira FDL explica 76.7% da inéricia entre-grupos (λ_1 = 9.214) e a segunda os restantes 23.3% (λ_2 = 2.799).

Uma primeira análise dos coeficientes padronizados, sugere que a primeira FDL representa essencialmente um contraste entre a *Taxa Média das Aplicações* versus a *Taxa Média de Recursos* e a *Margem Financeira*. No entanto, esta interpretação é discutível, uma vez que sendo a *Margem Financeira* aproximadamente igual à diferença entre a *Taxa Média das Aplicações* e a *Taxa Média de Recursos*, estas três variáveis tendem a anular-se. Seguem-se em ordem de importância, os coeficientes relativos à *Solvabilidade Bruta* e à *Margem de Negócio*, ambos com sinal negativo, e aos logaritmos da *Transformação dos Recursos de Clientes em Crédito* e da *Relevância dos Custos do Produto* com sinal positivo. Nenhuma dos variáveis originais está fortemente correlacionada com a primeira FDL. As correlações estruturais mais importantes dizem respeito ao logaritmo da *Transformação dos Recursos de Clientes em Crédito* (r = 0.325) e à *Relevância dos Custos com o Pessoal* (r = -0.285). Conjugando estes resultados, Z1 poder-se-à interpretar como um indicador ligado à estrutura de exploração. "Scores" elevados nesta variável indicam estruturas mais leves e gestões mais activas dos créditos a clientes. Esta dimensão permite sobretudo distinguir os bancos estrangeiros dos bancos nacionais.

Os três coeficientes padronizados mais importantes na segunda FDL sugerem igualmente um contraste entre a *Taxa Média das Aplicações* versus a *Taxa Média de Recursos* e a *Margem Financeira*. No entanto, pelas razões já acima apontadas a importância destas variáveis é mais aparente que real. O quarto coeficiente mais importante é o coeficiente associado à *Solvabilidade Bruta* com sinal negativo, seguindo-se a *Rendibilidade Bruta dos Capitais Próprios* com sinal positivo. A variável mais fortemente correlacionada com Z2 é a *Rendibilidade Bruta dos Capitais Próprios* (r = 0.286). Estes resultados sugerem que Z2 é essencialmente um indicador de estrutura patrimonial. Scores elevados nesta variável indicam uma estrutura caracterizada por um peso reduzido peso dos capitais próprios. Esta dimensão permite distinguir sobretudo os bancos nacionais posteriores a 1984 dos bancos nacionais anteriores a esta data, verificando-se que os bancos mais recentes mostram uma maior tendência a recorrer a capitais próprios.

2.3 TESTES DE INFORMAÇÃO ADICIONAL

A existência de FDLs expressas a partir de um número elevado de variáveis, grande parte das quais acabam na prática por ser ignoradas, levanta a questão de saber se não se poderá efectuar a análise com base unicamente nas variáveis que de facto parecem contribuir para as diferenças observadas. Esta questão é importante na medida em que a inclusão de variáveis irrelevantes ou redundantes pode introduzir um grau considerável de variabilidade amostral dificultando o reconhecimento das verdadeiras causas de separação. Para determinar se um determinado subconjunto de variáveis é irrelevante podem utilizar-se os chamados testes de informação adicional (Rao 1973, Seber 1984, 471-472) que são na realidade casos particulares de testes multivariados de análise de covariância (MANCOVA). Nomeadamente, suponha-se que se pretende testar se um determinado subconjunto de variáveis, *Q*, contem toda a informação relevante para a separação dos grupos, ou de forma equivalente se o seu

complementar, $\overline{Q}$, não contribui para as diferenças. Esta hipótese pode formalizar-se da seguinte forma

$$H_0 : \mathbf{E}(\mathbf{X}_{g\overline{Q}} \mid \mathbf{X}_{gQ}) = \mathbf{E}(\mathbf{X}_{g'\overline{Q}} \mid \mathbf{X}_{g'Q}) \qquad \forall g, g' = 1,2,..., k$$

ou seja, as médias condicionais das variáveis não incluídas em *Q* são idênticas para todos os grupos. No caso de só existirem dois grupos (*k* = 2) esta hipótese é equivalente à hipótese de que as distâncias de Mahalanobis (populacionais) entre as médias de grupos baseadas em *Q*, sejam idênticas às distâncias de Mahalanobis que consideram o conjunto completo de variáveis (Krishnaia 1982). Quando se verifica H_0, diz-se que $\overline{Q}$ não contem informação adicional para a separação entre os grupos e que o conjunto *Q* é um "conjunto adequado".

Para testar H_0 é conveniente estabelecer as partições de **W** e **T**

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{QQ} & \boldsymbol{W}_{Q\overline{Q}} \\ \boldsymbol{W}_{\overline{Q}Q} & \boldsymbol{W}_{\overline{Q}\overline{Q}} \end{bmatrix} \qquad \boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{QQ} & \boldsymbol{T}_{Q\overline{Q}} \\ \boldsymbol{T}_{\overline{Q}Q} & \boldsymbol{T}_{\overline{Q}\overline{Q}} \end{bmatrix}$$

e definir a matrizes de somas de desvios quadráticos e cruzados condicionais

$$\mathbf{W}_{\overline{Q}|Q} = \mathbf{W}_{\overline{Q}\overline{Q}} - \mathbf{W}_{\overline{Q}Q}\mathbf{W}_{QQ}^{-1}\mathbf{W}_{Q\overline{Q}} ; \mathbf{T}_{\overline{Q}|Q} = \mathbf{T}_{\overline{Q}\overline{Q}} - \mathbf{T}_{\overline{Q}Q}\mathbf{T}_{QQ}^{-1}\mathbf{T}_{Q\overline{Q}} ; \mathbf{B}_{\overline{Q}|Q} = \mathbf{T}_{\overline{Q}|Q} - \mathbf{W}_{\overline{Q}|Q}$$

Admitindo que os vectores aleatórios $X_g = [X_{g1},...,X_{gp}]^T$ seguem distribuições normais multivariadas com matrizes de covariância idênticas, H_0 pode ser testada com base nas estatísticas MANCOVA habituais, nomeadamente o maior valor próprio de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{W}_{\overline{Q}|Q}^{-1}$ (primeiro valor próprio de Roy), a soma dos valores próprios de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{W}_{\overline{Q}|Q}^{-1}$ (traço de Lawley-Hotelling), a soma dos valores próprios de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{T}_{\overline{Q}|Q}^{-1}$ (traço de Bartlett-Pillai), ou o produto dos complementares para a unidade dos valores próprios de $\boldsymbol{B}_{\overline{Q}|Q}\boldsymbol{T}_{\overline{Q}|Q}^{-1}$ (lambda de Wilks).

Para que estes testes sejam estatisticamente válidos, para alem da verificação dos pressupostos clássicos atrás referidos, é ainda necessário que o subconjunto *Q* tenha sido escolhido à partida. Na prática, testes de informação adicional são frequentemente utilizados com o conjunto *Q* sugerido por uma análise de coeficientes padronizados ou correlações estruturais. O principal problema desta estratégia reside no facto que quando *Q* é sugerido pelos dados, a probabilidade (sob H_0) de se rejeitar a sua adequação tende a ser menor do que a probabilidade expressa no nível de significância nominal.

Voltando ao exemplo da secção anterior, a interpretação final de FDL_1 e FDL_2 baseou-se fundamentalmente no conjunto *Q1* formado pelas seguintes variáveis: ln TRCC, SB, MN, RCPE, ln RCPD, RBCP. Levanta-se a questão de saber se as restantes variáveis acrescentam alguma informação à capacidade explicativa de *Q1*. O

quadro 6 apresenta o valor das estatísticas MANCOVA referentes aos testes sobre a informação adicional de $\overline{Q}$. Qualquer que seja a estatística considerada, rejeita-se sempre, ao nível de significância de 5%[9], a hipótese de *Q1* ser um conjunto adequado. Por conseguinte, há evidência de que o conjunto de variáveis que foi, de facto, utilizado para interpretar as duas dimensões de separação não contem toda a informação relevante para explicar as diferenças entre estes três grupos.

QUADRO 6

TESTES DE INFORMAÇÃO ADICIONAL

VARIÁVEIS: LR, CCG, GE, TMA, TMR, MF, ln EB, ALE, RBA, RA, RBCP

CRITÉRIO	EST. MULTIV.	F*	**GR. DE LIB.**	P.VALUE
ROY	2.808	4.212	10 ; 15	0.005
WILKS	0.115	2.736	20 ; 28	0.007
BARTLETT-PILLAI	1.301	2.792	20 ; 30	0.010
LAWLEY- HOTELLING	4.100	2.665	20 ; 26	0.006

2.4 MÉTODOS DE SELECÇÃO PASSO A PASSO

Muitas vezes, à partida não é claro quais serão os melhores candidatos para menores conjuntos adequados. Pretendendo-se minimizar a realização de escolhas subjectivas no decorrer da análise, é comum recorrer a métodos de selecção automática, que num certo sentido pretendem deixar "os dados falar por si". Os mais populares destes métodos são métodos de selecção passo a passo baseados em testes de informação adicional. Nomeadamente, considere-se a seguinte factorização da estatística global de Wilks, $\Lambda = |\mathbf{W}| / |\mathbf{T}|$ (ou seja, Λ é a estatística de Wilks relativa à hipótese nula de não existência de diferenças entre os grupos)

$$\ddot{E} = \ddot{E}_{j_1} * \ddot{E}_{j_2|j_1} * \ddot{E}_{j_3|j_1,j_2} * \ldots * \ddot{E}_{j_p|j_1,j_2,\ldots,j_{p-1}}$$

Em que

$$\ddot{E}_{j_1} = \frac{W_{j_1 j_1}}{\hat{O}_{j_1 j_1}}, \quad \ddot{E}_{j_i|Q} = \frac{W_{j_i|Q}}{T_{j_i|Q}} = \frac{W_{j_i j_i} - W_{j_i Q} W_{QQ}^{-1} W_{Q j_i}}{T_{j_i j_i} - T_{j_i Q} T_{QQ}^{-1} T_{Q j_i}}$$

Aqui, $\Lambda_{j_i|Q}$ é a estatística de Wilks relativa a um teste sobre a informação adicional de X_j dado *Q*. Se um teste, dado um conjunto, *Q* com *q* variáveis, tivesse

[9] Note-se no entanto, que nenhum dos valores de prova (p-values) indicados no quadro 6 são exactos porque, para além das aproximações assintóticas habituais (que neste caso só não necessárias para a estatística de Wilks), os pressupostos de normalidade multivariada e igualdade de matrizes de covariâncias não se verificam. Com efeito, vimos na secção 2.2 que as condições necessárias (mas não suficientes) de normalidade univariada e igualdade de variâncias não se verificavam para várias variáveis. Se essas condições se verificassem, para garantir a validade destes testes teríamos ainda que verificar a normalidade multivariada e a igualdade das matrizes populacionais de covariância.

sido decidido à partida, F* = [(*N-k-q) / (k*-1)] * [(1-$\ddot{E}_{j_i|Q}$) / $\ddot{E}_{j_i|Q}$], seguiria sob a hipótese nula uma distribuição F de Snedecor com *k-1* e *N-k-q* graus de liberdade. Métodos ascendentes de selecção partem do conjunto vazio e escolhem a cada passo a variável $\mathbf{X}_{j_i} \in \overline{Q}$ que maximiza o valor de F*, para possível inclusão em *Q*. Métodos descendentes, partem do conjunto formado por todas as variáveis e escolhem a cada passo a variável $\mathbf{X}_{j_i} \in Q$ que minimiza o valor de F*, para possível eliminação. Existem ainda métodos mistos que combinam estas duas estratégias. Em qualquer dos casos, pontos críticos das respectivas distribuições F, são muitas vezes utilizados como valores de referência para decidir quando prosseguir com a inclusão/eliminação de variáveis ou terminar o processo. No entanto, devido à existência de selecção e à realização de várias comparações, inferências baseadas na distribuição F não são de facto válidas e estes procedimentos só podem ser justificados de uma forma heurística. Note-se que apesar de este método ter sido apresentado em termos da estatística de Wilks, ele poderia ter sido formulado em função de qualquer das outras estatísticas de informação adicional habituais. Como $\mathbf{W}_{j_i|Q}$, $\mathbf{T}_{j_i|Q}$ e $\mathbf{B}_{j_i|Q} = \mathbf{T}_{j_i|Q} - \mathbf{W}_{j_i|Q}$ tem dimensão (1*1), $\mathbf{B}_{j_i|Q}\ \mathbf{W}^{-1}_{j_i|Q}$ só tem um valor próprio e todas as estatísticas clássicas conduzem a resultados equivalentes. Existem no entanto métodos de selecção passo a passo que utilizam critérios de selecção diferentes, nomeadamente a maximização da estatística global de Lawley-Hotelling (este critério também é referido como critério de Rao) ou critérios baseados em distâncias de Mahalanobis (ver McKay e Campbell, 1982, 13-14).

No exemplo que temos vindo a utilizar, um método ascendente (puro) de selecção passo a passo usando $\mathbf{\Lambda}_{j_i|Q}$ como critério de selecção, e a comparação entre F* e o 90º percentil da respectiva distribuição F como critério de paragem sugere o conjunto *Q2* = {CCG, ln TRCC, MN}. Os coeficientes padronizados da primeira FDL são respectivamente 0.748 (CCG), 0.585 (ln TRCC) e -0.828 (MN). As correlações estruturais são (pela mesma ordem), 0.453, 0.695 e -0.308. Estes valores são compatíveis com a interpretação anterior da primeira dimensão de separação, que continua a distinguir sobretudo os bancos estrangeiros dos bancos nacionais anteriores a 1984 (scores médios de 1.862 e -1.319 respectivamente) ocupando os bancos nacionais posteriores a 1984 uma posição intermédia (score médio -0.493). No entanto, a segunda dimensão de separação tem agora uma expressão muito reduzida (explica apenas 1.8% da inércia entre-grupos) parecendo ser apenas o resultado de variação amostral.

Um método descendente (puro) de selecção usando os mesmos critérios, sugere o conjunto *Q3* = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP}. Os coeficientes das respectivas FDLs, correlações estruturais e posições relativas dos seus scores médios encontram-se nos quadros 7 e 8. A primeira FDL explica 77.6 % (λ_1 = 8.346) da inércia entre-grupos e a segunda os restantes 22.4 % (λ_2 = 2.408). Um teste de informação adicional revela que este conjunto pode ser considerado adequado (p-value > 0.50). As interpretações das FDLs e as posições relativas dos scores médios de cada grupo são, neste caso, idênticas ás interpretações baseadas na análise do conjunto completo de 17 variáveis.

QUADRO 7

FUNÇÕES DISCRIMINANTES LINEARES

VARIÁVEIS: LR, ln TRCC, GE, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP

		FDL1			FDL2	
VAR	Coef. não Padron.	Coef. Padron.	Corr.	Coef. não Padron.	Coef. Padron.	Corr.
LR	0.32409	1.883	-0.137	0.05549	0.322	0.099
ln TRCC	6.99993	3.063	0.344	2.67158	1.169	-0.201
GE	1.29104	1.108	0.049	-0.70319	-0.604	-0.142
SB	-0.44495	-2.730	0.124	-0.58599	-3.596	-0.228
TMA	-4.55795	-8.502	-0.087	-3.71700	-6.933	-0.046
TMR	4.21365	7.554	0.082	3.43491	6.158	-0.136
MF	6.00060	8.493	-0.152	5.71432	8.088	0.042
MN	-1.26609	-1.989	-0.163	-1.52258	-2.392	-0.025
RCPE	-0.05767	-0.406	-0.303	-0.18357	-1.293	0.246
ln RCPD	4.60097	1.993	0.017	1.02920	0.446	0.100
Ln EB	-1.64765	-0.792	-0.029	1.29428	0.622	0.038
ALE	0.00068	0.643	0.122	-0.00159	-1.514	-0.141
RBCP	3.55562	0.707	-0.230	7.58427	1.507	0.298

QUADRO 8

SCORES MÉDIOS DAS FDLs EM CADA GRUPO (SCORES CENTRADOS NA ORIGEM)

VARIÁVEIS: LR, ln TRCC, GE, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP

	NOVOS	ANTIGOS	ESTRANGEIROS
Z1	-1.804	-2.216	3.638
Z2	-2.682	1.246	0.110

2.5 COMPARAÇÕES ENTRE TODOS OS SUBCONJUNTOS POSSIVEIS

McCabe (1975) defendeu que uma comparação entre todos os subconjuntos possíveis de variáveis, seria preferível à utilização de algoritmos de selecção passo a passo, uma vez que estes algoritmos podem não identificar contribuições para a separação decorrentes de combinações de variáveis que sejam substancialmente diferentes da soma das respectivas contribuições individuais. Este problema é mais marcante em algoritmos ascendentes do que em algoritmos descendentes (McKay e Campel 1982, 15). McCabe (1975) sugeriu usar o valor da estatística global de Wilks, entendido neste contexto como um mero índice de proximidade entre os grupos, como

critério de comparação e mostrou como o algoritmo de Furnival (1971) para comparação entre todos os subconjuntos possíveis em Análise de Regressão, pode ser adaptado para esse efeito. De acordo com McCabe (1975), é possível identificar os melhores (segundo o valor de Λ) subconjuntos de variáveis num "tempo razoável", desde que o número total de variáveis (p) não ultrapasse 20. Utilizando modernos computadores pessoais e substituindo o algoritmo de Furnival pelo algoritmo de Furnival e Wilson (1974), este limite poderá ser estendido, pelo menos até às 30 variáveis[10] (ver Duarte Silva 1998, para detalhes).

Para o exemplo que temos vindo a descrever, o quadro 9a) apresenta os dois melhores subconjuntos de cada dimensão de acordo com o valor de Λ. Os gráficos 1 e 2, mostram a evolução do índice $\tau^2 = 1-\Lambda^{1/r}$ para o melhor subconjunto de cada dimensão. Da análise do quadro 9a) ressaltam os seguintes resultados: O valor de Λ tende a decrescer de forma clara desde $q = 1$ até $q = 13$. A partir de $q = 13$, o valor de Λ parece estabilizar. As diferenças entre os valores de Λ para os dois melhores subconjuntos de cada dimensão também são claramente mais marcadas para $q = 13$ (assumindo o valor 0.0065) do que para $q > 13$ (onde nunca ultrapassa 0.0006). Estes resultados sugerem o melhor (de acordo com Λ) subconjunto com 13 variáveis, *Q3*, como um bom candidato para selecção. Curiosamente, *Q3* foi igualmente o conjunto seleccionado pelo método descendente de selecção passo a passo referido na secção 2.4. Em geral, não é possível no entanto garantir que um método de selecção passo a passo vá escolher o melhor subconjunto de uma dada dimensão de acordo com algum índice de separação (ou proximidade) nem um número de variáveis a partir do qual esse índice tenda a estabilizar. Ainda neste exemplo, o método ascendente de selecção passo a passo descrito em 2.4 tinha seleccionado o conjunto *Q2* = {CCG, ln TRCC, MN} que além de parecer ignorar a segunda dimensão de separação e utilizar um número de variáveis aparentemente bastante reduzido (Λ está longe de estabilizar para $q = 3$), está, de acordo com Λ, algo distante do melhor subconjunto da sua dimensão, *Q4* = {CCG, SB, MF}, uma vez que Λ assume os valores 0.274 e 0.300 para os conjuntos *Q4* e *Q2*, respectivamente. Finalmente, convém referir que testes de informação adicional permitem rejeitar a adequação de *Q2*, qualquer que seja a estatística utilizada (p-value < 0.02).

[10] O esforço computacional de algoritmos de pesquisa entre todos os subconjuntos possiveis cresce exponencialmente com p. Para o algoritmo de Furnival, que é um algoritmo de pesquisa exaustiva, o esforço dobra com a introdução de cada nova variável. Para $p = 20$, dependendo do computador utilizado, uma pesquisa completa poderá levar de algumas dezenas de segundos a alguns minutos, enquanto para $p = 30$ o tempo requerido será da ordem das horas ou das dezenas de horas. O algoritmo de Furnival e Wilson é um algoritmo de enumeração implícita onde o esforço computacional depende das diferenças entre os vários subconjuntos de variáveis quanto ao critério de comparação. Nas pior das hipóteses, se todos os subconjuntos tiverem contribuições semelhantes para a separação dos grupos, o esforço destes dois algoritmos será da mesma ordem de grandeza. No entanto, se alguns subconjuntos se destacarem dos restantes, pesquisas para $p = 30$ poderão demorar poucos minutos, e para p entre 40 e 50 poderão ser feitas em menos de 48 horas. Para detalhes sobre algoritmos de pesquisa entre todos os subconjuntos possíveis ver Duarte Silva (1998).

3. IDENTIFICAÇÃO DE SUBCONJUNTOS ADEQUADOS

3.1 TESTES DE HIPÓTESES SIMULTÂNEOS: A METODOLOGIA STP

Como se viu pela discussão apresentada na secção anterior, os métodos tradicionais de selecção de variáveis são fundamentalmente métodos heurísticos, que não são capazes de responder de uma forma estatisticamente válida ao problema da identificação dos subconjuntos adequados. Designe-se por $\boldsymbol{A} = \{Q : Q \text{ é adequado}\}$ o conjunto dos subconjuntos de variáveis que contem toda informação relevante para explicar as diferenças entre os grupos. Nesta secção ir-se-á discutir o problema de fazer inferências sobre $\boldsymbol{A}$. Em particular, far-se-á uma revisão de métodos que permitem encontrar conjuntos, $\boldsymbol{A}_\alpha$, que incluem todos os elementos de $\boldsymbol{A}$ com uma probabilidade não inferior a $1-\alpha$. Ao longo da discussão, admitir-se-ão os pressupostos de normalidade multivariada e igualdade das matrizes de covariância (populacionais) intra-grupos.

Como vimos na secção anterior, a adequação de um subconjunto particular pode ser testada com base numa estatística MANCOVA. Ao tentar fazer inferências sobre $\boldsymbol{A}$ é necessário testar simultaneamente a adequação de vários subconjuntos, o que requer o ajustamento dos níveis de significância individuais. Uma primeira abordagem para este problema consistiria em fazer testes de informação adicional para todos os subconjuntos possíveis com níveis de significância individuais, $\alpha_{IND} = \alpha / (2^P-1)$, ajustados pelo método de Bonferroni. No entanto, dado o elevado numero (2^P-1) de testes simultâneos a realizar, este procedimento seria demasiado conservador, o que neste caso teria como consequência a inclusão em $\boldsymbol{A}_\alpha$ de muitos conjuntos inadequados, para ser útil na prática. Existem no entanto métodos menos conservadores. Nomeadamente, Mckay e Campbell (1982) descrevem uma metodologia designada por STP (do inglês, Simultaneous Test Procedure) desenvolvida inicialmente por Gabriel (1968, 1969) e adaptado por Mckay (1977) para problemas de selecção de variáveis Análise Discriminante. O primeiro passo desta metodologia consiste em testar, com base numa das estatísticas MANOVA habituais, se a totalidade das variáveis originais revelam de facto diferenças entre os grupos. Este teste é conduzido ao nível global de significância, α. Se a hipótese nula, de médias idênticas para todos os grupos, for aceite não faz sentido continuar a análise. No caso de se concluir pela existência de diferenças significativas, em seguida prossegue-se com o cálculo, para cada subconjunto Q, de uma estatística MANCOVA respeitante à informação adicional contida em $\overline{Q}$. Essa estatística é comparada com o mesmo ponto crítico usado para o teste MANOVA inicial. Note-se que este ponto crítico é diferente daquele que seria usado para um teste MANCOVA individual decidido à partida. Por exemplo, se os testes forem conduzidos com base na estatística de Wilks, o nível de significância individual (α_{IND}) pode ser calculado através da igualdade

$$\Lambda(p\text{-}q,k\text{-}1,N\text{-}k\text{-}q;\ \alpha_{IND}) = \Lambda(p,k\text{-}1,N\text{-}k;\ \alpha)$$

onde $\Lambda(d,m_H,m_E;\ \alpha)$ é o ponto critico da distribuição de uma estatística Λ com parâmetros d, m_H e m_E (ver Seber 1984, 40-42), a um nível de significância α. Como o ponto crítico, $\Lambda(p,k\text{-}1,n\text{-}k;\ \alpha)$, para o teste MANOVA global, é inferior ao ponto critico $\Lambda(p\text{-}q,k\text{-}1,n\text{-}k\text{-}q;\ \alpha)$, para o teste MANCOVA sobre a adequação de Q, α_{IND} é inferior a α, sendo esta diferença entre α_{IND} e α que garante a protecção global da

bateria de testes. Testes baseados em qualquer das outras três estatísticas habituais são igualmente possíveis, sendo nesse caso o ponto critico efectivamente utilizado superior àquele que corresponderia a um único teste de informação adicional. Em qualquer dos casos, α_{IND} é sempre inferior a α, ou seja utilizam-se testes individuais conservadores a fim de se garantir uma protecção global ao nível α. No entanto, para se evitarem testes individuais demasiado conservadores é habitual escolherem-se valores relativamente altos para o nível de significância global.

Há três aspectos que se devem ter em consideração para a escolha da estatística em que se baseiam os STP. Em primeiro lugar, as diferenças entre α_{IND} e α variam consoante a estatística escolhida. Em particular, quando os STP são baseados no primeiro valor próprio de Roy, os testes individuais são menos conservadores do que em STP baseados em qualquer outra estatística (McCabe 1982, 19). Com base nesta característica, Gabriel (1968) descreve os STP baseados no primeiro valor próprio como sendo os STP mais *resolventes*. Os STP baseados na estatística de Wilks tendem a ser os menos *resolventes* enquanto os STP baseados nos traços de Lawley-Hotteling ou Bartllet-Pillai tendem a ter um comportamento intermédio. O segundo aspecto a considerar é a potência dos testes individuais. É sabido que o facto da estatística de Roy ignorar todos os valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ à excepção do primeiro, pode prejudicar seriamente a sua potência (Seber 1984, 414-416). Em particular, se as diferenças entre grupos não poderem ser explicadas ao longo de uma única dimensão, a potência da estatística de Wilks pode ser substancialmente superior à potência da estatística de Roy, sendo esta diferença de potências geralmente mais importante do que as diferenças devidas à *resolução* (McCabe 1982,). Finalmente, um terceiro aspecto de ordem mais prática tem a ver com razões de ordem computacional. Nomeadamente, dado que os STP consideram implicitamente todos os subconjuntos possíveis a sua aplicação pode tornar-se impraticável para um número moderado de variáveis. Existem no entanto duas formas de tentar ultrapassar este problema. Em primeiro lugar, reconhecendo uma propriedade de coerência dos STP, nenhum subconjunto de um conjunto excluído de $\mathbf{A}_\alpha$ pode fazer parte de $\mathbf{A}_\alpha$, nem todos os subconjuntos tem que ser testados quanto à sua adequação. Esta propriedade traduz o principio intuitivo de que a informação adicional de um subconjunto $\overline{Q}$ nunca pode diminuir quando se adicionam variáveis a $\overline{Q}$, e é uma consequência do facto de as estatísticas de informação adicional de Bartlett-Pillai, Lawley-Hotelling e Roy, nunca diminuírem quando se eliminam variáveis de *Q*, enquanto a estatística de Wilks nunca aumenta nas mesmas circunstâncias. Em segundo lugar, utilizando o algoritmo de Furnival-McCabe é possível calcular o valor da estatística de Wilks para todos os subconjuntos possíveis (Λ_Q)[11] de forma eficiente. Por sua vez, as estatísticas de Wilks sobre a informação adicional de $\overline{Q}$ dado *Q* ($\Lambda_{\overline{Q}|Q}$) podem obter-se a partir da factorização $\Lambda = \Lambda_Q * \Lambda_{\overline{Q}|Q}$. Para as restantes estatísticas multivariadas não existem factorizações equivalentes que permitam realizar todos os testes de informação adicional de uma forma eficiente.

[11] Formalmente Λ_Q define-se como $|\mathbf{W}_{QQ}| / |\mathbf{T}_{QQ}|$, ou seja Λ_Q é a estatística de Wilks para testar a existência de diferenças entre-grupos nas médias de *Q*.

3.2 EXEMPLO

Para o exemplo descrito na secção 2.2, a estatística de Wilks relativa à hipótese nula de igualdade dos valores esperados de todos os indicadores para os três grupos de bancos assume o valor $\Lambda^* = 0.02577$. Os parâmetros da distribuição de Λ são neste caso, $d = p = 17$, $m_H = k\text{-}1 = 2$ e $m_E = N\text{-}k = 30$. Recordando que, para uma estatística $\Lambda(d,m_H,,m_E)$, quando $\min(d,m_H) = 2$, a transformação $F^* = [(1 - \Lambda^{1/2}) / \Lambda^{1/2}] * [(m_E - d+1)/(|m_H - d|+2)]$ segue sob a hipótese nula uma distribuição F de Snedecor com $2(|m_H - d|+2)$ e $2(m_E - d+1)$ graus de liberdade (Seber 1984, 43), o ponto crítico da distribuição de Λ para um teste ao nível de significância, $\alpha = 0.10$, é igual a 0.1145. Neste caso, há evidência clara (p-value < 0.001) para se concluir pela existência de diferenças reais entre os grupos. A metodologia STP a um nível global de significância de 10%, leva a incluir no conjunto $\boldsymbol{A}_{0.10}$ todos os subconjuntos, Q , para os quais $\Lambda_{\overline{Q}|Q}$ seja superior a 0.1145, ou de forma equivalente Λ_Q seja inferior a 0.2251. Esta estratégia corresponde a incluir subconjuntos em $\mathbf{A}_{0.10}$, em função de testes de informação adicional conduzidos aos níveis individuais de significância de 0.074 (q = 1), 0.058 (q = 2), 0.040 (q = 3), 0.028 (q = 4) e a níveis inferiores a 0.02 para $q \geq 5$. De acordo com este critério $\boldsymbol{A}_{0.10}$ não deverá incluir nenhum conjunto formado por menos de 4 variáveis. Incluem-se em $\boldsymbol{A}_{0.10}$ dois subconjuntos formados por 4 variáveis, a saber *Q5* = {CCG, SB, MF, ln RCPD} e *Q6* = {CCG, SB, MF, RA}. De entre os conjuntos com um número de variáveis superior a 4 dever-se-ão incluir (entre outros) todos aqueles que contenham *Q5* ou *Q6* como subconjuntos. Uma descrição exaustiva de $\boldsymbol{A}_{0.10}$ seria por um lado extremamente complexa devido ao elevado número de conjuntos envolvidos, e por outro lado de reduzida utilidade dado o objectivo habitual de procurar subconjuntos simples que contenham toda a informação relevante.

As variáveis incluídas em *Q5* e *Q6*, *Capacidade Crediticia Geral*, *Solvabilidade Bruta*, *Margem Financeira* e o logaritmo da *Relevância dos Custos no Produto* (*Q5)* ou a *Rendibilidade do Activo* (*Q6)*, são variáveis que capturam os aspectos fundamentais das duas dimensões de separação. Aparentemente, estas variáveis seriam suficientes para explicar todas as diferenças entre os grupos considerados. Impõem-se no entanto alguns comentários. Em primeiro lugar, convém notar que nos testes de informação adicional, o erro de 1ª espécie, consiste em considerar como inadequado um conjunto que é adequado. Ou seja, a probabilidade que é controlada, é a probabilidade de não se identificarem alguns conjuntos inadequados. No nosso exemplo, podemos nomeadamente afirmar (com a probabilidade de erro controlada a 10%) que nenhum conjunto formado por menos de 4 indicadores é por si só suficiente para explicar todas as diferenças entre os grupos considerados. Já a afirmação de que *Q5* ou *Q6* são subconjuntos adequados, corresponde a hipóteses cuja probabilidade de erro não foi controlada. Esta é, bem entendido, a estratégia habitual, considera-se um conjunto como adequado a não ser que haja evidência em contrário. Porém, no caso da metodologia STP esta estratégia levanta alguns problemas. Por um lado, dado que esta abordagem utiliza testes individuais conservadores, a probabilidade de se cometerem erros de 2ª espécie pode, por essa razão, ser demasiado elevada. Neste exemplo, tal não parece ser um problema grave, dado que os níveis de significância individuais para subconjuntos com menos de 5 variáveis estão próximos dos 5% habituais. Por outro lado, em testes ligados a métodos de selecção de variáveis, as consequências de se cometerem erros de 2ª espécie podem ser mais graves do que noutras situações, uma vez que eles levam a que não se incluam na análise variáveis importantes. Em particular, quando se utilizam amostras de dimensão reduzida, não existindo evidência

suficiente para se concluir que um determinando subconjunto é inadequado, pode não ser claro se tal é devido de facto à adequação do respectivo conjunto. Em situações deste género, é conveniente tentar determinar de uma forma exploratória se a inclusão de variáveis adicionais é capaz de enriquecer a análise.

3.3 OUTRAS METODOLOGIAS DE TESTES SIMULTÂNEOS

Os STP de Gabriel e Mckay não constituem a única estratégia para fazer inferências sobre $\boldsymbol{A}$ controlando o nível de significância global de todos os testes realizados. Uma abordagem alternativa, sugerida por Spjøtvoll (1978) para análise de regressão mas directamente aplicável neste contexto, consiste em numa primeira fase realizar para todos os subconjuntos testes de informação adicional ao nível α ignorando que se estão a realizar vários testes em simultâneo. Em seguida, incluem-se em $\boldsymbol{A}_\alpha$ todos os conjuntos que foram considerados adequados por estes testes e ainda todos os conjuntos que incluam membros de $\boldsymbol{A}_\alpha$ como subconjuntos, ainda que a hipótese de adequação tenha sido inicialmente rejeitada. Desta forma, a propriedade da coerência, que era verificada automaticamente nos STP é aqui assegurada "à força". Pode-se provar (Spjøtvoll 1978) que os $\boldsymbol{A}_\alpha$ assim construídos incluem todos os conjuntos adequados com uma probabilidade não inferior a $1-\alpha$. Este procedimento tem ainda a vantagem de ser menos conservador que os STP de Gabriel e Mckay. As suas principais desvantagens são as seguintes : (i) A inclusão em $\boldsymbol{A}_\alpha$ para cada subconjunto, Q, não depedende apenas de Q mas também de todos os subconjuntos de Q. (ii) Ao contrário do que acontece nos STP, os níveis de significância individuais de cada teste não são conhecidos.

Relativamente ao exemplo que temos vindo a utilizar, a estratégia de Spjøtvoll baseada na estatística de Wilks levaria a que se incluíssem em $\boldsymbol{A}_\alpha$ todos os conjuntos para os quais $\Lambda_{\overline{Q}|Q}$ seja superior aos pontos críticos de uma distribuição de Λ com parâmetros $d = p–q = 17– q$, $m_H = k –1 = 2$ e $m_E = N\text{-}k–q = 30–q$, Para $\alpha = 0.10$ e $q \leq 4$ os pontos críticos são 0.123 ($q = 1$), 0.133 ($q = 2$), 0.145 ($q = 3$), 0.157 ($q = 4$), 0.170 ($q = 5$), 0.187 ($q = 6$) e 0.205 ($q = 7$). Os menores conjuntos a incluir em $\boldsymbol{A}_{0.10}$ são os seguintes conjuntos de 7 variáveis: *Q7* = {ln TRCC, SB, TMA, TMR, MF, ln RCPD, ALE}, {ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD}, *Q8* = {ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD} e *Q9* = {CCG, ln TRCC, SB, TMA, TMR, MF, ln RCPD}. Verificamos então, que a utilização de testes menos conservadores permite neste caso reduzir consideravelmente o número de conjuntos que se aceitam como adequados.

4. DIMENSÕES DE SEPARAÇÃO E ÍNDICES DE COMPARAÇÃO ENTRE SUBCONJUNTOS

Um segundo problema dos métodos tradicionais de selecção de variáveis é o facto de estes métodos não partirem de um conceito claro e universal de "separação entre grupos" conducente a uma única medida objectiva deste conceito. Com efeito, iremos aqui argumentar que em ADD o conceito da "separação" providenciada por um determinado conjunto de variáveis deverá ser entendido tendo em atenção a capacidade desse conjunto para satisfazer os objectivos da análise, ou seja a sua capacidade para explicar cabalmente todas as diferenças entre os grupos que se considerem relevantes. Tendo em atenção esses objectivos, para problemas diferentes os índices de separação mais adequados poderão também ser diferentes. Nesta secção iremos discutir o problema de escolher um índice adequado para comparar subconjuntos de variáveis em ADD. Ao longo da discussão adoptaremos uma perspectiva geométrica evitando recorrer a pressupostos baseados em modelos probabilisticos.

4.1 ADD NUMA PERSPECTIVA GEOMÉTRICA

Suponha-se que se pretende avaliar a qualidade de um determinado subconjunto, *Q*, formado por *q* variáveis tendo em vista a explicação das diferenças entre os grupos. Sejam $\mathbf{X}_Q$ a matriz ($N*q$) das observações em *Q* ,**G** uma matriz ($N*k$) de indicadores do grupo de origem e $\mathbf{1_N} = [1 \ldots 1]^T$ um vector coluna (com N linhas) constituído unicamente por 1's. Sejam ainda, Ω o subspaço de $\Re^N$ gerado pelas colunas de **G**, ω o subspaço de Ω gerado por $\mathbf{1_N}$, $\Omega^\perp$ e $\omega^\perp$os complementos ortogonais de Ω e ω em $\Re^N$, $\omega^p = \omega^\perp \cap \Omega$ o complemento ortogonal de ω em Ω , e γ o subspaço de $\omega^\perp$ gerado pelas colunas centradas (em relação às médias globais) de $\mathbf{X}_Q$. Então, $\mathbf{X}_Q$ tem uma representação única em termos das suas projecções ortogonais em Ω ($\mathbf{P}_\Omega \mathbf{X}_Q$) e $\Omega^\perp$ ($\mathbf{P}_{\Omega^\perp} \mathbf{X}_Q$) enquanto $\mathbf{P}_\Omega \mathbf{X}_Q$ tem uma representação única nas suas projecções ortogonais em ω ($\mathbf{P}_\omega \mathbf{P}_\Omega \mathbf{X}_Q = \mathbf{P}_\omega \mathbf{X}_Q$) e ω^p ($\mathbf{P}_{\omega^p} \mathbf{P}_\Omega \mathbf{X}_Q = \mathbf{P}_{\omega^p} \mathbf{X}_Q$). Essas representações permitem escrever $\mathbf{X}_Q = \mathbf{P}_\Omega \mathbf{X}_Q + \mathbf{P}_{\Omega^\perp} \mathbf{X}_Q = \mathbf{P}_\omega \mathbf{X}_Q + \mathbf{P}_{\omega^p} \mathbf{X}_Q + \mathbf{P}_{\Omega^\perp} \mathbf{X}_Q$, em que $\mathbf{P}_\Omega \mathbf{X}_Q$ é a matriz das médias por grupo, $\mathbf{P}_{\Omega^\perp} \mathbf{X}_Q$ é a matriz dos desvios em relação às médias por grupo, $\mathbf{P}_\omega \mathbf{X}_Q$ é a matriz das médias globais e $\mathbf{P}_{\omega^p} \mathbf{X}_Q$ é a matriz dos desvios das médias por grupo em relação às médias globais.

Quanto maior for a separação entre grupos maior será a importância da parcela $\mathbf{P}_{\omega^p} \mathbf{X}_Q$ nesta representação. Com efeito, as técnicas clássicas de ADD podem ser apresentadas no contexto de uma Análise de Correlação Canónica entre γ e ω^p. Nomeadamente, nessa análise o vector de $\Re^N$ definido pelos scores (centrados) dos indivíduos na primeira FDL, define a primeira direcção canónica em γ, e o primeiro valor próprio de $\mathbf{B}\,\mathbf{T}^{-1}$ (l_1) iguala o cosseno quadrado do ângulo entre a primeira FDL e ω^p (Masson 1990). Nesse sentido l_1 pode ser entendido como uma medida da separação ao longo da dimensão definida pela primeira FDL. De igual modo, a i-ésima FDL (FDL_i) define a i-ésima direcção canónica em γ, enquanto o i-ésimo valor próprio de $\mathbf{B}\,\mathbf{T}^{-1}$ (l_i) iguala o cosseno quadrado entre

FDL_i e o vector que define a i-ésima direcção canónica em ω^{p12}. Desta forma, é possível definir índices de separação entre os grupos a partir dos valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$. Estes índices medem a proximidade em $\Re^N$ entre as posições relativas de ω^p e do espaço gerado pelos elementos (centrados) de $\mathbf{X}_Q$, havendo uma correspondência entre a importância dada a l_i e o ênfase atribuído à dimensão de separação definida pela i-ésima FDL.

4.2 INDICES DE COMPARAÇÃO ENTRE SUBCONJUNTOS

Para situar neste contexto a estratégia habitual de avaliar a separação entre grupos através do valor da estatística de Wilks (Λ), convém notar a equivalência entre a minimização de Λ e a maximização do índice de separação que lhe está associado

$$\tau^2 = 1 - \Lambda^{1/r} = 1 - \left(\prod_{i=1}^{r} (1 - l_i) \right)^{1/r}$$

e que é simplesmente o complementar para a unidade da média geométrica dos senos quadrados dos ângulos canónicos entre γ e ω^p. Ou seja, τ^2 é um índice que por um lado considera todas as dimensões de separação possíveis, mas por outro lado tende a dar um maior ênfase às primeiras dimensões, uma vez que uma média geométrica de valores compreendidos entre 0 e 1, tende a dar maior importância (pelo menos em comparação com uma média aritmética) aos valores mais próximos de zero, e as FDLs estão ordenadas pela da sua "proximidade" (medida aqui pelo ângulo canónico respectivo) com ω^p. Põe-se então a questão de saber se τ^2 é o índice mais adequado para todos os problemas de ADD. Vai-se aqui argumentar que sendo em geral, τ^2 um índice "razoável", para alguns problemas poderão haver outros índices mais apropriados. Nomeadamente, se todas as diferenças relevantes poderem ser explicadas ao longo de uma única dimensão de separação, um índice mais apropriado será l_1, o cosseno quadrado do ângulo entre FDL_1 e ω^p, uma vez que nesse caso os restantes valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$ resultam ou de ruído resultante da variação amostral, ou de dimensões de separação consideradas como pouco interessantes pelo analista. Por outro lado, se todas as dimensões de separação possíveis representarem diferenças reais e forem consideradas como igualmente importantes, um índice mais adequado será

$$\xi^2 = \frac{\mathrm{tr}(\mathbf{B}\mathbf{T}^{-1})}{r} = \sum_{i=1}^{r} \frac{l_i}{r},$$

[12] Esta identidade resulta da igualdade entre os quadrados dos coeficientes de correlação canónica e os valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$ e da interpretação geométrica de um coeficiente de correlação.

que iguala a média aritmética dos cossenos quadrados de todos os ângulos canónicos. O leitor mais atento, certamente reconhecerá que enquanto τ^2 é o índice multivariado associado com a estatística de Wilks, l_1 e ξ^2 são os índices associados respectivamente com as estatísticas de Roy e de Bartlett-Pillai. Torna-se então natural indagar como se situa neste contexto o índice

$$\varsigma^2 = \frac{\mathrm{tr}(\mathbf{BW}^{-1})}{\mathrm{r} + \mathrm{tr}(\mathbf{BW}^{-1})} = 1 - \frac{\mathrm{r}}{\sum_{\mathrm{i}=1}^{\mathrm{r}} \frac{1}{1-\mathrm{l}_\mathrm{i}}}$$

associado com a estatística de Lawley-Hotelling. Notando que a relação entre ς^2 e os valores próprios de $\mathbf{B\ T^{-1}}$ permite exprimir ς^2 como o complementar para a unidade da média harmónica dos senos quadrados de todos os ângulos canónicos, verificamos que ς^2 é um índice semelhante a τ^2, mas que ainda põe maior ênfase nas dimensões definidas pelas primeiras FDLs.

4.3 NÚMERO E IMPORTÂNCIA DAS DIMENSÕES DE SEPARAÇÃO

Vemos então que a questão da escolha de um índice apropriado da contribuição de $\mathbf{X}_Q$ para a separação entre os grupos está intimamente ligada ao ênfase que se pretende dar a cada dimensão de separação. No caso de só existirem dois grupos, só é possível definir uma dimensão de separação e todos os índices dão resultados equivalentes[13]. Porém no caso de existirem mais do que dois grupos o problema pode não ter uma resposta evidente. Por um lado, é importante distinguir as FDLs que representam dimensões reais de separação, das FDLs que estão apenas associadas a variabilidade amostral. Por outro lado, é necessário determinar a importância que cada dimensão real tem para o analista.

Quando se assumem os pressupostos clássicos de normalidade multivariada e igualdade das variâncias e covariâncias intra-grupos a primeira destas questões pode ser respondida com a ajuda de testes de hipóteses conhecidos, os chamados testes de dimensionalidade. Estes testes formalizam a hipótese de que só *t* FDLs correspondem a "dimensões reais" de separação da seguinte forma. Sejam

$$\boldsymbol{\mu}_\mathbf{g} = \mathbf{E}(\mathbf{X}_\mathbf{g}) \qquad \bar{\boldsymbol{\mu}} = \sum_{\mathbf{g}=1}^{\mathbf{k}} \frac{\mathrm{n_g}\,\boldsymbol{\mu}_\mathbf{g}}{\mathrm{N}}$$

[13] Que são ainda equivalentes aos resultados que se obtêm se tomar como índice de separação a distância de Mahalanobis entre os centroides de cada grupo, $\mathrm{D} = \left[\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathbf{T}} \mathbf{S}^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)\right]^{1/2}$ em que **S** = **W**/(*N*-2) é a matriz amostral de covariâncias intra-grupos.

Esta equivalência resulta das igualdades $\mathrm{D}^2 = \frac{\mathrm{N}(\mathrm{N}-2)}{\mathrm{n}_1\mathrm{n}_2}\ddot{\mathrm{e}}_1 = \frac{\mathrm{N}(\mathrm{N}-2)}{\mathrm{n}_1\mathrm{n}_2}\frac{\mathrm{l}_1}{1-\mathrm{l}_1}$ que definem relações monótonas entre D e os valores próprios positivos de $\mathbf{B\ W^{-1}}$ e $\mathbf{B\ T^{-1}}$.

as médias populacionais, por grupo e globais dos vectores aleatórios $\mathbf{X_g}$. Sejam ainda

$$\mathbf{\Sigma} = E\left[\left(X_g - \boldsymbol{\mu}_g\right)\left(X_g - \boldsymbol{\mu}_g\right)^T\right] \quad \mathbf{\Theta} = \sum_{g=1}^{k} n_g \left(\boldsymbol{\mu}_g - \bar{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}_g - \bar{\boldsymbol{\mu}}\right)^T$$

as matrizes de desvios populacionais "intra" e "entre" grupos. Então o problema de determinar o número de "dimensões reais" de separação é equivalente à identificação da característica de matriz de não-centralidade $\mathbf{\Psi} = \mathbf{\Theta}\,\mathbf{\Sigma}^{-1}$ (ver, por exemplo, Krishnaiah, 1982)[14]. Com efeito, se $\mathbf{\Psi}$ tiver característica t ($t \leq r$), então as t primeiras correlações canónicas populacionais são positivas enquanto as r-t últimas são iguais a 0. Esse resultado é geralmente interpretado como significando que existem exactamente t "dimensões reais" de separação. A hipótese nula de que a característica de $\mathbf{\Psi}$ é igual ou inferior a t pode ser testada com base em várias funções dos últimos r-t valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ (λ_{t+1}, ..., λ_r). Por exemplo, uma estatística comum, devida a Bartlett (1947), é

$$T = \left(N - 1 - \frac{q+k}{2}\right)\sum_{i=t+1}^{r} ln(1+\boldsymbol{\lambda}_i)$$

Sob a hipótese nula, T segue assintóticamente uma distribuição do Qui-quadrado com $(q - t)*(k - 1 - t)$ graus de liberdade (ver Krishnaiah, 1982, para uma descrição de outras funções de λ_{t+1}, ..., λ_r que podem ser usadas como estatísticas em testes de dimensionalidade).

A realização de testes de dimensionalidade pode, no entanto, não ser suficiente para determinar quais (nem a importância) as dimensões a considerar. Em primeiro lugar, os pressupostos clássicos dos testes de dimensionalidade raramente estão satisfeitos na prática, pelo que as suas conclusões são frequentemente questionáveis. Apesar destes testes serem relativamente robustos e por conseguinte defensáveis desde que não existam desvios substanciais aos seus pressupostos, há problemas a que eles não conseguem responder de uma forma completamente satisfatória. Nomeadamente, tal como acontece com os testes de informação adicional, os testes de dimensionalidade só são válidos quando a análise é baseada num conjunto de variáveis escolhido à partida havendo enviesamentos de efeitos mal conhecidos quando se efectua uma selecção prévia de variáveis. Finalmente, pode acontecer que algumas dimensões de separação, ainda que estatisticamente significativas, sejam consideradas como "pouco interessantes" e sejam por essa razão ignoradas pelo analista, ou que não haja evidência suficiente para concluir pela existência real de algumas dimensões que o analista considera suficientemente importantes para considerar. Situações do primeiro tipo poderão acontecer quando as ultimas FDLs, embora representando

[14] Estamos a supor que $\mathbf{\Sigma}$ é uma matriz não singular. Note-se ainda que a característica de $\mathbf{\Psi}$ é idêntica à característica de $\mathbf{\Theta}$. No entanto, dado que por um lado a separação entre os grupos pode ser descrita a partir dos valores e vectores próprios de $\mathbf{\Psi}$ e por outro lado os testes de dimensionalidade são baseados em valores próprios de $\mathbf{B}\,\mathbf{W}^{-1}$ que pode ser considerado como uma estimativa de $\mathbf{\Psi}$ (a menos de uma constante de proprocionalidade), é habitual caracterizar o número de dimensões de separação a partir de $\mathbf{\Psi}$.

dimensões reais, tenham uma contribuição negligenciável para a separação dos grupos, ou quando elas estejam associadas a conceitos que não sejam relevantes para o estudo em questão. Situações do segundo tipo acontecem tipicamente, quando o reduzido numero de observações disponíveis não permita estabelecer a existência de dimensões associadas a FDLs cuja interpretação é teoricamente consistente e que caracterizem conceitos importantes para o estudo.

Por conseguinte, a escolha de dimensões a considerar é, no nosso entender, uma escolha eminentemente subjectiva em que se deve ter em atenção a natureza do problema e os objectivos do análise. Em geral, das r dimensões de separação possíveis, as primeiras s ($s \leq t \leq r$) corresponderão a dimensões reais com interesse e as seguintes t-s corresponderão a dimensões reais mas sem interesse, e as ultimas serão apenas o resultado de variação amostral. Um índice de separação ideal, deverá considerar as s dimensões relevantes de forma equilibrada e ignorar as restantes. Havendo dúvidas quanto ao número de dimensões a reter, o que poderá acontecer devido a testes de dimensionalidade inconclusivos, ou à existência de interpretações alternativas para as FDLs, índices que tal como τ^2 consideram todas as dimensões possíveis mas atribuem maior ênfase às primeiras, podem ser considerados como um compromisso razoável. No entanto, por vezes poderá haver interesse em utilizar índices que atribuiam maior ou menor ênfase nas primeiras dimensões de separação. Dois índices conhecidos com essas características são respectivamente ζ^2 e ξ^2 . Frequentemente, mais do que a escolha de um único índice pode ser particularmente útil a comparação das ordenações dos "melhores" subconjuntos, resultantes da escolha de vários índices que dão ênfases diferentes a cada dimensão de separação. Essa comparação pode dar pistas importantes para uma melhor compreensão das diferenças entre os grupos, compreensão essa que constitui o objectivo último da análise.

Finalmente, um aspecto prático a ter em consideração reside na possibilidade de existirem limitações em termos computacionais quanto à capacidade de se identificarem e ordenarem os "melhores" subconjuntos de acordo com cada índice, num tempo razoável. Como vimos atrás, para ordenações baseadas em τ^2 o algoritmo de *Furnival--Furnival e Wilson--McCabe* pode ser utilizado, o que permite ultrapassar este problema para um número "moderado" de variáveis. Como é demonstrado em Duarte Silva (1998), estes algoritmos continuam a ser adaptáveis para ordenações baseadas em ζ^2 ou ξ^2, e no caso de r ser menor ou igual a 3, para ordenações baseadas em qualquer função monótona dos valores próprios de $\mathbf{B}\,\mathbf{T}^{-1}$.

4.4 EXEMPLO

Para o exemplo relativo às diferenças entre as instituições bancárias a operar em Portugal em 1993 os quadros 9 a) -- 9 d) mostram os dois melhores subconjuntos de cada dimensão de acordo com os índices τ^2 , ζ^2 , ξ^2 e l_1. A evolução destes índices para os melhores subconjuntos está representada nas figuras 1 e 2. É particularmente interessante comparar a evolução e os melhores subconjuntos seleccionados pelos índices l_1 e ξ^2. Nomeadamente, l_1 tem um crescimento claro desde $q = 1$ até $q = 4$, continuando a crescer de forma mais moderada de $q = 5$ até $q = 11$ e estabilizando para $q > 11$. Recordando que l_1 é um índice que só considera a primeira dimensão de separação, a sua evolução sugere que é possível capturar os aspectos principais dessa dimensão com 4 variáveis, sendo necessárias 11 variáveis para a descrever de uma forma completa.

Figura 1
EVOLUÇÃO DOS INDICES DE SEPARAÇÃO
(melhores subconjuntos)

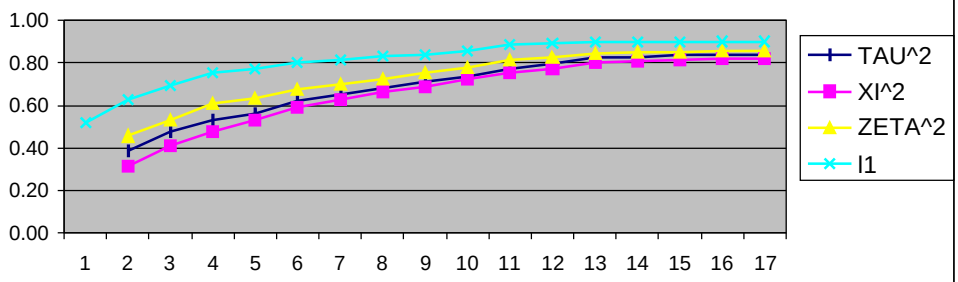
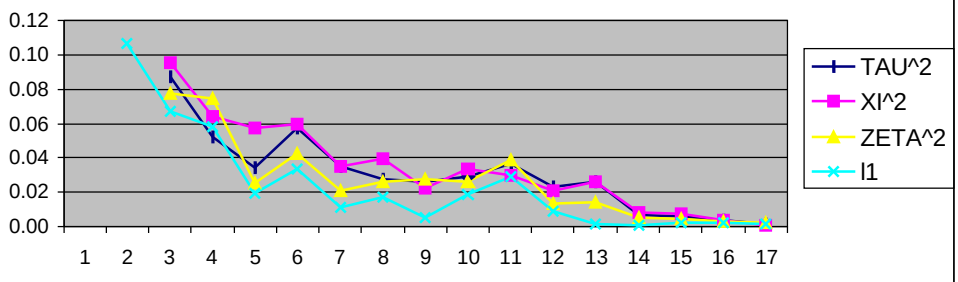


Figura 2
VARIAÇÃO DOS INDICES DE SEPARAÇÃO
(melhores subconjuntos)



A escolha das variáveis incluídas nos melhores subconjuntos (de acordo com l_1) para $q = 4$, *Q10* = {LR, ln TRCC, GE, MN}, e $q = 11$, *Q11* = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE}, é consistente com essa hipótese. A evolução do índice ξ^2 é no entanto, diferente (ver figuras 1 e 2). Este índice mantém um claro crescimento até $q = 13$ só começando a estabilizar a partir deste ponto. Recordando que ξ^2 é o índice que pondera as duas dimensões de separação de forma mais equilibrada, estes resultados sugerem que são necessárias 13 variáveis para as representar a ambas de forma completa. É de realçar que o melhor subconjunto com 13 variáveis de acordo com ξ^2 é o conjunto *Q3* que já havia sido proposto por um método descendente de selecção passo a passo, e pela comparação entre todos os subconjuntos baseada em Λ (ou τ^2). Este conjunto incluiu todas as variáveis importantes para a interpretação das duas dimensões de separação, nomeadamente todos os elementos de *Q11* e a *Rendibilidade Bruta dos Capitais Próprios*, que não fazendo parte de *Q11* é particularmente útil para a interpretação de FDL_2. Curiosamente o melhor conjunto de 13 variáveis de acordo com ζ^2 também é *Q3*, o que já não acontece quando se utiliza l_1 para índice de comparação. No entanto, a evolução de ζ^2 já não sugere a escolha de *Q3* de forma tão clara, podendo igualmente optar-se por um conjunto com 11 variáveis (ver figuras 1 e 2), aparecendo neste caso *Q10* em primeiro lugar. Como tínhamos visto na secção 4.2 de entre os índices τ^2, ξ^2 e

ζ^2 este último é aquele que está mais próximo de l_1, não sendo portanto surpreendente que l_1 e ζ^2 sugiram a escolha de subconjuntos semelhantes.

QUADRO 9
DOIS MELHORES SUBCONJUNTOS DE CADA DIMENSÃO

a) SEGUNDO O CRITÉRIO DE WILKS

q	Q	Λ	TAU^2
	ln TRCC	0.4793	0.521
1	RCPE	0.5234	0.477
	CCG, MF	0.3726	0.390
2	SB, MF	0.3819	0.382
	CCG, SB, MF	0.2740	0.477
3	CCG, MF, RBCP	0.2957	0.456
	CCG, SB, MF, ln RCPD	0.2223	0.529
4	CCG, SB, MF, RA	0.2238	0.527
	LR, ln TRCC, GE, MN, RBCP	0.1909	0.563
5	CCG, SB, MF, ln RCPD, ALE	0.1919	0.562
	ln TRCC, SB, TMA, TMR, MF, ln RCPD	0.1442	0.620
6	LR, ln TRCC, GE, MF, MN, RBCP	0.1522	0.610
	ln TRCC, SB, TMA, TMR, MF, ln RCPD, ALE	0.1191	0.655
7	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD	0.1199	0.654
	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE	0.1008	0.683
8	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, RBA	0.1016	0.681
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD	0.0847	0.709
9	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE, RBCP	0.0863	0.706
	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, RBCP	0.0689	0.738
10	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	0.0695	0.736
	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	0.0512	0.774
11	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, RBCP	0.0538	0.768
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	0.0412	0.797
12	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBA, RBCP	0.0446	0.789
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	0.0314	0.823
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBA, RBCP	0.0379	0.805
14	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	0.0292	0.829
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	0.0294	0.828
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	0.0273	0.835
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RCP	0.0278	0.833
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	0.0262	0.838
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA, RCP	0.0268	0.836

b) SEGUNDO O CRITÉRIO DE BARTLETT-PILLAI

q	Q	tr (BT^{-1})	XI^2
	ln TRCC	0.5207	0.521
1	RCPE	0.4766	0.477
	SB, MN	0.6328	0.316
2	SB, MF	0.6327	0.316
	CCG, SB, MF	0.8231	0.412
3	CCG, MF, RBCP	0.7894	0.395
	CCG, GE, MN, RBCP	0.9513	0.476
4	CCG, MN, ALE, RBCP	0.9344	0.467
	CCG, GE, MF, MN, RBCP	1.0660	0.533
5	SB, TMA, TMR, MF, RA	1.0450	0.523
	ln TRCC, SB, TMA, TMR, MF, ln RCPD	1.1858	0.593
6	CCG, ln TRCC, MF, MN, ALE, RBCP	1.1399	0.570
	ln TRCC, SB, TMA, TMR, MF, ln RCPD, ALE	1.2559	0.628
7	SB, TMA, TMR, MF, RCPE, ALE, RA	1.2390	0.620
	ln TRCC, SB, TMA, TMR, MF, RCPE, ALE, RA	1.3348	0.667
8	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE	1.3089	0.654
	ln TRCC, SB, TMA, TMR, MF, RCPE, ln RCPD, ALE, RA	1.3791	0.690
9	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, ALE, RBCP	1.3735	0.687
	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	1.4463	0.723
10	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ALE, RBCP, RA	1.4163	0.708
	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	1.5056	0.753
11	ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ALE, RBA, RBCP, RA	1.4817	0.741
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBCP	1.5469	0.773
12	LR, ln TRCC, SB, TMA, TMR, MF, MN, RCPE, ln RCPD, ALE, RBA, RBCP	1.5410	0.770
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	1.5996	0.800
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln EB, ALE, RBCP, RA	1.5725	0.786
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA	1.6149	0.807
14	CCG, LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	1.6145	0.807
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	1.6290	0.815
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RCP	1.6270	0.814
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	1.6370	0.819
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RA	1.6334	0.817

c) SEGUNDO O CRITÉRIO DE LAWLEY-HOTELLING

q	Q	tr (BW^{-1})	$ZETA^2$
	ln TRCC	1.0863	0.521
1	RCPE	0.9104	0.477
	CCG, MF	1.6838	0.457
2	SB, MF	1.5799	0.441
	CCG, SB, MF	2.2969	0.535
3	CCG, MF, RCP	2.2736	0.532
	LR, ln TRCC, GE, MN	3.1176	0.609
4	CCG, SB, MF, ln RCPD	3.0111	0.601
	CCG, SB, MF, ln RCPD, ALE	3.4697	0.634
5	CCG, ln TRCC, MN, ln RCPD, RBA	3.4283	0.632
	LR, ln TRCC, GE, ln RCPD, ln EB, ALE	4.1914	0.677
6	LR, CCG, ln TRCC, MN, ln RCPD, RBA	4.1833	0.677
	LR, CCG, ln TRCC, GE, MN, ln RCPD, RBA	4.6158	0.698
7	LR, ln TRCC, GE, ln RCPD, ln EB, ALE, RBCP	4.6026	0.697
	LR, ln TRCC, SB, GE, MN, ln RCPD, ln EB, ALE	5.2277	0.723
8	ln TRCC, SB, TMA, TMR, MF, MN, ln RCPD, RBA	5.1843	0.722
	LR, ln TRCC, GE, TMA, TMR, MF, MN, ln RCPD, RBCP	6.0374	0.751
9	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD	6.0081	0.750
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, RBA	6.9734	0.777
10	LR, ln TRCC, SB, GE, TMA, TMR, MF, ln RCPD, ln EB, ALE	6.9483	0.776
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE	8.8418	0.816
11	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, RBCP	8.2584	0.805
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBA	9.6887	0.829
12	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBCP	9.5196	0.826
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	10.7540	0.843
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA	10.1110	0.835
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP	11.1857	0.848
14	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	11.1685	0.848
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	11.5756	0.853
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RCP	11.4097	0.851
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	11.8339	0.855
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA, RCP	11.7670	0.855

d) SEGUNDO O CRITÉRIO DE ROY

q	Q	l_1	l_1
	ln TRCC	1.0863	0.521
1	RCPE	0.9104	0.477
	CCG, MF	1.6840	0.627
2	SB, MF	1.5552	0.609
	CCG, MF, RCP	2.2705	0.694
3	CCG, ln TRCC, MN	2.2027	0.688
	LR, ln TRCC, GE, MN	3.0324	0.752
4	CCG, SB, MF, ln RCPD	2.8394	0.740
	CCG, ln TRCC, MN, ln RCPD, RBA	3.3722	0.771
5	LR, ln TRCC, GE, MN, ln EB	3.2470	0.765
	LR, CCG, ln TRCC, MN, ln RCPD, RBA	4.1267	0.805
6	LR, ln TRCC, GE, ln RCPD, ln EB, ALE	4.0928	0.804
	LR, ln TRCC, GE, ln RCPD ln EB, ALE, RBCP	4.4245	0.816
7	LR, CCG, ln TRCC, MN, ln RCPD, ALE, RBA	4.3727	0.814
	LR, ln TRCC, SB, GE, MN, ln RCPD, ln EB, ALE	4.9925	0.833
8	LR, CCG, ln TRCC, GE, MN, ln RCPD, ln EB, ALE	4.8121	0.828
	LR, ln TRCC, GE, TMA, TMR, MF, MN, ln RCPD, RBCP	5.1958	0.839
9	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD	5.0599	0.835
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, RBA	5.9790	0.857
10	LR, ln TRCC, SB, GE, TMA, TMR, MF, ln RCPD, ln EB, ALE	5.9701	0.857
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE	7.7497	0.886
11	LR, ln TRCC, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBCP	7.0717	0.876
	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBA	8.5044	0.895
12	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBCP	8.0755	0.890
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE, RBA	8.6428	0.896
13	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA	8.5632	0.895
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA	8.7152	0.897
14	LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RCP	8.6723	0.897
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RCP	8.8891	0.899
15	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP	8.8670	0.899
	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP, RA, RCP	9.0560	0.901
16	LR, CCG, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBA, RBCP, RCP	9.0368	0.900

Estes resultados ilustram a forma como a escolha do índice utilizado para comparar subconjuntos, influencia a importância dada às diferentes dimensões de separação. Antes de se escolher um índice particular, é conveniente tentar determinar quantas dimensões são de facto necessárias para separar os grupos. Tal poderá ser feito com a ajuda dos testes de dimensionalidade discutidos na secção 4.3. Neste caso, um teste de dimensionalidade (utilizando todas as variáveis) relativo à hipótese nula de que a matriz de não-centralidade, $\boldsymbol{\Psi}$, tem característica 1, resulta num valor para a estatística T de 29.362 (p-value = 0.022). Por conseguinte, a um nível de significância de 5%, pode concluir-se pela existência de duas dimensões reais de separação. Havendo interesse em distinguir os bancos relativamente a estas duas dimensões que foram consideradas como igualmente importantes, privilegiaram-se neste caso as comparações baseadas em ξ^2. A comparação entre as ordenações de subconjuntos motivadas por ξ^2 e l_1 ajuda por sua vez a isolar as variáveis mais fortemente associadas a cada dimensão de separação. No caso de análises centradas na interpretações das diferenças associadas unicamente à estrutura de exploração, o índice privilegiado seria l_1.

5. CONCLUSÕES

A descrição de diferenças entre grupos a partir de um grande número de variáveis é um problema complexo que requer uma combinação inteligente de técnicas de ADD, com metodologias de selecção e comparação de subconjuntos de variáveis e conhecimentos substantivos sobre o problema a analisar. Este problema é tradicionalmente abordado ou através de métodos informais que partem da análise do conjunto completo de variáveis, ignorando posteriormente aquelas variáveis que se revelarem menos interessantes, ou através de selecções prévias baseadas em métodos de selecção passo a passo. A análise informal do conjunto completo de variáveis é geralmente um bom ponto de partida para compreender a natureza das diferenças mais importantes. No entanto, esta análise é eminentemente subjectiva e pode ser influenciada por um grande número de variáveis irrelevantes ou redundantes que apenas acrescentam ruído. Métodos automáticos de selecção passo a passo podem contribuir para ultrapassar estes dois problemas. Porém, estes métodos usam algoritmos heurísticos que não garantem a identificação dos subconjuntos mais apropriados. Por outro lado, embora alguns do métodos de selecção passo a passo mais importantes se baseiem em testes de hipóteses formais, esses testes são aí utilizados de uma forma que não permite a realização de inferências estatisticamente válidas. Metodologias de testes simultâneos, permitem realizar inferências sobre o conjunto *A* formado por todos os subconjuntos "adequados", os seja, os subconjuntos que incluem toda a informação relevante para explicar as diferenças observadas. Estas metodologias são particularmente úteis para identificar subconjuntos inadequados. Para o problema de identificar os subconjuntos adequados, estes métodos não conseguem dar uma resposta completamente satisfatória, uma vez que apenas controlam a probabilidade de se considerar como inadequados conjuntos adequados (erro de 1ª espécie). A escolha entre vários candidatos a conjuntos adequados pode fazer-se de uma forma exploratória comparando o valor de vários índices de separação. Por sua vez, a comparação de ordenações de subconjuntos sugeridas por índices que dão ênfases diferenciados às diferentes dimensões de separação, pode igualmente ajudar a compreender melhor a natureza das diferenças observadas.

Todas as técnicas descritas neste artigo podem ser facilmente generalizáveis a um contexto mais geral. Suponha-se que se dispõe de um conjunto de *N* observações em *p* variáveis, X_1, X_2, ..., X_p, cujos valores esperados podem ser representados num espaço vectorial $\Omega \subseteq \Re^N$ de dimensão *u*. Seja ω um subspaço de Ω com dimensão *s*, ω^p com dimensão $t = u - s$, o complemento ortogonal de ω em Ω e γ o espaço de dimensão $r = \min(t, p)$ gerado pelas projecções de X_1, X_2, ..., X_p no complemento ortogonal de ω em $\Re^N$. Então, é bem sabido que qualquer vector $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]$ tem representações únicas, $x_i = a_i + b_i = a_i + c_i + d_i$ com $a_i \in \omega$, $b_i \in \gamma$, $c_i \in \omega^p$ e $d_i \in \Omega^{\perp}$, em que a_i, b_i, c_i e d_i são as projecções ortogonais de x_i em ω, γ, ω^p e $\Omega^{\perp}$. Nesta formulação, Ω pode ser definido a partir de um modelo linear apropriado,

$E(X) \in \omega$ define uma hipótese de referência supostamente falsa pretendendo-se descrever e interpretar desvios em relação a essa hipótese. Tal poderá ser feito a partir da análise dos vectores próprios de HE^{-1} ou HT^{-1} (que são idênticos), em que H, E e T = H + E são as matrizes das somas de quadrados e produtos cruzados para os vectores c_i (H), d_i (E), e b_i (T). Estes vectores próprios definem as direcções em γ, resultantes de uma análise de correlação canónica entre γ e ω^p e os valores próprios de HT^{-1} igualam os cosenos quadrados dos respectivos ângulos canónicos.

Esta formulação estende as técnicas de ADD à análise de qualquer "efeito multivariado" definido a partir de modelos e hipóteses lineares. No caso dos problemas clássicos de Análise Discriminante, Ω é o espaço gerado por *k* variáveis binárias que indicam os grupos de origem e a condição $E(X) \in \Omega$, especifica que $E(X_i)$ apenas depende do grupo a que o indivíduo *i* pertence. A hipótese de referência, é a igualdade das médias por grupo, o que equivale a impor que todos os vectores de valores esperados pertençam a um espaço unidimensional, ω. "Desvios" em relação à hipótese de referência são equivalentes a diferenças entre grupos. Notando a ligação entre a ADD clássica e os modelos MANOVA a um factor, alguns autores (por exemplo, Masson 1990) mostram como a formulação geral apresentado no parágrafo anterior, permite estender a aplicação das técnicas de ADD à interpretação de qualquer "efeito" considerado como significativo por uma MANCOVA ou MANOVA a mais de um factor. Para aplicar as técnicas discutidas neste artigo a problemas de selecção de variáveis visando a interpretação de qualquer "efeito" neste contexto geral, basta substituir as matrizes B e W por matrizes H e E apropriadas, mantendo-se a validade de todos os resultados e técnicas apresentadas.

REFERÊNCIAS BIBLIOGRÁFICAS

BARTLETT, M.S. (1947), "Multivariate Analysis", *Journal of the Royal Statistical Society Suppl.*, Vol 9, 176-190.

DUARTE SILVA, A.P. (1998), "Efficient Screening of Variable Subsets in Multivariate Statistical Models", *FCEE – Universidade Católica Portuguesa, C.R. Porto*, WP-98-004.

FURNIVAL, G.M. (1971), "All Possible Regressions with Less Computation" *Technometrics*, Vol. 13, 403-408.

FURNIVAL, G.M. E WILSON, R.W. (1974), "Regressions by Leaps and Bounds", *Technometrics*, Vol 16, 499-511.

GABRIEL, K.R. (1968), "Simultaneous Test Procedures in Multivariate Analysis of Variance", *Biometrika*, Vol 55, 489-504.

GABRIEL, K.R. (1969), "Simultaneous Test Procedures – Some Theory of Multiple Comparisons", *Annals of Mathematical Statistics*, Vol 40, 224-250.

HUBERTY, C.J. (1994), *Applied Discriminant Analysis*, Nova Iorque, John Wiley.

KOBILINSKY, A. (1990), "Analyse Factoriel Discriminante", in *Analyse Discriminante sur Variables Continues, G. Celeux (ed)*, 65-80, INRIA.

KRISHNAIAH, P.R. (1982), "Selection of Variables in Discriminant Analysis", in *Handbook of Statistics, Vol 2, P.R. Krishnaiah e L.N. Kanal (ed)*, 883-892, North-Holland Publishing Company.

MASSON, J.P. (1990), "Discrimination e Analyse de Variance", in *Analyse Discriminante sur Variables Continues, G. Celeux (ed)*, 81-99, INRIA.

MCCABE, G.P. (1975), "Computations for Variable Selection in Discriminant Analysis", *Technometrics*, Vol 17, 103-109.

MCKAY, R.J. (1977), "Simultaneous Procedures for Variable Selection in Multiple Discriminant Analysis", *Biometrika*, Vol 64, 283-290.

MCKAY, R.J. E CAMPBELL, N.A. (1982), "Variable Selection Techniques in Discriminant Analysis I. Description", *British Journal of Mathematical and Statistical Psychology*, Vol 35, 1-29.

RAO, C.R. (1973), *Linear Statistical Inference and its Applications, 2nd Ed.*, Nova Iorque, John Wiley.

SEBER, J.A.F. (1984), *Multivariate Observations*, Nova Iorque, John Wiley.

SPJØTVOLL, E. (1977), "Alternatives to Plotting C_p in Multiple Regression", *Biometrika*, Vol 64, 1-8.