

FERNANDA MARIA CAMPOS DE SOUSA

**NOVAS METODOLOGIAS E VALIDAÇÃO EM
CLASSIFICAÇÃO HIERÁRQUICA ASCENDENTE**

Lisboa
2000

nº de arquivo
“copyright”

FERNANDA MARIA CAMPOS DE SOUSA

**NOVAS METODOLOGIAS E VALIDAÇÃO EM
CLASSIFICAÇÃO HIERÁRQUICA ASCENDENTE**

**Dissertação apresentada para obtenção do
Grau de Doutor em Matemática, Especialidade de Estatística,
pela Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia**

LISBOA

2000

Universidade do Porto	
Faculdade de Engenharia	
Biblioteca 7	
Nº	
CDU	
Data	/ /20

*A presente dissertação foi apoiada financeiramente
pela unidade de I&D "CEDEC - Centro de Estudos de Engenharia Civil",
projecto de "Técnicas Matemáticas em Engenharia", da Universidade do Porto
e parcialmente apoiada
pelo projecto de investigação do "Centro de Matemática e Aplicações",
Linha de "Matemática do Aleatório", da Universidade Nova de Lisboa
e pelo LEAD - Laboratório de Estatística e Análise de Dados
da Faculdade de Psicologia e Ciências da Educação da Universidade de Lisboa.*

*A meus pais
e
às minhas filhas*

Agradecimentos

Ao terminar este trabalho, pretendo agradecer a todas as pessoas e entidades que, de formas bem diferentes, contribuíram para a sua realização. Na impossibilidade de ser exaustiva e correndo o risco de, involuntariamente, omitir algum contributo significativo, desejo referir aqueles apoios que considero mais importantes.

Ao Professor Fernando Nicolau, meu orientador, quero expressar o meu sincero obrigado pelas longas sessões de trabalho, ricas de ideias e sugestões. Quero ainda agradecer-lhe o carinho amigo com que sempre me acolheu.

À Professora Helena Bacelar-Nicolau, agradeço o interesse e apoio dispensados durante a realização deste trabalho.

Ao Professor Francisco Calheiros, responsável por me despertar o gosto pela Estatística, quero manifestar o meu apreço, pela amizade e disponibilidade que sempre me dispensou.

À Faculdade de Engenharia da Universidade do Porto, e em particular, ao Departamento de Engenharia Civil, agradeço as condições dadas, nomeadamente a dispensa de serviço docente que me foi concedida.

Aos meus colegas de Secção em geral, mas de forma muito particular ao Dr. Luís Vieira, agradeço a disponibilidade para troca de impressões e o incentivo amigo manifestado durante a realização deste trabalho. Ao meu colega de cadeira Dr. Rui Gonçalves, desejo agradecer a ajuda sempre pronta, quando solicitada.

À minha amiga Professora Paula Brito, agradeço o apoio concedido, nomeadamente em termos bibliográficos.

À minha família, principalmente aos meus pais, expresso a minha gratidão pelo apoio constante e pela confiança que tornaram possível este trabalho.

Às minhas filhas, Margarida e Inês Sofia, quero agradecer a compreensão da minha ausência e menor disponibilidade dos últimos tempos.

Ao meu marido, José Afonso, agradeço o incentivo e apoio que, muito à sua maneira, me concedeu ao longo deste trabalho.

Resumo

A dissertação aqui apresentada começa por situar os problemas da Classificação e da Classificação Hierárquica no contexto da Estatística Moderna, referindo as fontes matemáticas que lhe dão suporte. Assim a primeira preocupação foi fazer uma síntese das bases matemáticas da Análise Classificatória, e em particular da Análise Classificatória Hierárquica, e um levantamento dos principais problemas de investigação, pura e aplicada, desta área da actualidade. Desta atitude relevou-se o problema da validação e do controlo das estruturas classificatórias.

O resultado de um método de Classificação Hierárquica Ascendente depende de dois aspectos: a eventual existência de uma estrutura nos dados e do(s) método(s) utilizado(s), sendo de admitir que cada método tem um modelo de estrutura subjacente, que lhe é próprio e que ele é capaz de otimizar em cada situação. A aplicação de um critério a um conjunto de dados com estrutura mal definida, ou mesmo ausente, tem como resultado uma estrutura que está associada ao modelo subjacente ao método.

A necessidade de produzir dados com tipos de estruturas, sobre as quais se pudesse actuar, motivou a dedução de processos para a geração aleatória de árvores de classificação, tendo daí surgido a proposta de duas metodologias: a Geração com Parâmetro de Forma, que permite gerar árvores predominantemente de um tipo prefixado, e a Geração Uniforme.

Na sequência da análise dos vários critérios de Classificação Hierárquica Ascendente, foi realizado o estudo de um conjunto de métodos directos e os correspondentes de metodologia VL – Validade da Ligação, que recorrem a estatísticas de tendência central. Foram estudados e programados os métodos baseados na média geométrica, já propostos anteriormente, e os baseados no centro e na mediana.

Numa Classificação Hierárquica Ascendente ocorrem diferentes tipos de estruturas classificatórias, cuja pertinência de comparação é indiscutível. Foi realizado um estudo, de propriedades e de distribuição, dos diferentes coeficientes estatísticos ordinais disponíveis para tal fim e foi feita uma averiguação da possibilidade de obter coeficientes únicos para as diferentes comparações.

Com o conjunto de ferramentas mencionadas foram definidos protocolos capazes de informar sobre os problemas da comparação de índices, de hierarquias e de índices e hierarquias e sobre os efeitos que perturbações aleatórias têm sobre as mesmas.

Abstract

This thesis begins to situate Classification and Hierarchical Classification problems in the Modern Statistics context, relating mathematics sources that support them. Elaborate a summary of the Cluster Analysis and in particular of the Hierarchical Cluster Analysis mathematics bases, has been the first preoccupation, and one lifting of the main pure and applied actual investigation problems of this area. From this attitude the validation problem and the clustering structures control has been unfolded important.

The Ascendant Hierarchical Cluster method result depends of two aspects: the possible existence of data structure and of the employed method(s). We know that each method has a proper subjacent structure model and that he has capability to optimise in each situation. When a data has no structure or a bad defined structure, the result is a structure that is associated with the subjacent model method.

The data production requirement with structures types about witch we could control, motivated the deduction of classification trees random generation processes, two methodologies had there been proposed: a Parameter Form Generation, to generate trees with a preponderate prefixed type, and a Uniform Generation.

From the analysis of Hierarchical Ascendant Cluster several criterions, has been done the study of a direct methods set and the correspondents with the VL – Validity of the Link methodology, that are based in central character statistics. The methods based in geometric mean, proposed previously, in the middle and in the median were studied and programmed.

Different types of clustering structures take place in Ascendant Hierarchical Cluster, and is indisputable to compare them. A distribution and proprieties study, of the available different ordinal statistics coefficients, has been done and is presented, and the possibility to obtain unique coefficients for the different comparisons was searched.

With the mentioned tools set, protocols able to inform about the indexes, about hierarchies, and about indexes and hierarchies comparison problems and about effects that random disturbances have over them were defined.

Résumé

La dissertation présentée ici commence par situer les problèmes de la Classification et de la Classification Hierarchique dans le contexte de la Statistique Moderne, en rapportant les notions mathématiques que les supportent. Donc la première préoccupation a été de faire une résumée des bases mathématiques d'Analyse Classificatoire, et en particulier d'Analyse Classificatoire Hierarchique, et identifier les problèmes fondamentaux d'investigation, pure et appliquée, de ce domaine de l'actualité. De cette attitude on a relevé le problème de la validation et du contrôle des structures classificatoires.

Le résultat d'une méthode de Classification Hierarchique Ascendante dépend de deux aspects: l'éventuel existence de structure aux données et la(les) méthode(s) utilisée(s), soit d'admettre que chaque méthode a un model propre de structure subjacente, qu'elle est capable d'optimiser à chaque situation. L'application d'un critère classificatoire à un ensemble de données avec une structure mal définie ou même inexistante, a comme résultat une structure qui est associée au modèle subjacent à la méthode.

La nécessité de produire données avec types de structures, sur lesquelles on pouvait opérer, a motivée la déduction des processus pour la génération aléatoire d'arbres de classification. Ont été proposées deux méthodologies: la Génération avec Paramètre de Forme, qui permet générer arbres de types préfixés, et la Génération Uniforme.

À la séquence de l'analyse des différents critères de Classification Hierarchique Ascendante, a été réalisé une étude d'un ensemble de méthodes directes et les correspondants de méthodologie VL – Validité de la Liaison, qui utilisent des statistiques de tendance central. Ont été étudiées et programmées les méthodes basées à la moyenne géométrique, déjà proposées antérieurement, et les basées au centre et a la médiane.

Dans une Classification Hierarchique Ascendante existent différents types de structures classificatoires, qu'il faut comparer. A été réalisé une étude, de propriétés et de distribution, des différents coefficients statistiques ordinaux disponibles pour ce but et a été faite l'étude de la possibilité d'obtenir coefficients uniques pour effectuer les différentes comparaisons.

En utilisant l'ensemble d'outils referaient avant ont été définis protocoles capables d'informer sur les problèmes de la comparaison d'indices, de hierarchies et d'indices et hierarchies et sur les effets qui perturbations aléatoires ont sur les mêmes.

SIMBOLOGIA

$\text{card} ()$	Cardinal de
$\text{int} ()$	Maior inteiro contido em
$\text{sgn} ()$	Sinal de
$\text{exp} ()$	Exponencial de
$\max\{$	Máximo dos elementos de
$\min\{ \}$	Mínimo dos elementos de
$P ()$	Conjunto das partições de
$\tilde{P} ()$	Conjunto das partes de
$\text{gr} ()$	Grafo de uma relação de equivalência
E	Conjunto dos elementos a classificar
F	Conjunto dos pares de elementos de E
Δ	Conjunto dos pares de elementos iguais de F
$\overline{X}$	Média amostral
S^2	Variância amostral
$N(0,1)$	Distribuição normal de média zero e variância um
$\text{Exp} (k)$	Distribuição exponencial de parâmetro k
$\chi^2(k)$	Distribuição qui-quadrado com k graus de liberdade
$U(0,1)$	Distribuição uniforme no intervalo $]0,1[$
$G(k,r)$	Distribuição gama de parâmetros k e r
$\text{Beta}(k,r)$	Distribuição beta reduzida de parâmetros k e r
Φ	Função de distribuição da $N(0,1)$
$B_x ()$	Função beta no ponto x
$B ()$	Função beta no ponto 1
$I_x ()$	Função beta incompleta no ponto x
$\sim$	Segue uma distribuição

$\equiv$	Segue uma distribuição assintótica
VL	Validade da Ligação
C.H.A.	Classificação Hierárquica Ascendente
R_S	Coeficiente Ró de Spearman
T_K	Coeficiente Tau de Kendall
T_{Ler}	Coeficiente de Lerman
T_{GK}	Coeficiente de Goodman - Kruskal
R_S^*	Coeficiente Ró de Spearman reduzido
T_K^*	Coeficiente Tau de Kendall reduzido
T_{Ler}^*	Coeficiente de Lerman reduzido

ÍNDICE DE MATÉRIAS

Capítulo 1 – Introdução

1.1	Motivação e Objectivos	3
1.2	Estrutura do Trabalho	5

Capítulo 2 – Classificação

Introdução	11
2.1 Algumas Definições	13
2.1.1 Os Dados	13
2.1.2 Tipos de Variáveis	13
2.1.3 Funções de Comparação	14
2.1.3.1 Funções de Comparação entre Elementos	14
2.1.3.2 Funções de Comparação entre Classes	18
2.1.4 Métodos Hierárquicos e de Partição	19
2.2 Classificação Hierárquica	21
2.2.1 Hierarquia – Sua Representação	22
2.2.2 Classificação Hierárquica Ascendente.....	25
2.2.2.1 Abordagem Clássica	26
2.2.2.2 Abordagem VL – Validade da Ligação	30
2.3 Árvores de Classificação	33
2.3.1 Conceitos Gerais	33
2.3.2 Geração Aleatória de Árvores	38
2.4 Validação em Classificação	41
2.4.1 Razões para Validar uma Classificação	41
2.4.2 Diferentes Abordagens	42
Referências	49

Capítulo 3 – Novos Métodos de Tendência Central em C.H.A.

Introdução	57
3.1 Métodos Directos – Abordagem Clássica	59
3.1.1 Critério AMG	59
3.1.2 Critério AMed	61
3.1.3 Critério AC	63
3.2 Métodos VL – Abordagem Validade da Ligação	64
3.2.1 Critério AVMg	65
3.2.2 Critério AVMed	69
3.2.2.1 Resultados Preliminares	69
3.2.2.2 Dimensão Amostral Ímpar	72
3.2.2.3 Dimensão Amostral Par	74
3.2.3 Critério AVC	79
Referências	81

Capítulo 4 – Geração Aleatória de Árvores de Classificação

Introdução	85
4.1 Metodologias para a Geração de Árvores	87
4.1.1 Geração de Árvores com Parâmetro de Forma	88
4.1.2 Geração Uniforme de Árvores	89
4.2 Análise das Diferentes Metodologias	89
4.2.1 Algumas Considerações	89
4.2.2 Comportamento Teórico das Diferentes Metodologias	98
4.2.3 Algumas Simulações	103
Referências	109

Capítulo 5 – Comparação Ordinal de Estruturas Classificatórias

Introdução	113
5.1 Comparação de Ordenações	119
5.1.1 Coeficientes de Comparação	120
5.1.2 Comparação Analítica dos Vários Coeficientes	122

5.1.3	Comportamento Assintótico dos Coeficientes	124
5.1.4	A Distribuição Exacta da Estatística C	128
5.2	Comparação de Preordenações	138
5.2.1	Coeficientes de Comparação	139
5.3	Comparação de Uma Ordenação e Uma Preordenação	144
5.3.1	Análise dos Vários Coeficientes	144
5.3.2	A Distribuição Exacta da Estatística C	148
5.3.2.1	Preordenação Associada a Uma Cadeia Pura	148
5.3.2.2	Preordenação Associada a Uma Árvore Equilibrada	153
	Referências	157

Capítulo 6 – Validação Estocástica em Classificação Hierárquica

	Introdução	161
6.1	Procedimento de Simulação	163
6.2	Alguns Resultados	167
6.2.1	Comparação $M_0 - H_0$	167
6.2.2	Comparação $M_0 - M_i$	169
6.2.3	Comparação $M_i - H_i$	171
6.2.4	Comparação $H_0 - H_i$	180
6.3	Conclusões	189

Capítulo 7 – Notas Finais e Desenvolvimentos Futuros

7.1	Notas Finais	195
7.1.1	Aspectos Computacionais.....	196
7.1.2	Conclusões	197
7.2	Desenvolvimentos Futuros.....	197

Anexo 1	– Demonstração do Resultado 3.4	199
---------	---------------------------------------	-----

Anexo 2 – Distribuições Exacta e Aproximada da Estatística C para

Ordenações	205
-------------------------	------------

ÍNDICE DE FIGURAS

Capítulo 2

2.1 – Dendrograma e respectiva hierarquia de partições	23
2.2 – Matriz cofenética associada ao dendrograma da Fig. 2.1	24

Capítulo 4

4.1 – Esquema de procedimento de actuação num nó	87
4.2 – Representação esquemática dos dendrogramas para $m = 4$	91
4.3 – Representação esquemática dos dendrogramas para $m = 5$	91
4.4 – Representação esquemática dos dendrogramas para $m = 6$	92
4.5 – Representação esquemática dos dendrogramas para $m = 7$	93
4.6 – Tipos de dendrogramas para $m = 4$	94
4.7 – Tipos de dendrogramas para $m = 5$	94
4.8 – Tipos de dendrogramas para $m = 6$	94
4.9 – Tipos de dendrogramas para $m = 7$	95
4.10 – Representação da Árvore	97
4.11 – Representação Polaca da Árvore	97

Capítulo 5

5.1 – Distribuições exacta e aproximada de C_M , para $M = 4$	137
5.2 – Distribuições exacta e aproximada de C_M , para $M = 10$	137
5.3 – Distribuições exacta e aproximada de C_M (caso cadeia pura), para $M = 6$	154
5.4 – Distribuições exacta e aproximada de C_M (caso equilibrado), para $M = 6$	155

Capítulo 6

6.1 – Extremos e quartis de $\overline{TGK_i^3(0.05)}$, na <i>Geração Uniforme</i>	172
6.2 – Extremos e quartis de $\overline{TGK_i^3(0.15)}$, na <i>Geração Uniforme</i>	173
6.3 – Extremos e quartis de $\overline{TGK_i^3(0.25)}$, na <i>Geração Uniforme</i>	173
6.4 – Extremos e quartis de $\overline{TGK_i^3(0.5)}$, na <i>Geração Uniforme</i>	174
6.5 – Extremos e quartis de $\overline{TGK_i^3(0.05)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	175
6.6 – Extremos e quartis de $\overline{TGK_i^3(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	175
6.7 – Extremos e quartis de $\overline{TGK_i^3(0.25)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	176
6.8 – Extremos e quartis de $\overline{TGK_i^3(0.5)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	177
6.9 – Extremos e quartis de $\overline{TGK_i^3(0.05)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	177
6.10 – Extremos e quartis de $\overline{TGK_i^3(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	178
6.11 – Extremos e quartis de $\overline{TGK_i^3(0.25)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	179
6.12 – Extremos e quartis de $\overline{TGK_i^3(0.5)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	179
6.13 – Extremos e quartis de $\overline{TGK_i^4(0.05)}$, na <i>Geração Uniforme</i>	181
6.14 – Extremos e quartis de $\overline{TGK_i^4(0.15)}$, na <i>Geração Uniforme</i>	182
6.15 – Extremos e quartis de $\overline{TGK_i^4(0.25)}$, na <i>Geração Uniforme</i>	183
6.16 – Extremos e quartis de $\overline{TGK_i^4(0.5)}$, na <i>Geração Uniforme</i>	183
6.17 – Extremos e quartis de $\overline{TGK_i^4(0.05)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	184
6.18 – Extremos e quartis de $\overline{TGK_i^4(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	185
6.19 – Extremos e quartis de $\overline{TGK_i^4(0.25)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	185
6.20 – Extremos e quartis de $\overline{TGK_i^4(0.5)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	186
6.21 – Extremos e quartis de $\overline{TGK_i^4(0.05)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	187
6.22 – Extremos e quartis de $\overline{TGK_i^4(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	187
6.23 – Extremos e quartis de $\overline{TGK_i^4(0.25)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	188

6.24 – Extremos e quartis de $\overline{TGK_i^4(0.5)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	189
---	-----

ÍNDICE DE QUADROS

Capítulo 2

2.1 – Valores de constantes na fórmula de Lance & Williams para alguns critérios	29
2.2 – Número de árvores não isomórficas de três categorias diferentes	37

Capítulo 4

4.1 – Valores de probabilidade para $m = 4$	100
4.2 – Valores de probabilidade para $m = 5$	101
4.3 – Valores de probabilidade para $m = 6$	101
4.4 – Valores de probabilidade para $m = 7$	102
4.5 – Geração Uniforme, $m = 4$	104
4.6 – Geração Uniforme, $m = 5$	104
4.7 – Geração Uniforme, $m = 6$	105
4.8 – Geração Uniforme, $m = 7$	105
4.9 – Permutação Dupla, $m = 5$	106
4.10 – Permutação Dupla, $m = 6$	106

Capítulo 5

5.1 – Valores médios e variâncias obtidos para R_S^*	125
5.2 – Valores médios e variâncias obtidos para $T_K^* \equiv T_{Ler}^*$	125
5.3 – Função de probabilidade de C_2	130
5.4 – Função de probabilidade de C_3	130
5.5 – Função de probabilidade de C_4	132
5.6 – Valores de Λ e de $V(T_{Ler})$	142

5.7 – Função de probabilidade de C_3 (caso cadeia pura)	151
5.8 – Valores médios e de variância de C_M (caso cadeia pura).....	154
5.9 – Valores médios e de variância de C_M (caso equilibrado)	155

Capítulo 6

6.1 – Medidas descritivas dos valores de TGK^1	168
6.2 – Frequência de cada método na melhor recuperação da estrutura inicial.....	168
6.3 – Medidas descritivas dos valores de $\overline{TGK^2(\delta)}$ na Geração Uniforme	170
6.4 – Medidas descritivas dos valores de $\overline{TGK^2(\delta)}$ na Geração com Parâmetro de Forma com $p = 0.025$	170
6.5 – Medidas descritivas dos valores de $\overline{TGK^2(\delta)}$ na Geração com Parâmetro de Forma com $p = 0.5$	170
6.6 – Medidas descritivas de $\overline{TGK_i^3(0.05)}$, na Geração Uniforme	172
6.7 – Medidas descritivas de $\overline{TGK_i^3(0.15)}$, na Geração Uniforme	172
6.8 – Medidas descritivas de $\overline{TGK_i^3(0.25)}$, na Geração Uniforme	173
6.9 – Medidas descritivas de $\overline{TGK_i^3(0.5)}$, na Geração Uniforme	174
6.10 – Medidas descritivas de $\overline{TGK_i^3(0.05)}$, na G. Par. de Forma com $p = 0.025$	174
6.11 – Medidas descritivas de $\overline{TGK_i^3(0.15)}$, na G. Par. de Forma com $p = 0.025$	175
6.12 – Medidas descritivas de $\overline{TGK_i^3(0.25)}$, na G. Par. de Forma com $p = 0.025$	176
6.13 – Medidas descritivas de $\overline{TGK_i^3(0.5)}$, na G. Par. de Forma com $p = 0.025$	176
6.14 – Medidas descritivas de $\overline{TGK_i^3(0.05)}$, na G. Par. de Forma com $p = 0.5$	177
6.15 – Medidas descritivas de $\overline{TGK_i^3(0.15)}$, na G. Par. de Forma com $p = 0.5$	178
6.16 – Medidas descritivas de $\overline{TGK_i^3(0.25)}$, na G. Par. de Forma com $p = 0.5$	178
6.17 – Medidas descritivas de $\overline{TGK_i^3(0.5)}$, na G. Par. de Forma com $p = 0.5$	179
6.18 – Medidas descritivas de $\overline{TGK_i^4(0.05)}$, na Geração Uniforme	181
6.19 – Medidas descritivas de $\overline{TGK_i^4(0.15)}$, na Geração Uniforme	182

6.20 – Medidas descritivas de $\overline{TGK_i^4(0.15)}$, na <i>Geração Uniforme</i>	182
6.21 – Medidas descritivas de $\overline{TGK_i^4(0.5)}$, na <i>Geração Uniforme</i>	183
6.22 – Medidas descritivas de $\overline{TGK_i^4(0.05)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	184
6.23 – Medidas descritivas de $\overline{TGK_i^4(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	184
6.24 – Medidas descritivas de $\overline{TGK_i^4(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	185
6.25 – Medidas descritivas de $\overline{TGK_i^4(0.5)}$, na <i>G. Par. de Forma</i> com $p = 0.025$	186
6.26 – Medidas descritivas de $\overline{TGK_i^4(0.05)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	186
6.27 – Medidas descritivas de $\overline{TGK_i^4(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	187
6.28 – Medidas descritivas de $\overline{TGK_i^4(0.15)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	188
6.29 – Medidas descritivas de $\overline{TGK_i^4(0.5)}$, na <i>G. Par. de Forma</i> com $p = 0.5$	188

Anexo 2

1 – Distribuições exacta e aproximada de C para $M = 4$	207
2 – Distribuições exacta e aproximada de C para $M = 5$	208
3 – Distribuições exacta e aproximada de C para $M = 6$	208
4 – Distribuições exacta e aproximada de C para $M = 7$	209
5 – Distribuições exacta e aproximada de C para $M = 8$	210
6 – Distribuições exacta e aproximada de C para $M = 9$	211
7 – Distribuições exacta e aproximada de C para $M = 10$	212

CAPÍTULO 1
INTRODUÇÃO

CAPÍTULO 1

INTRODUÇÃO

ÍNDICE

1.1 MOTIVAÇÃO E OBJECTIVOS

1.2 ESTRUTURA DO TRABALHO

CAPÍTULO 1

INTRODUÇÃO

1.1 MOTIVAÇÃO E OBJECTIVOS

O trabalho aqui apresentado, insere-se em Classificação Hierárquica, e é uma abordagem às metodologias e bases probabilístico-estatísticas da validação de estruturas classificatórias hierárquicas.

Análise Classificatória será aqui entendida como uma área da Análise de Dados, em particular, e da Estatística em geral, que tem por objectivo agrupar o conjunto de objectos a classificar num número relativamente pequeno de classes, satisfazendo a condição de que objectos de uma mesma classe sejam semelhantes entre si e objectos de classes diferentes sejam dissemelhantes. Note-se que os grupos ou classes de elementos são desconhecidos à partida e são construídos ao longo do desenvolvimento do método. Análise Classificatória não é Análise Discriminante. Uma Classificação Hierárquica é uma sucessão encaixada de partições, com propriedades particulares, como será enunciado mais adiante.

Na literatura muitos têm sido os algoritmos ou critérios propostos para a obtenção de hierarquias de classificações. O acentuado desenvolvimento na área da informática, nas últimas décadas, contribuiu de forma decisiva, para a simplificação da implementação e execução desses algoritmos. Paralelamente a evolução tecnológica, permite armazenar grandes conjuntos de dados, de difícil compreensão no seu todo e cuja utilidade depende da capacidade de extrair deles a informação principal.

Sendo a Classificação um instrumento poderoso no estudo e tratamento de dados, é necessário garantir que a divisão em classes, sugerida por esse procedimento automático, não distorça a estrutura do conjunto de dados inicial, indicando a existência de classes sem validade real. Lembre-se o facto inconveniente, de a aplicação de um algoritmo de classificação hierárquica,

originar sempre uma hierarquia de partições, mesmo quando os dados em estudo não têm qualquer estrutura definida.

Os aspectos anteriormente referidos, não devem ser vistos como limitações da metodologia, antes funcionam como alerta para a necessidade de considerar os resultados de certos métodos, como propostas a confrontar, numa perspectiva científica, com a realidade.

O recurso crescente a metodologias de Classificação, por estatísticos e por simples utilizadores, levou alguns estudiosos a preocuparem-se com a necessidade de implementar medidas da qualidade do produto final e de ajuda à interpretação e à tomada de decisões. A envolvente a este tipo de preocupações e consequentes contribuições para as suas resoluções, será aqui designada por Validação em Classificação.

Têm surgido diversos trabalhos nesta área, com perspectivas e formalização bem distintas, conforme a qualidade dos dados e a natureza dos problemas. Algumas abordagens recorrem a noções e metodologias da Estatística Matemática, outras validam por simulação. A validação de uma classificação pode ser realizada a níveis diferentes; têm surgido trabalhos para avaliar a qualidade e consistência de uma classe da hierarquia, validar uma certa partição ou mesmo toda a hierarquia produzida. Esses trabalhos referem-se sobretudo a classes ou partições. No entanto o número de trabalhos neste tema é insuficiente, face à complexidade e dimensão do problema, e os resultados são, congresso após congresso, considerados como pobres e não adequados, apesar de sistematicamente se considerar o assunto como prioritário.

A tese aqui apresentada, pretende ser uma contribuição para o tema Validação em Classificação, os objectos a validar são as hierarquias de classificações e recorre ao uso da simulação. A metodologia desenvolvida tem por base três componentes, a geração aleatória de árvores, os algoritmos de Classificação Hierárquica Ascendente e a comparação dos diversos tipos de estruturas que ocorrem em Classificação.

Vários factores intervêm na obtenção de uma classificação hierárquica e é importante criar sensibilidade ao grau e tipo de dependência do produto final em relação a cada um dos factores.

1.2 ESTRUTURA DO TRABALHO

A presente dissertação é composta por esta apresentação do trabalho, em linhas bastante gerais, e por mais seis capítulos, sendo o primeiro destes uma síntese do estado do conhecimento das matérias abordadas.

A estrutura deste trabalho está fortemente ligada às áreas de desenvolvimento e de investigação necessárias para atingir os objectivos a que nos propusemos. Como já foi referido na secção anterior, a metodologia de validação proposta no capítulo 6, assenta em três componentes distintas, mas interligadas, que correspondem aos temas dos capítulos 3, 4 e 5.

No capítulo 2, intitulado de “Classificação”, são inicialmente apresentadas diversas noções matemáticas, estatísticas e de classificação, que intervêm no desenvolvimento da tese. Em seguida é exposta a metodologia de Classificação Hierárquica, abordando as várias representações associadas. Os métodos hierárquicos ascendentes são divididos em dois tipos: os que seguem uma abordagem, aqui denominada de clássica, em oposição à outra, referida por VL – Validade da Ligação, que recorre a noções probabilísticas para definir as funções de comparação. A secção 2.3 é dedicada ao enquadramento da noção de árvore em Classificação e são introduzidos os conceitos e a necessidade da geração aleatória de árvores. Ainda no mesmo capítulo é feita uma síntese das várias abordagens de Validação em Classificação.

Um conjunto de novos métodos de Classificação Hierárquica Ascendente, tendo em comum o recurso a estatísticas de tendência central, para a definição dos critérios de agregação, são propostos no capítulo 3. Estes métodos encontram-se divididos em duas categorias, de acordo com o tipo de abordagem subjacente. Na primeira secção são apresentados os métodos que recorrem a uma metodologia directa, abordagem clássica, e são analisadas as suas propriedades teóricas. A segunda secção é dedicada aos critérios que utilizam uma metodologia VL, os quais conduzem à dedução de funções de distribuição das estatísticas envolvidas. Associado a este último aspecto aparecem as questões de programação desses métodos, que são também aqui analisadas.

A geração aleatória de árvores de classificação é o assunto do capítulo 4. Aí são apresentadas duas metodologias novas para a simulação de árvores de classificação, tendo em conta a sua

topologia, os seus níveis ou patamares e as etiquetas das folhas ou nós terminais, isto é as árvores aqui geradas são dendrogramas. O primeiro método gera árvores em função de um parâmetro de forma, o que permite privilegiar uns tipos topológicos relativamente a outros, de acordo com o objectivo pretendido. O segundo processo gera árvores aleatória e uniformemente do conjunto de árvores distintas que é possível definir para um determinado número de nós terminais. São estudadas as características teóricas das árvores geradas por estes dois processos e o seu desempenho é analisado via simulação. Para analisar o comportamento do segundo processo recorre-se à comparação deste com uma metodologia de geração de árvores, proposta por Lapointe & Legendre.

No capítulo 5 é feito um estudo sobre a comparação ordinal de estruturas classificatórias. Atendendo a que as relações de ordem desempenham um papel relevante em Estatística e ao facto de que às várias estruturas de classificação podemos associar sempre uma ordenação ou uma preordenação, reduz-se o problema da comparação ordinal de estruturas classificatórias à comparação de duas ordenações, duas preordenações ou uma ordenação com uma preordenação. São estudados vários coeficientes, em particular os propostos por Kendall e por Lerman, para medir a concordância entre as duas estruturas, para cada tipo de comparação, bem como as respectivas distribuições assintóticas. No caso da comparação de duas ordenações, é feita a comparação analítica dos vários coeficientes, mostrando que todos eles são função de um único, do qual é deduzida a distribuição exacta. Para os vários coeficientes analisados, mostrou-se que se pode considerar o coeficiente de comparação de duas preordenações, como um coeficiente geral para os três tipos de comparações.

Uma metodologia de validação de classificação é apresentada no capítulo 6. A geração de dendrogramas é usada para obter matrizes de dados ultramétricos, os quais são submetidos a diferentes critérios de Classificação Hierárquica Ascendente, e consequente comparação entre a estrutura inicial existente nos dados e a estrutura hierárquica obtida para cada critério. Esta comparação permite medir a capacidade de cada método em recuperar a estrutura existente. Posteriormente cada matriz de dados é submetida a perturbação aleatória, a partir da qual são obtidas novas classificações hierárquicas e efectuados vários tipos de comparações. Com estas comparações pode medir-se a capacidade dos vários métodos em recuperar a estrutura, quando esta se afasta de uma estrutura ultramétrica, e a influência que o tipo de estrutura dos dados tem no comportamento dos diferentes métodos.

Finalmente, no capítulo 7, é feita uma avaliação do trabalho realizado, focando aspectos de programação subjacentes ao desenvolvimento da metodologia aqui apresentada e alguns trabalhos de aplicação realizados e não incluídos na tese por opção. Uma secção de conclusões finais e uma de trabalho futuro a desenvolver completam este capítulo.

Subjacente ao trabalho desenvolvido, embora não haja referências muito específicas, está o ajustamento, a integração e a verificação dos novos métodos na biblioteca protótipo de programas de Classificação Hierárquica, CLUSTER, que vai presentemente na versão 7.

CAPÍTULO 2
CLASSIFICAÇÃO

CAPÍTULO 2

CLASSIFICAÇÃO

ÍNDICE

INTRODUÇÃO

2.1 ALGUMAS DEFINIÇÕES

2.1.1 Os Dados

2.1.2 Tipos de Variáveis

2.1.3 Funções de Comparação

2.1.3.1 Função de Comparação entre Elementos

2.1.3.2 Função de Comparação entre Classes

2.1.4 Métodos Hierárquicos e de Partição

2.2 CLASSIFICAÇÃO HIERÁRQUICA

2.2.1 Hierarquia – Sua Representação

2.2.2 Classificação Hierárquica Ascendente

2.2.2.1 Abordagem Clássica

2.2.2.2 Abordagem VL – Validade da Ligação

2.3 ÁRVORES DE CLASSIFICAÇÃO

2.3.1 Conceitos Gerais

2.3.2 Geração Aleatória de Árvores

2.4 VALIDAÇÃO EM CLASSIFICAÇÃO

2.4.1 Razões para Validar uma Classificação

2.4.2 Diferentes Abordagens

REFERÊNCIAS

CAPÍTULO 2

CLASSIFICAÇÃO

INTRODUÇÃO

Comparar e classificar são duas actividades complementares, quase sempre ligadas na prática. Comparam-se objectos ou indivíduos entre si, segundo determinadas características analisadas ou de acordo com um dado ponto de vista. Esta comparação conduz à formação de grupos de elementos que se assemelham mais entre si. As classes de objectos equivalentes, assim constituídas, são em seguida ordenadas de acordo com relações adequadas.

De uma forma geral, pode afirmar-se que a Classificação tem por objectivo definir sobre um conjunto de objectos, comparáveis dois a dois, partições ou hierarquias de partições, que respeitem o melhor possível as analogias (semelhanças) entre esses objectos.

Neste contexto ao referir o termo Classificação, não estarão incluídas metodologias que, partindo de um conjunto de classes pré-definidas e de um conjunto de elementos a classificar, atribuem cada elemento a uma das classes. Nesta perspectiva as classes não são conhecidas *a priori*, mas são conhecidas passo a passo.

Nas últimas décadas o acentuado desenvolvimento computacional levou a um aparecimento de novas metodologias em Classificação e a uma crescente aplicação em áreas muito diversas. Vários estatísticos se têm preocupado em alertar para alguns problemas metodológicos existentes nesta área e para alguns cuidados necessários quando da utilização dos métodos. Neste contexto pode referir-se algumas importantes contribuições; Bock [12] e [13], Cormack [18], Gordon [30] e [31].

Este capítulo sobre Classificação está organizado em quatro partes distintas. Na primeira secção são abordadas noções de base para este trabalho e apresentadas algumas definições

matemáticas associadas, de forma a tornar claro o desenvolvimento posterior do tema em questão.

A segunda secção deste capítulo é dedicada à Classificação Hierárquica, inicialmente em termos gerais, sendo depois apresentada com mais pormenor a metodologia associada à Classificação Hierárquica Ascendente, que ao longo deste texto será representada por *C.H.A.*. Referem-se alguns critérios de agregação, que pelo seu desempenho ou relevância histórica, são considerados importantes. É apresentada a metodologia de *C.H.A.* que recorre a noções probabilísticas, para a definição das funções de comparação, conhecida por metodologia *VL* (Validade da Ligação).

Na terceira secção, dedicada a árvores de classificação, são abordados vários conceitos e definições sobre árvores em geral, e sobre árvores de classificação, ou dendrogramas, em particular. Neste trabalho recorre-se à geração aleatória de árvores, sendo este tema aqui abordado.

A validação em Classificação é o assunto da última secção. São expostas algumas definições e a terminologia mais pertinentes ligadas a este tema. Referem-se os trabalhos e abordagens mais relevantes, do ponto de vista deste trabalho.

2.1 ALGUMAS DEFINIÇÕES

Nesta secção são apresentados alguns tópicos essenciais à introdução da Classificação, como tema inserido na Análise de Dados. Esta abordagem envolve noções matemáticas que, sendo elementares, constituem um conjunto de suporte de ferramentas essenciais para o desenvolvimento deste trabalho.

2.1.1 Os Dados

Podemos considerar que os métodos classificatórios operam sobre dois tipos de dados: um quadro de dados iniciais e um quadro de proximidades. Em geral admite-se que o quadro de dados iniciais é do tipo dados indivíduos - variáveis, apresentando-se sob a forma de matriz, $n \times p$, $X = (x_{ik})$ em que x_{ik} representa o valor da variável de ordem k observada no indivíduo i . Muitos métodos de Classificação estão preparados para trabalhar com uma matriz de proximidades, exigindo uma transformação prévia da matriz de dados iniciais numa matriz de proximidades. Este quadro de proximidades pode ser calculado para os n indivíduos, dando origem a uma matriz $n \times n$, em que o elemento (i, j) representa a proximidade entre os indivíduos i e j , ou para as variáveis, matriz $p \times p$, em que o elemento (i, j) representa a proximidade entre as variáveis i e j . Estas matrizes de proximidades são geralmente admitidas como simétricas e podem traduzir semelhanças ou dissemelhanças de acordo com o tipo de função que lhe dá origem.

2.1.2 Tipos de Variáveis

O tratamento estatístico de uma ou mais variáveis depende fortemente da sua(s) natureza(s), e diferentes tipos de variáveis podem ser usados para descrever um objecto. Um quadro de dados pode ser homogéneo ou heterogéneo quanto ao tipo de variáveis que o compõem. Neste trabalho, por simplicidade de apresentação e se nada for dito em contrário, irá considerar-se que as variáveis de um quadro de dados são todas do mesmo tipo.

Diferentes tipologias de caracterização de variáveis podem ser encontradas na literatura, Anderberg [1], Bacelar-Nicolau [3], Benzecri [11], Celeux et al [15], Gordon [26], Lerman

[50], Nicolau [57], ect. Na tipologia aqui apresentada optou-se por definir os grandes grupos, sendo assim bastante consensual.

- *Variáveis numéricas ou quantitativas* – o conjunto de chegada da aplicação associada a cada variável aleatória é $\mathbb{R}$, as variáveis aleatórias induzem preordens.
- *Variáveis nominais ou ordinais* – a cada variável aleatória está associado um conjunto de estados, ou modalidades, e a um objecto corresponde um e um só estado de cada variável. Sobre os estados de uma variável ordinal pode impor-se uma ordem, o que já não acontece para os estados de uma variável nominal que não podem ser ordenados.
- *Variáveis binárias* – é um caso particular das variáveis nominais ou ordinais, que ocorre quando o número de estados possíveis de uma variável é igual a dois. Este tipo de variável, quer pela ocorrência frequente na prática quer pela simplicidade de tratamento, tem sido objecto de diversos estudos estatísticos.

Como já foi mencionado, caracterizações mais pormenorizadas podem ser encontradas nos trabalhos acima referenciados.

2.1.3 Funções de Comparação

2.1.3.1 Função de Comparação entre Elementos

Dado um conjunto E de m elementos a classificar, define-se função de comparação entre elementos de E como uma aplicação $\gamma : E \times E \rightarrow \mathbb{R}_0^+$. Uma função de comparação pode ser do tipo dissemelhança, para dois elementos quaisquer i e j de E $\gamma(i, j) = d_{ij}$, em que um pequeno valor de d_{ij} significa que i e j são elementos bastante análogos, ou do tipo semelhança com $\gamma(i, j) = s_{ij}$, onde um pequeno valor de s_{ij} traduz grandes diferenças entre os elementos em causa. Assim a semelhança entre dois elementos é tanto mais forte quanto mais semelhantes eles são e vale $s_{\max}$ se e só se os elementos “são, sob todos os pontos de vista analisados, idênticos”¹.

¹ Lembre-se que nem todos os coeficientes são bem comportados, podendo por isso ter elementos que “são sobre todos os pontos de vista analisados idênticos” mas que não conduzem ao valor máximo de semelhança. Na metodologia de classificação VL, este inconveniente é eliminado.

Em seguida apresentam-se algumas propriedades desejáveis para uma função de comparação, quando esta é do tipo dissemelhança.

Seja $d : E \times E \rightarrow \mathbb{R}_0^+$ uma aplicação e considerem-se as condições:

- i. $d(x, x) = 0, \forall x \in E$
- ii. $d(x, y) = d(y, x), \forall x, y \in E$
- iii. $d(x, y) = 0 \Rightarrow x = y$ (2.1)
- iv. $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in E$
- v. $d(x, y) \leq \max(d(x, z), d(z, y)), \forall x, y, z \in E$

Se d satisfaz i. e ii. diz-se um *índice de dissemelhança* (ou apenas uma *dissemelhança*).

Se d é um índice de dissemelhança e satisfaz iii. diz-se um *índice de distância*.

Se d é um índice de distância e satisfaz iv. diz-se uma *distância*.

Se d é um índice de distância que verifica v. diz-se uma *distância ultramétrica*.

A condição v., chamada de condição ultramétrica, é mais restritiva que a desigualdade triangular e tem uma grande aplicabilidade em Classificação Hierárquica.

A condição ultramétrica tem uma interpretação geométrica simples; dados três pontos quaisquer de E eles formam sempre um triângulo que ou é equilátero ou isósceles em que a medida da base é inferior à dos dois lados iguais. Um espaço ultramétrico goza de propriedades específicas, das quais se referem as seguintes:

- Seja B uma bola, aberta ou fechada, num espaço ultramétrico. Então qualquer ponto de B pode ser o seu centro.
- Dadas duas bolas B_1 e B_2 num espaço ultramétrico, cada uma podendo ser aberta ou fechada, cuja intersecção é não vazia, então uma delas está contida na outra.

Obtém-se uma caracterização paralela, à feita anteriormente, usando semelhanças e considerando $s(x, x) = s_{\max}, \forall x \in E$.

A escolha do tipo de função de comparação depende essencialmente da natureza dos elementos envolvidos. Neste trabalho, salvo situação devidamente clarificada, usaremos semelhanças; esta opção é pouco restritiva dado que a passagem de um tipo de medida ao outro é em geral uma transformação simples¹.

Há vantagens na escolha de uma função de comparação que seja um índice de dissemelhança (semelhança), nomeadamente são suficientes $m(m-1)/2$ valores de $d_{ij}(s_{ij})$ para definir a matriz dos valores da função de comparação associada ao conjunto E de cardinal m .

Com frequência surgem definições de dissemelhanças (semelhanças) admitindo que as aplicações se encontram reduzidas, isto é, assumindo valores em $[0,1]$.

Muitas medidas de dissemelhanças e de semelhanças têm sido propostas, para um resumo ver Anderberg [1] (capítulos 4 e 5), Gower [33]. Considerações e algumas recomendações sobre a selecção de uma medida de dissemelhança podem ser vistas em Gordon [26], [28] e [30].

Em Bacelar-Nicolau [3] é feito um estudo sobre funções de comparação entre elementos para dados binários, no qual é apresentada uma vasta lista de coeficientes, agrupando-os em classes. Nesse trabalho é ainda realizado um estudo das distribuições dos referidos coeficientes.

A escolha da função de comparação é um passo importante e dela pode depender bastante o resultado final. Uma forma de reduzir a susceptibilidade da escolha do índice de dissemelhança (semelhança) é conseguida se substituirmos os valores numéricos do índice pela ordenação desses valores.

Seja E o conjunto a classificar, escolhido um índice de dissemelhança associa-se um valor numérico a cada elemento de $E \times E$. Ao reter apenas a ordenação associada a esses valores a informação fica menos rica. Apesar da ordenação depender da função de comparação escolhida, é de esperar que seja mais estável do que a utilização directa dos respectivos valores numéricos. Em Lerman [50] são apresentados alguns resultados sobre a substituição dos valores numéricos pelas respectivas ordenações, e respectiva estabilidade, no caso particular de dados binários.

¹Apesar da estreita relação existente entre os valores de dissemelhanças e os de semelhanças, usualmente $d_{ij} + s_{ij} = \text{constante}$, tal relação não pode ser assumida sempre como verdadeira.

As relações de ordem e de preordem desempenham um papel bastante importante em Classificação e em particular no estudo aqui desenvolvido, o que justifica as definições que se seguem.

Seja R uma relação binária sobre um conjunto E e considerem-se as seguintes condições:

- i. $x \neq y \wedge xRy \Rightarrow y \text{ não } Rx$
 - ii. $x, y, z \in E, xRy \wedge yRz \Rightarrow xRz$
 - iii. $x \in E \Rightarrow x \text{ não } Rx$
 - iv. $x \in E \Rightarrow xRx$
- (2.2)

Se R satisfaz i. diz-se uma *relação binária antisimétrica*.

Se R satisfaz i. e ii. diz-se uma *relação de ordem*.

Se R satisfaz ii. diz-se uma *relação de preordem*.

Se R é uma relação de ordem e satisfaz iii. diz-se uma *ordem estrita*.

Se R é uma relação de preordem e satisfaz iii. diz-se uma *preordem estrita*.

Se R é uma relação de ordem e satisfaz iv. diz-se uma *ordem reflexiva* (ou apenas *ordem*).

Se R é uma relação de preordem e satisfaz iv. diz-se uma *preordem reflexiva* (ou apenas *preordem*).

No que se segue, de uma forma geral, considera-se que uma ordem é uma relação binária reflexiva, antisimétrica e transitiva e que uma preordem é uma relação binária reflexiva e transitiva.

Outro tipo de classificação de uma ordem (preordem) tem a ver com o número de elementos comparáveis. Assim uma ordem (preordem) para a qual dois elementos distintos quaisquer são comparáveis, isto é: $x, y \in E \wedge x \neq y \Rightarrow xRy \vee yRx$, diz-se uma *ordem (preordem) total*. Uma ordem (preordem) para a qual existem elementos não comparáveis diz-se uma *ordem (preordem) parcial*.

Considerem-se agora as relações binárias entre pares de elementos de E .

Para a resolução de muitos problemas não é necessário conhecer os valores de distância entre os vários pares de elementos de E , sendo suficiente o conhecimento das desigualdades entre os valores dessas distâncias. Assim surge a noção de ordenação de um conjunto de valores.

Seja O uma relação binária em $E \times E$ que verifica as seguintes condições¹:

- i. $\forall x, y, z \in E \Rightarrow (x, x)O(y, z)$
 - ii. $\forall x, y \in E \Rightarrow (x, y)O(y, x)$
 - iii. $\forall x, y, z \in E \Rightarrow (y, z)O(x, x) \Rightarrow (y = z)$
- (2.3)

Se O for uma relação de ordem, preordem, em $E \times E$, então O diz-se uma *ordenação*, respectivamente uma *preordenação*, em E . Além disso O diz-se uma *ordenação total*, *preordenação total*, em E se for uma relação de ordem total, preordem total em $E \times E$.

Seja G o conjunto das partes de dois elementos de E , $G = \{\{x, y\} : x \in E \wedge y \in E\}$. Dar uma preordenação sobre E é equivalente a dar uma preordem sobre G , Benzecri [11] (Tome I, Parte B, nº 3).

Alternativamente, seja F o conjunto dos pares de elementos distintos de E , isto é, $F = \{(x, y) \in E \times E : x \neq y\}$. Pode então dizer-se que qualquer preordem (ordem) total sobre F é uma *preordenação (ordenação)* sobre E .

Definição 2.1

Duas medidas de dissemelhança / semelhança, sobre E , dizem-se equivalentes se e só se as preordenações associadas, sobre E , forem idênticas.

2.1.3.2 Função de Comparação entre Classes

Uma função de comparação entre os elementos de E dá origem a uma matriz de semelhanças ou dissemelhanças, matriz esta que foi chamada de matriz de proximidades em 2.1.1, quando se referiram os tipos de dados de uma classificação. Sobre os valores desta matriz tomam-se decisões de agregação dos elementos em classes, segundo um determinado critério. O objectivo de uma classificação é formar classes de elementos mais semelhantes entre si do que entre elementos de classes diferentes, sendo por isso natural que comece por agregar o par

¹ Atendendo a que a relação $(x, y)O(z, w)$ tem a leitura “ x está mais próximo de y do que z está de w ”, as condições i., ii. e iii. tornam-se claras.

de elementos de E onde a função de comparação é máxima (mínima), quando esta é do tipo semelhança (dissemelhança). Para as comparações seguintes é necessário prolongar a noção de função de comparação a classes de elementos de E .

Define-se função de comparação entre partes de E como sendo uma aplicação $\Gamma: \tilde{P}(E) \times \tilde{P}(E) \rightarrow \mathfrak{R}_0^+$, onde $\tilde{P}(E)$ representa o conjunto das partes de E . Γ deve verificar algumas propriedades, algumas delas já referidas no ponto anterior. Assim dadas duas partes disjuntas, A e B , de elementos de E , a função de comparação entre A e B , $\Gamma(A, B)$, traduz o resultado da aplicação do critério de agregação a essas partes.

Existe na literatura um elevado número de critérios de agregação. Em 2.2 serão considerados os mais relevantes, quer pelas suas propriedades, quer pela sua relação com este trabalho.

2.1.4 Métodos Hierárquicos e de Partição

Como já foi referido anteriormente um procedimento classificatório define partições ou hierarquias de partições óptimas, relativamente a um dado critério estabelecido. Antes de prosseguir com o desenvolvimento deste assunto, para que a leitura seja mais clara, lembrem-se algumas noções matemáticas.

Uma *partição* de um conjunto E é um conjunto de partes de E , mutuamente disjuntas, chamadas de classes da partição, cuja reunião é E . Designe-se por $P(E)$ o conjunto das partições de E .

Definir uma partição π de E corresponde a estabelecer uma relação de equivalência¹ em E (dois elementos x e y de E são equivalentes se e só se pertencerem a uma mesma classe de π). Analogamente toda a relação de equivalência sobre E define uma partição.

Partição e relação de equivalência são duas maneiras distintas, mas logicamente equivalentes, de formalizar matematicamente o conceito de classificação.

¹ Relação binária reflexiva, simétrica e transitiva. Dada uma relação de equivalência R , ela define classes de equivalência. Uma *classe de equivalência* X , tal que $x \in X$, é uma parte de E constituída por todos os elementos y de E tais que xRy .

Seja R uma relação de equivalência sobre o conjunto E . *Conjunto quociente de E por R* , representado por E/R , é o conjunto cujos elementos são as classes induzidas por R sobre E . Uma equivalência, partição e quociente, sobre um conjunto E , podem ser obtidos a partir de uma aplicação (ou função) de E num conjunto A .

Seja $f : E \rightarrow A$, então f define:

- Uma relação de equivalência sobre E ,

$$xRy \Leftrightarrow f(x) = f(y) \quad (x \text{ "tem a mesma imagem por } f \text{ que " } y).$$

- Uma partição de E . Cada classe da partição é constituída pelos elementos de E que têm a mesma imagem por f .

- Um conjunto quociente E/f . Coincide com o conjunto das classes.

Para o desenvolvimento deste assunto sugere-se Barbut & Monjardet [6].

Seja π uma partição sobre E e R a relação de equivalência por ela definida. Represente-se por $gr(R)$ o grafo de R , que é uma parte de $E \times E$ e é definido por:

$$gr(R) = \{(x, y) : x \in E, y \in E \wedge xRy\}. \quad (2.4)$$

A relação de inclusão em $E \times E$ permite definir uma ordem, representada por $<$, sobre $P(E)$; assim $R < R'$ significa que $gr(R) \subset gr(R')$ ou de forma equivalente que $\forall x, y \in E : xRy \Rightarrow xR'y$. A partição π diz-se um *refinamento da partição π'* associada a R' , ou ainda que π é mais fina que π' .

O conjunto ordenado $P(E)$ está munido de uma estrutura de reticulado, isto é, todo o par de partições $\{\pi, \pi'\}$ admite um maior minorante comum, $\pi \wedge \pi'$, e um menor majorante comum $\pi \vee \pi'$. A partição $\pi \wedge \pi'$ pode ser definida pelo seu grafo:

$$gr(R \wedge R') = gr(R) \cap gr(R'), \quad (2.5)$$

ou ainda,

$$x(R \wedge R')y \text{ se e só se } x(R)y \text{ e } x(R')y. \quad (2.6)$$

Dado um conjunto E de elementos a classificar, existem procedimentos classificatórios para obtenção da melhor, em algum sentido e de acordo com o critério correspondente, partição de E em c classes, em que c pode ser ou não predefinido. Não é desta classe de métodos que trata este texto, mas sim de métodos de Classificação Hierárquica. Em Classificação Hierárquica obtém-se uma sucessão de partições encaixadas, permitindo conhecer a história da formação de cada uma das classes da partição escolhida, o que pode ser de grande utilidade em termos de interpretação.

Muitos têm sido os algoritmos propostos para construir classificações de forma automática. Para uma abordagem geral ver Lerman [48], Hartigan [35], Jain & Dubes [37] e Gordon [26].

2.2 CLASSIFICAÇÃO HIERÁRQUICA

Os diferentes métodos de Classificação Hierárquica podem ser agrupados em duas categorias; os métodos ascendentes ou aglomerativos, sem dúvida os mais usados, e os descendentes ou divisíveis.

Um método aglomerativo, parte das m classes formadas pelos elementos isolados do conjunto a classificar e em cada passo do algoritmo reúne as duas classes mais semelhantes, atribuindo um valor a esta agregação. Geralmente não há nenhum critério de paragem específico e o algoritmo procede até que os elementos de E estejam todos reunidos numa só classe. Note-se que, em cada passo as duas classes que se agregam são cada vez menos homogêneas, podendo até ser constituídas por elementos com comportamentos opostos. O resultado deste procedimento designa-se por Classificação Hierárquica Ascendente.

Um processo divisível consiste em dicotomias sucessivas do conjunto total até à obtenção dos elementos isolados que o constituem. A este procedimento associa-se uma Classificação Hierárquica Descendente.

Note-se que num algoritmo aglomerativo, as classes a agregar em cada etapa são escolhidas de acordo com um dado critério de optimização. Contudo, este facto não implica que a classificação hierárquica produzida, no seu todo, satisfaça um critério de optimização. Para alguns aspectos combinatórios em Classificação refere-se Saporta [65], onde algumas fórmulas e quadros alertam para a impossibilidade de obter soluções óptimas.

2.2.1 Hierarquia – Sua Representação

Dois conceitos muito úteis em Classificação Hierárquica são o de hierarquia e o de hierarquia indiciada ou árvore indexada. Uma classificação hierárquica produz uma hierarquia indiciada cuja representação gráfica mais frequente é o dendrograma¹. Apesar de dendrograma ser uma representação gráfica, no que pode ser visto como um abuso de linguagem, usar-se-á o termo dendrograma quer para a representação gráfica da hierarquia indiciada quer para a própria hierarquia.

Seja E um conjunto de m elementos, uma hierarquia de partições é um conjunto ou sucessão de partições encaixadas. Seja H uma família de partes não vazias de E , diz-se que H é uma hierarquia sobre E se:

- i. $E \in H$
 - ii. $\{x\} \in H, \forall x \in E$
 - iii. $\forall A, B \in H, A \cap B \in \{A, B, \emptyset\}$.
- (2.7)

Na Fig. 2.1 apresenta-se graficamente uma hierarquia e a respectiva sucessão de partições encaixadas.

Uma árvore indiciada ou dendrograma pode ser caracterizado de várias formas, Boorman & Olivier [14], Constantinescu [17], Hartigan [34], uma das mais correntes é defini-la pela associação de um valor h a cada nó interno (subconjunto). Assim para todos os subconjuntos com objectos em comum, quando a função de comparação é do tipo distância, verifica-se a seguinte relação:

$$h(A) \leq h(B) \Leftrightarrow A \subset B. \quad (2.8)$$

Define-se uma função h no conjunto dos pares de objectos de E , que a cada par (i, j) associa $h_{i,j} = h(i, j)$, como sendo o valor de h do menor subconjunto que contém simultaneamente os objectos i e j . Nestas condições o valor $h_{i,j}$ indica uma distância entre esses objectos.

Entre outras, uma propriedade desejável para os valores $h_{i,j}$ é satisfazerem a condição ultramétrica:

$$h_{i,j} \leq \max(h_{i,k}, h_{j,k}), \forall i, j, k \in E. \quad (2.9)$$

Os valores desta função h , associada a um dendrograma, devem ser usados com precaução já que eles dependem da medida de semelhança (dissemelhança) e do critério de agregação usados. A comparação de diferentes dendrogramas, obtidos a partir do mesmo conjunto de dados, pode ser no entanto, um dos objectivos do estudo. Para ultrapassar esta dificuldade pode recorrer-se à noção de *árvore ordenada* (ranked tree), na qual se retém como informação sobre os valores de h , apenas a sua relação de ordem, Boorman & Olivier [14], Murtagh [56]. Os valores de h serão designados por índices de nível e os valores ordinais correspondentes por níveis de agregação. Os valores ordinais estão representados no exemplo da Fig. 2.1.

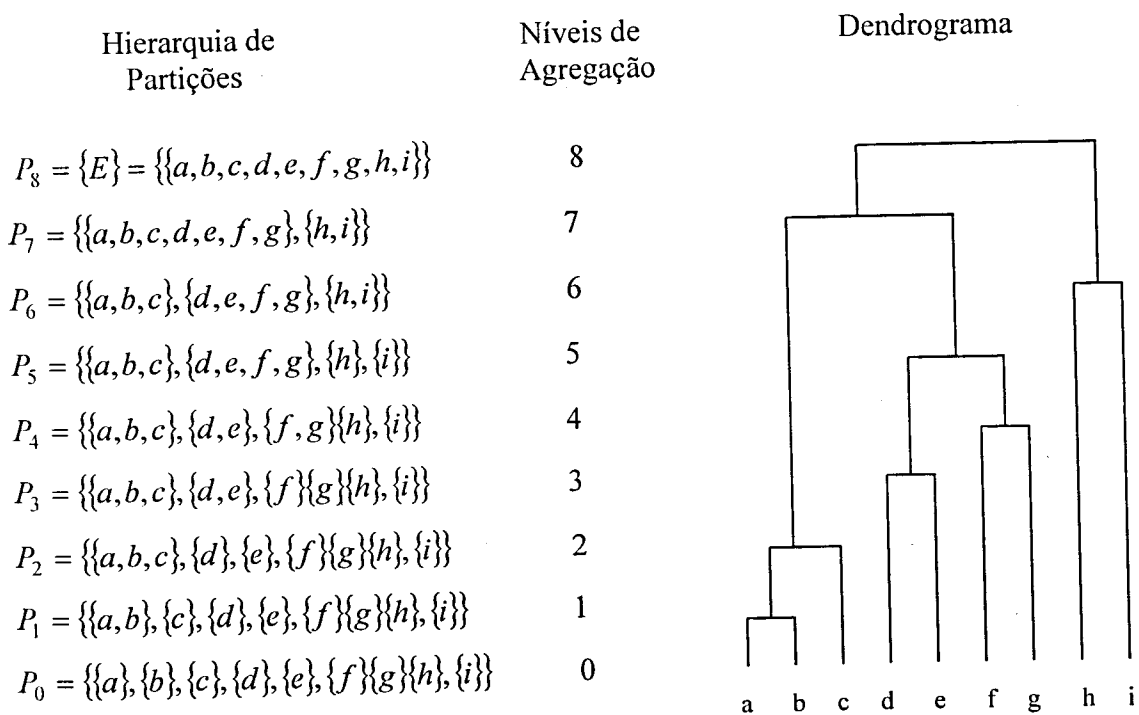


Fig. 2.1 – Dendrograma e respectiva hierarquia de partições

¹ Embora existam outras formas de representação gráfica que podem ser vistas na literatura deste tema.

Alternativamente pode definir-se um dendrograma como uma hierarquia indiciada de partições, isto é uma sucessão de partições cujas classes são parcialmente ordenadas por inclusão. Numa partição, cada classe corresponde a um nó no dendrograma.

Seja $P_0, P_1, \dots, P_i, P_{i+1}, \dots, P_k$ a sucessão de partições do conjunto E . De uma forma geral dizemos que, P_i é *mais fina* que P_{i+1} ou que P_i está encaixada em P_{i+1} , dado que para toda a classe, A , de P_i existe uma classe, B , de P_{i+1} tal que $A \subset B$. P_0 é a partição *mais fina*, tem um número de classes igual ao cardinal de E , já que cada classe é constituída por um único elemento. P_k é a partição *menos fina*, constituída por uma só classe, E .

A representação de uma hierarquia pode ainda ser feita à custa de uma matriz, designada por *matriz cofenética*, em que o elemento (i, j) é o valor $h_{i,j}$ já referido, ou o valor ordinal correspondente. É uma matriz quadrada, de dimensão m , simétrica e são suficientes $m(m-1)/2$ valores para a definir.

Matriz cofenética e dendrograma contêm a mesma informação. Na Fig. 2.2 representa-se esta matriz para o dendrograma da Fig. 2.1, usando os níveis de agregação.

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>b</i>	1							
<i>c</i>	2	2						
<i>d</i>	7	7	7					
<i>e</i>	7	7	7	3				
<i>f</i>	7	7	7	5	5			
<i>g</i>	7	7	7	5	5	4		
<i>h</i>	8	8	8	8	8	8	8	
<i>i</i>	8	8	8	8	8	8	8	6

Fig. 2.2 – Matriz cofenética associada ao dendrograma da Fig. 2.1

2.2.2 Classificação Hierárquica Ascendente

Na aplicação de um método de *C.H.A.* a um conjunto E , de cardinal m , estão implícitas duas escolhas:

- a função de comparação entre os elementos de E , γ , já definida em 2.1.3.1.
- a função de comparação entre partes de E , Γ , definida em 2.1.3.2.

De forma sucinta apresenta-se de seguida a sequência de passos num algoritmo de *C.H.A.*, considerando que as funções de comparação são do tipo distância:

- Condições iniciais – Definição prévia das duas funções de comparação e consequente determinação de $\{\gamma(x, y) : x, y \in E\}$.
- Etapa 0 – Definir a partição inicial ou mais fina, $P_0 = \{\{x\} : x \in E\}$
- Etapa i (para $i = 1, \dots, k$):
 - Determinar o par ou pares de classes de P_{i-1} que verificam o critério de agregação associado a Γ , isto é obter o par ou pares de classes que correspondem ao mínimo de $\{\Gamma(A, B) : A, B \in P_{i-1}\}$.
 - Actualizar a partição, $P_i = \{P_{i-1} \setminus (\{A\} \cup \{B\})\} \cup \{A \cup B\}$.
 - Actualizar os valores de Γ .

Na prática, a agregação de um número de classes superior a dois num dado nível, tem resolução simples. Contudo, por simplificação de escrita, admite-se que em cada etapa apenas um par de classes se agrega.

Num algoritmo de *C.H.A.* podem ser introduzidos critérios de paragem, por exemplo, impondo um número mínimo (superior a um) fixo de classes para a última partição, ou limitando o valor da função de comparação entre partes. Um procedimento habitual é deixar o

algoritmo actuar até obter a partição menos fina, coincidente com E . Nestas condições e atendendo ao referido no parágrafo anterior tem-se $k = m - 1$.

Como já foi referido existe um elevado número de funções de comparação entre elementos disponíveis. É bastante comum recorrer a funções do tipo distância, quando os elementos de E são indivíduos, e do tipo coeficiente de correlação ou de associação, quando os elementos de E são variáveis.

Também para o cálculo de semelhança ou dissemelhança entre duas classes, o número de opções é vasto, sendo habitual associar a cada possibilidade um critério de classificação. Em 2.2.2.1 e 2.2.2.2 referem-se alguns destes critérios.

2.2.2.1 Abordagem Clássica

Um dos critérios mais conhecidos é o da *Ligação Mínima*, com sigla *S.L.* (Single Linkage). Neste critério, sendo γ do tipo dissemelhança, a distância entre duas classes A e B é dada por:

$$\Gamma(A, B) = \min\{\gamma(a, b) : a \in A \wedge b \in B\} \quad (2.10)$$

Este método não depende dos valores numéricos de γ , no sentido em que é apenas função das suas ordens. Como consequência deste facto, funções de comparação entre elementos, equivalentes¹, produzem a mesma hierarquia.

O critério *S.L.* agrega classes pelos elementos mais próximos, o que dificulta a distinção de classes pouco separadas. As classes obtidas pelo *S.L.* exibem um fenómeno frequentemente inconveniente, conhecido por cadeia, pelo qual à medida que a agregação de classes tem lugar, há uma tendência para formar classes numerosas e alongadas. Esta propriedade faz com que o *S.L.* seja um bom método na recuperação de classes alongadas.

Pode afirmar-se que o *S.L.* dá ênfase ao isolamento das classes e não valoriza a sua homogeneidade ou coesão interna. Estudos feitos para este método, Jardine & Sibson [39], indicaram que ele é estável para qualquer conjunto de dados, no sentido de que uma pequena

¹ Aquelas que produzem a mesma preordenação.

perturbação nos valores γ , provoca pequenas alterações nos valores de distâncias ultramétricas finais.

Outro critério clássico é o da *Ligação Máxima*, denotado por *C.L.* (Complete Linkage), no qual a distância entre duas classes é dada pela distância dos dois elementos, um de cada classe, mais afastados, isto é:

$$\Gamma(A, B) = \max\{\gamma(a, b) : a \in A \wedge b \in B\}. \quad (2.11)$$

O critério *C.L.* assegura que a distância entre quaisquer dois elementos de uma nova classe, formada por união de duas já existentes, é não superior a essa distância nas classes que lhe deram origem.

Tal como no critério anterior o *C.L.* só depende das ordens dos valores numéricos de γ , produzindo a mesma hierarquia para funções de comparação entre elementos equivalentes.

As classes formadas com esta metodologia são em geral compactas, bastante coesas internamente, mas pouco isoladas externamente.

Outro critério que importa referir é o da *Ligação Média*, representado na literatura por *A.L.* (Average Linkage), que tem um desempenho intermédio entre o *S.L.* e o *C.L.*, no tipo de classes formadas. Neste critério a distância entre duas classes é a distância média dos pares de elementos. A definição dos pares de elementos não é única, mas neste trabalho considera-se

$$\Gamma(A, B) = \sum_{a \in A \wedge b \in B} \frac{\gamma(a, b)}{n_A n_B}, \quad (2.12)$$

em que n_A e n_B são os cardinais das classes A e B .

Para distâncias euclidianas importa referir um critério de agregação de natureza um pouco diferente, o *método de Ward*, Ward [71]. Os m elementos de E são considerados como uma nuvem de pontos num espaço $\mathfrak{R}^p$ e o critério de agregação é minimizar a variação de inércia interclasse. Em cada etapa reúnem-se as classes que conduzem a uma perda de inércia interclasses, ou a um aumento da inércia intraclases, mínimo.

Sejam g , g_A , g_B e g_{AB} os centros de gravidade global e das classes A , B e $A \cup B$, respectivamente, e n_A e n_B os cardinais de A e B . A variação de inércia interclasses¹ provocada pela reunião de A com B é dada por:

$$\begin{aligned}\Gamma(A, B) &= n_A d^2(g_A, g) + n_B d^2(g_B, g) - (n_A + n_B) d^2(g_{AB}, g) \\ &= \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B).\end{aligned}\quad (2.13)$$

Os algoritmos dos critérios de agregação referidos, e muitos outros, podem ser descritos por uma fórmula de recorrência, proposta por Lance & Williams [41], para coeficientes de dissemelhança:

$$\Gamma(A \cup B, C) = \alpha_1 \Gamma(A, C) + \alpha_2 \Gamma(B, C) + \beta \Gamma(A, B) + \gamma |\Gamma(A, C) - \Gamma(B, C)|, \quad (2.14)$$

e generalizada por Jambu & Lebeaux [38].

$$\begin{aligned}\Gamma(A \cup B, C) &= \alpha_1 \Gamma(A, C) + \alpha_2 \Gamma(B, C) + \beta \Gamma(A, B) + \gamma |\Gamma(A, C) - \Gamma(B, C)| + \\ &\quad + \delta_1 h(A) + \delta_2 h(B) + \delta_3 h(C),\end{aligned}\quad (2.15)$$

onde α_1 , α_2 , β , γ , δ_1 , δ_2 e δ_3 são constantes que assumem valores particulares, de acordo com o critério que se pretende usar, $A \cup B$ representa a nova classe, resultante da agregação das classes A e B , e h é a função definida em 2.2.1, que traduz o índice de agregação no nível de formação da classe.

A fórmula de recorrência (2.15), relativamente a (2.14), permite incluir alguns critérios definidos em espaços euclidianos, que envolvem minimização de momentos centrados de segunda ordem, como por exemplo o critério de variância mínima da união de duas classes.

¹ Dada pela média dos quadrados das distâncias dos centros de gravidade de cada classe ao centro de gravidade global.

O quadro 2.1 apresenta os valores das constantes da fórmula de Lance & Williams para os quatro critérios referidos.

Critério	α_1	α_2	β	γ
Ligação Mínima	1/2	1/2	0	-1/2
Ligação Máxima	1/2	1/2	0	1/2
Ligação Média	$\frac{n_A}{n_A + n_B}$	$\frac{n_B}{n_A + n_B}$	0	0
Método de Ward	$\frac{n_A + n_C}{n_A + n_B + n_C}$	$\frac{n_B + n_C}{n_A + n_B + n_C}$	$-\frac{n_C}{n_A + n_B + n_C}$	0

Quadro 2.1 – Valores das constante na fórmula de Lance & Williams para alguns critérios.

Jardine & Sibson [39] propuseram propriedades que os métodos de classificação devem verificar e mostraram que o critério da ligação mínima, era o único critério de classificação que verificava todas essas propriedades. Posteriormente, em diversos trabalhos, foi mostrado que tais propriedades eram demasiado restritivas e foram propostas propriedades alternativas desejáveis para os métodos classificatórios.

Uma propriedade desejável é a não existência de inversões, isto é, pretende-se que os métodos de agregação produzam classes que verifiquem (2.8). Existem contudo métodos que não verificam esta propriedade. Milligan [51] e Batagelj [9], propuseram condições necessárias e suficientes à não existência de inversões, para os critérios incluídos na fórmula de Lance & Williams:

$$\text{i. } \alpha_1 + \alpha_2 \geq 0 \quad (2.16)$$

$$\text{ii. } \alpha_1 + \alpha_2 + \beta \geq 1 \quad (2.17)$$

$$\text{iii. } \gamma \geq -\min(\alpha_1, \alpha_2) \quad (2.18)$$

Outras características dos métodos de C.H.A., têm origem no conceito de distorção do espaço (space distortion) introduzido por Lance & Williams [42], e recorrem às noções de conservação (space-conserving), contractividade (space-contracting) e dilatação (space-dilating). Para definições formais, ou referências às mesmas, ver Chen & Van Ness [16].

Dado um algoritmo de *C.H.A.*, considere-se uma etapa em que as classes A e B se reúnem, uma outra classe C e a consequente actualização da função de comparação, do tipo distância. Admita-se ainda, sem perda de generalidade, que $\Gamma(A, C) < \Gamma(B, C)$. Então esse procedimento diz-se:

□ Conservativo se

$$\Gamma(A, C) < \Gamma(A \cup B, C) < \Gamma(B, C) \quad (2.19)$$

□ Contractivo se

$$\Gamma(A, C) \geq \Gamma(A \cup B, C) < \Gamma(B, C) \quad (2.20)$$

□ Dilatador se

$$\Gamma(A, C) < \Gamma(A \cup B, C) \geq \Gamma(B, C) \quad (2.21)$$

Em Chen & Van Ness [16] são apresentadas condições necessárias e suficientes, para que um algoritmo, pertencente à família de Lance & Williams, possua estas propriedades. Ainda neste trabalho e em Van Ness [69] é abordado o conceito de admissibilidade de um algoritmo, numa tentativa de eliminar alguns procedimentos considerados maus. Esta abordagem não dá contudo indicações para o melhor algoritmo.

2.2.2.2 Abordagem VL - Validade da Ligação

Dado um conjunto E , a classificar, de cardinal m , seja c uma aplicação de $E \times E$ em $\mathfrak{R}$, escolhida para medida de comparação entre elementos. Os $m(m-1)/2$ valores do conjunto $\{c(x, y) : x \neq y, x, y \in E\}$, podem ser considerados realizações de uma variável aleatória Z , em que Z representa o valor da medida de comparação entre elementos, cuja lei é definida por uma hipótese base de referência, em regra governada pelo princípio de Occam.

Nicolau [57] e Bacelar-Nicolau [2] propõem para função de comparação entre elementos x e y , uma função γ dada por:

i. $\gamma(x, y) = P(Z \leq c(x, y)) = F_Z(c(x, y))$, se c for uma medida de semelhança

ii. $\gamma(x, y) = P(Z \geq c(x, y)) = 1 - F_Z(c(x, y))$, se c for uma medida de dissemelhança

Admitindo, como hipótese natural ou de referência, independência entre os elementos de E , aceita-se que o conjunto $\{\gamma(x, y) | x \neq y, x, y \in E\}$ constitui uma amostra de $m(m-1)/2$ variáveis aleatórias independentes e identicamente distribuídas da lei uniforme no intervalo $[0,1]$. Trata-se de uma generalização da abordagem pioneira de Lerman [48].

Quanto à escolha da função de comparação entre partes de E , admite-se que, dadas duas classes A e B de E , respectivamente de cardinais n_A e n_B , toda a informação útil para decidir da sua semelhança/dissemelhança está contida no conjunto $\{\gamma(a, b) | (a, b) \in E \times E\}$.

Para função de comparação entre partes Lerman propôs usar, o valor da função de distribuição do valor observado do máximo das semelhanças, isto é:

$$\Gamma_{AVL}(A, B) = P(\gamma_{Max} \leq \gamma_{Max}(A, B)), \quad (2.22)$$

cujas distribuições Lerman deduzia em duas etapas utilizando um raciocínio combinatório.

Facilmente se conclui que neste método, baptizado pelo autor de *AVL - Agregação pela Verosimilhança da Ligação* (em francês *Vraisemblance du Lieu*), a função de comparação entre classes, é dada por:

$$\Gamma_{AVL}(A, B) = (\gamma_{Max}(A, B))^{n_A n_B}. \quad (2.23)$$

Este método foi exaustivamente estudado por Bacelar - Nicolau [2] e tem vindo desde então a ser frequentemente utilizado, directamente ou através da família paramétrica de métodos que lhe está associada, Nicolau [57] e Bacelar - Nicolau & Nicolau [4]. Estes autores chamam a esta abordagem, por ser imprópria a designação de Verosimilhança, a metodologia *VL - Validade da Ligação*.

Na referida família paramétrica de métodos a função de comparação entre classes é dada por

$\Gamma(A, B) = (\gamma_{Max}(A, B))^{1+\varepsilon[(n_A n_B)^\xi - 1]}$ e dela destaca-se o critério *AVB*, que se obtém para $\varepsilon = 1$ e $\xi = 0.5$, e que será usado no capítulo 6.

A passagem, das estatísticas tradicionais de comparação à semelhança *VL*, tem vantagens para as quais se sugere Nicolau [59]. De entre os vários pontos aí referidos salienta-se o carácter de uniformização da escala de medida e a simplificação da escolha da função de comparação entre elementos no caso de dados binários, consequência da equivalência distribucional desses coeficientes, Bacelar-Nicolau [3].

Posteriormente Nicolau [59] e [60] propôs-se estender esse princípio a outras estatísticas da amostra (mínimo, média aritmética entre classes, média aritmética da nova classe e média geométrica), daí resultando em particular o método *AVM*, no qual a semelhança entre duas classes *A* e *B* é calculada recorrendo ao valor da função de distribuição da estatística média aritmética entre classes.

Em Nicolau [60] é proposta uma fórmula de recorrência alargada, generalização da fórmula de Lance & Williams, que inclui a família de métodos *AVL*:

$$\begin{aligned} \Gamma(A \cup B, C) = & \alpha_1 \Gamma(A, C)^{g(A, C; \varepsilon, \xi)} + \alpha_2 \Gamma(B, C)^{g(B, C; \varepsilon, \xi)} + \beta \Gamma(A, B)^{g(A, B; \varepsilon, \xi)} + \\ & + \gamma \left| \Gamma(A, C)^{g(A, C; \varepsilon, \xi)} - \Gamma(B, C)^{g(B, C; \varepsilon, \xi)} \right|^{1/g(A, B; \varepsilon, \xi)} \end{aligned} \quad (2.24)$$

onde $g(A, C; \varepsilon, \xi) = \frac{1}{1 + \varepsilon((n_A \times n_B)^\xi - 1)}$, e ε e ξ tomam valores no intervalo $[0, 1]$.

Facilmente se verifica que, para $\varepsilon = 0$ (2.24) coincide com (2.14), e que vários critérios de metodologia *VL* podem ser obtidos atribuindo valores particulares às constantes.

As estatísticas acima citadas não são obviamente as únicas possíveis para esta metodologia *VL*. Outras estatísticas, como a mediana e o centro, serão estudados no próximo capítulo, constituindo novas boas opções dentro desta abordagem.

De notar que as preocupações ligadas á visão métrica clássica como a conservação, contractividade e dilatação, não são aplicadas nesta abordagem. Em particular as inversões são possíveis e podem ser usadas positivamente na ajuda à interpretação de uma árvore. Quando o número de elementos a classificar é elevado, a representação gráfica da árvore obtida é de difícil leitura e a eliminação dos níveis em que se registam inversões pode ser um processo de redução da mesma.

2.3 ÁRVORES DE CLASSIFICAÇÃO

2.3.1 Conceitos Gerais

Dada uma estrutura de árvore sobre um conjunto de elementos, existem várias maneiras de a representar, no entanto a representação por recurso a um grafo é a que melhor ilustra as relações de ramificação existentes. Este aspecto gráfico é o responsável pela adopção do nome árvore para este tipo de estrutura.

A estrutura de árvore tem um papel importante em áreas científicas muito diversas, não sendo portanto de estranhar as diferentes terminologias e definições que surgem na literatura em torno deste tema. Das várias características de uma estrutura de árvore, pode referir-se o facto de este tipo de estrutura ser facilmente definida recursivamente, o que lhe confere particular importância em computação.

De um modo bastante geral define-se uma árvore como um grafo conexo sem ciclos. Isto significa que, para dois elementos quaisquer de um conjunto de objectos representado pelos nós, ou vértices, da árvore, há exactamente um caminho, através dos ramos da árvore, que os liga.

Dada uma árvore define-se:

- *Grau de um nó da árvore* como o número de ramos ligados a ele:
 - . Um *nó terminal* ou *folha* tem sempre grau um.
 - . Um *nó interior* tem sempre grau superior a um.

- *Árvore com raiz*, quando, para introduzir a direcção nos ramos da árvore, um dos nós é classificado de raiz.

- Quanto ao grau dos nós interiores:
 - . *Árvore binária*, se nenhum nó interior tem grau superior a três.
 - . *Árvore completamente binária*¹, se todos os nós interiores têm grau igual a três.

- Quanto às etiquetas dos nós tem-se:
 - . *Árvore C - etiquetada*, se a todos os seus nós são atribuídas etiquetas.
 - . *Árvore T - etiquetada*, se apenas os nós terminais são etiquetados.
 - . *Árvore N - etiquetada*, se nenhum tipo de nó é etiquetado.

- *Árvore ponderada* se existir atribuição de valores aos vários ramos da árvore, permitindo associar uma função definida sobre o conjunto dos pares de nós. Um exemplo de árvores ponderadas são as árvores aditivas ou árvores com comprimento de caminho².

Neste estudo há um tipo particular de árvores, *os dendrogramas*, com grande interesse, pelo facto de serem a representação gráfica mais frequente, do resultado de um método Classificatório Hierárquico, como já foi referido no ponto 2.2.1.

Um dendrograma, é uma árvore com raiz, com grupos não sobrepostos, com níveis de fusão organizados num caminho hierárquico. Pode ainda afirmar-se que um dendrograma é uma árvore hierárquica ponderada em que todos os caminhos das folhas à raiz têm o mesmo comprimento.

Frequentemente os dendrogramas são referidos como árvores ultramétricas. Neste tipo de árvore os nós internos são ordenados com base na distância relativa à raiz. Os valores de níveis de fusão podem substituir as ordens dos índices de nível dos nós internos dos dendrogramas. Este último procedimento é no entanto frequentemente evitado ou criticado

¹ Neste trabalho, por simplificação, ao referir-se o termo árvore binária admi-se que ela é completamente binária.

² Nas quais se associa a qualquer par de nós o comprimento do caminho que os une, dado pela soma dos pesos dos ramos que constituem esse caminho.

por vários investigadores, que apelam para o facto dos valores dos níveis de fusão dependerem de um conjunto de decisões, já referidas aquando da exposição do algoritmo genérico de uma *C.H.A.*, podendo afectar a sua interpretação, Gordon [27]. Num dendrograma a raiz tem sempre o maior valor de fusão, ou maior ordem, dado que representa o nó interno mais afastado das folhas.

Dada uma árvore, um aspecto que interessa também referir, é a sua característica relativamente à ordenação dos ramos de cada nó, isto é, o facto da posição relativa dos descendentes¹ de um nó ser ou não importante. No caso particular dos dendrogramas são possíveis várias representações diferentes para um mesmo dendrograma.

Define-se um par de dendrogramas como isomorfos, se eles diferem apenas na ordem dos braços dos seus nós. Dois dendrogramas, sobre um mesmo conjunto, serão distinguíveis se não forem isomorfos, e serão coincidentes se forem isomorfos.

A representação de um dendrograma tem um elevado grau de indeterminação, já que a ordenação esquerda-direita dos braços de cada nó interno pode mudar. Assim há 2^{n-1} maneiras diferentes de representar um mesmo dendrograma binário de m objectos.

A descrição completa de um dendrograma passa por três propriedades formais:

- Identificação da topologia da árvore e da posição da sua raiz.
- Fornecimento da escala, na qual os níveis de fusão estão representados hierarquicamente.
- A posição das unidades estatísticas nas folhas terminais da árvore deve ser dada, isto é, a árvore deve ser *T - etiquetada*.

Sibson [68] distinguiu duas categorias de dendrogramas: *invariantes de ordem local* e *invariantes de ordem global*. No primeiro tipo a informação retida sobre um dendrograma é apenas a topologia (tipo) e a atribuição de etiquetas às folhas, o que corresponde à definição dada atrás de árvore com raiz, *T - etiquetada*, não ponderada. Num dendrograma invariante de ordem global as classes na hierarquia, ou nós dos dendrogramas, têm uma única ordem ou

¹ Um nó y , que está directamente abaixo do nó x , na árvore, denomina-se descendente de x .

valor de nível associado, isto é, as distâncias relativas entre os nós internos e a raiz são completamente ordenadas.

Como já foi referido, num dendrograma pode substituir-se a ordenação dos nós pelos valores de nível de fusão associados aos nós internos, enriquecendo a informação fornecida pelo gráfico, apesar da incerteza quanto à qualidade dessa informação. A este tipo de dendrograma Lapointe & Legendre [45] deram o nome de dendrogramas completamente ponderados.

Neste trabalho reserva-se o termo dendrograma para a categoria de dendrogramas invariantes de ordem global, não fazendo contudo distinção de terminologia quando se retiverem apenas as ordens de formação de um nó (dendrograma invariante de ordem global), ou quando se retiverem os valores de nível de fusão (dendrograma completamente ponderado).

Ao longo deste trabalho, pela sua importância e por simplificação de apresentação, dá-se particular atenção às árvores binárias. Num dendrograma binário o número de sucessores de qualquer nó não terminal é constante, igual a dois. Assim na construção de um dendrograma, usando um algoritmo do tipo aglomerativo, a sucessão de reuniões de classes e/ou objectos combina exactamente dois destes elementos em cada nível. Pode concluir-se que, num dendrograma binário, de ordem invariante global, com m nós terminais, há precisamente $m-1$ níveis ou ordens associados.

A enumeração exaustiva do conjunto de árvores não isomórficas de um determinado tipo, ou pelo menos o seu cardinal, é um problema importante que tem aplicações em diferentes áreas. Em Murtagh [56] e em Gordon [27] é feita uma síntese dos resultados sobre o número, ou relações recursivas para o determinar, de árvores não isomórficas para diferentes categorias de árvores.

Neste contexto refira-se o trabalho de Felsenstein [22], onde se mostra que o número de árvores binárias de ordem m , $m \geq 2$, com raiz, não ponderadas e *T - etiquetadas* é dado por:

$$a(m) = \frac{(2m-3)!}{2^{m-2}(m-2)!}. \quad (2.25)$$

Por outro lado Frank & Svensson [24] mostraram que, quando não há empates nos níveis de fusão, o número de dendrogramas binários de ordem m distinguíveis é dado por:

$$d(m) = \frac{m!(m-1)!}{2^{m-1}}. \quad (2.26)$$

Em Murtagh [56] é referida uma fórmula de recorrência para determinar o número de árvores binárias, não *T* - *etiquetadas* e não ponderadas, que coincide com o número de tipos topológicos distintos, num dendrograma de ordem m . Essa fórmula difere para m par e ímpar:

$$t(m) = \begin{cases} \sum_{k=1}^{m/2-1} t(k)t(m-k) + \frac{1}{2}t(m/2)(t(m/2)+1) & \text{se } m \text{ par} \\ \sum_{k=1}^{(m-1)/2} t(k)t(m-k) & \text{se } m \text{ ímpar} \end{cases} \quad (2.27)$$

$$e \quad t(1) = t(2) = 1.$$

No quadro 2.2 apresenta-se o número de árvores não isomórficas, para as três situações aqui consideradas e para alguns valores de m .

m	$t(m)$	$a(m)$	$d(m)$
1	1	1	1
2	1	1	1
3	1	3	3
4	2	15	18
5	3	105	180
6	6	945	2700
7	11	10395	56700
8	23	135135	1587600
9	46	2027025	57153600
10	98	34459425	2571912000
15	4850	2.13×10^{14}	5.14×10^{18}
20	293547	8.21×10^{21}	1.53×10^{29}
30	1.41×10^9	4.95×10^{38}	4.58×10^{52}
40	8.10×10^{12}	1.01×10^{57}	4.83×10^{78}
50	5.15×10^{16}	2.75×10^{76}	$> 10^{100}$

Quadro 2.2 – Número de árvores não isomórficas de três categorias diferentes

O quadro anterior mostra bem que a rapidez de crescimento detectada nas primeiras linhas, é dramaticamente acentuada para os últimos valores de m considerados. De facto, fixada uma categoria de árvore, verifica-se que o cardinal do conjunto de árvores não isomórficas aumenta rapidamente com m , conduzindo a alguns problemas, quando se pretende determinar a distribuição, de uma dada categoria de árvores com elevado número de objectos.

Em muitas situações a enumeração exaustiva pode ser, necessariamente, substituída por amostragem aleatória. A geração aleatória de árvores é o assunto da próxima sub-secção.

2.3.2 Geração Aleatória de Árvores

A geração aleatória de árvores é um instrumento de grande utilidade em Classificação Hierárquica, nomeadamente para:

- ❑ A comparação empírica de diferentes algoritmos
- ❑ A validação de resultados
- ❑ Estudos de sensibilidade / robustez.

Foi já referido que o termo árvore engloba categorias bem distintas que, por sua vez, se associam a metodologias bem diferentes de geração:

- ❑ Árvores com nós terminais etiquetados ou não
- ❑ Árvores com nós interiores etiquetados ou não
- ❑ Árvores com ou sem raiz
- ❑ Árvores binárias ou não
- ❑ Árvores ponderadas ou não.

Várias têm sido as contribuições para o desenvolvimento deste tema. Em seguida referem-se alguns trabalhos mais relevantes, ou com interesse para o desenvolvimento do capítulo 4. Quanto à geração ou enumeração de árvores sem raiz, referem-se apenas os trabalhos de Knuth [40], por ser um tipo de árvore não importante para o desenvolvimento deste trabalho. Para referências mais detalhadas neste assunto refere-se Murtagh [56], Barthélemy & Guénoche [8] e Lapointe & Legendre [45].

Quanto à geração de árvores com raiz, binárias ou não, com os nós terminais etiquetados e não ponderadas refiram-se os trabalhos de Felsenstein [22], associando o nome de árvores evolutivas, Leclerc [47], utilizando o termo hierarquia de partes e Oden & Shao [61]. Neste último é apresentado um algoritmo para gerar equiprovavelmente todas as árvores, fixado o número de nós terminais.

Em De Soete *et al.* [20] é apresentado um algoritmo para gerar árvores aleatórias *T* - etiquetadas, com parâmetro de forma. As árvores são geradas em dois passos: a geração de formas topológicas e a atribuição de comprimentos aos ramos. Contudo, os modelos de árvores aqui abordados são mais gerais do que as árvores obtidas numa *C.H.A.*

Murtagh [55] considera a geração de árvores binárias ponderadas e não etiquetadas, dando-lhe o nome de dendrogramas. Considera a noção de *estrutura de um dendrograma* e define um *coeficiente de estrutura* que mede a "simetria" ou "balanço" dessa classificação hierárquica. Esse coeficiente assume valores entre 0 e 1, correspondendo, respectivamente, a um dendrograma equilibrado (tipo *C.L.*) e a um dendrograma o mais assimétrico possível, a cadeia pura (tipo *S.L.*). Sob a hipótese de que todos os dendrogramas (não etiquetados) distintos são equiprováveis, determina a distribuição estatística desse coeficiente de estrutura, por enumeração exaustiva para *m* pequeno ($m < 12$) e por simulação para *m* grande.

Para a geração de árvores não ponderadas, referem-se os trabalhos de Furnas [25], Shao & Sokal [67] e Quiroz [62].

Em Furnas [25] são discutidos métodos e apresentados algoritmos para a geração de seis categorias de árvores binárias não ponderadas. As árvores aqui consideradas são definidas pela sua forma topológica e pela sua etiquetagem mas não pelos valores de níveis de fusão, isto é são apenas invariantes de ordem local. As seis categorias de árvores consideradas são com e sem raiz, e dentro de cada categoria as *C* - etiquetadas, *T* - etiquetadas e *N* - etiquetadas. Os vários procedimentos de geração de árvores são uniformes, isto é, qualquer árvore possível tem igual probabilidade de ser gerada. Em termos gerais, estes procedimentos incluem a determinação do número de árvores da categoria pretendida, seja *N*, a ordenação conveniente desses *N* elementos, e posterior selecção da árvore correspondente ao número inteiro gerado uniforme e aleatoriamente do conjunto dos inteiros $\{1, 2, \dots, N\}$. Este procedimento é referido em vários trabalhos, como sendo geração aleatória uniforme de árvores no sentido de Furnas.

A geração de árvores, não necessariamente binárias, *T* - *etiquetadas* e não ponderadas, surge em Shao & Sokal [67], para o estudo da distribuição amostral de vários índices consensuais associados a diferentes metodologias de construção de árvores de consenso.

Quiroz [62] apresenta algoritmos de geração aleatória de árvores “*n*-árias”¹, *C* - *etiquetadas* e *T* - *etiquetadas*, mas não ponderadas. A geração aleatória de árvores é uniforme, no sentido de Furnas.

De seguida apresenta-se, com algum detalhe, a geração de dendrogramas introduzida por Lapointe & Legendre, que será usada para comparação com os processos de geração de árvores propostos no capítulo 4. O trabalho destes autores, pela sua relevância, recebeu o prémio da *IFCS* – International Federation of Classification Societies de 1996.

A geração aleatória de árvores realizada por este procedimento tem por base o algoritmo de permutação dupla de Lapointe & Legendre [44] e [45]. Os autores referem esta metodologia como uniforme no sentido de Furnas, quer isto dizer que a geração é uniforme no conjunto dos dendrogramas distintos que é possível definir, para um número fixo de nós terminais. Este facto, faz com que as diferentes topologias que resultam do processo aleatório, surjam em proporção ao número de árvores distinguíveis com a mesma topologia.

A geração aleatória de dendrogramas implica, além da geração de topologias aleatórias, a geração aleatória de níveis de fusão. Os valores dos níveis de fusão, devem ser gerados aleatoriamente, de acordo com a distribuição dos coeficientes de semelhança usados na construção dos respectivos dendrogramas. A existência de uma grande variedade de coeficientes de semelhança para obtenção de dendrogramas, dá origem a um leque diversificado de distribuições a considerar para a amostragem de níveis de fusão. Este aspecto pode complicar-se bastante. Uma maneira de evitar este tipo de problema é substituir os níveis de fusão pelas suas ordens.

Em 2.2.1, definiu-se matriz cofenética e referiu-se que esta e dendrograma contêm a mesma informação. A geração de dendrogramas aleatórios apresentada por esses autores, tem por base esta equivalência de informação, a representação compacta (*packed representation*) da topologia de uma árvore e o procedimento de *Permutação Dupla*. A noção de representação compacta foi introduzida por Murtagh [56], mostrando que *m-1* valores, correspondendo aos

¹ Generalização do termo binária para valores de *n* diferentes de dois.

níveis de fusão, são necessários e suficientes para definir a topologia de uma árvore binária de m nós terminais. Na representação matricial da hierarquia, estes valores situam-se na segunda diagonal principal.

Esse procedimento de geração aleatória de uma árvore de dimensão m , começa por definir um vector de $m-1$ níveis de fusão, permutando-o aleatoriamente para gerar uma topologia. Aplicando a propriedade ultramétrica aos elementos desse vector determina a matriz cofenética correspondente, obtendo assim uma árvore ponderada, com topologia aleatória. Uma permutação do vector que contém as etiquetas dos nós terminais faz dessa árvore um dendrograma aleatório.

Poucos trabalhos têm surgido associados a geração não uniforme de árvores. Este aspecto é contudo importante em Classificação Hierárquica, já que são bem conhecidas as tendências de métodos classificatórios para produzir tipos particulares de árvores. Dos métodos de geração de dendrogramas propostos no capítulo 4, um deles atende a este aspecto.

2.4 VALIDAÇÃO EM CLASSIFICAÇÃO

2.4.1 Razões para Validar uma Classificação

Foi já referida a existência de um elevado número de algoritmos de *C.H.A.*, e que este depende da escolha de duas funções de comparação¹. A aplicação de um qualquer desses procedimentos conduz sempre a uma hierarquia², que pode variar com o método.

Estes aspectos levantam várias questões, algumas de índole prática que se colocam ao simples utilizador, e outras que se colocam ao estatístico responsável pela qualidade do produto final, de carácter mais teórico.

A possibilidade de poder aplicar vários critérios de classificação a um conjunto de dados, levanta, naturalmente, a questão de identificar o melhor (se existir). O próprio conceito de “boa árvore” não é pacífico, pois trata-se de uma noção relativa. Escolhida uma hierarquia,

¹ O resultado depende ainda de outros factores como a escolha dos indivíduos e das variáveis intervenientes e da normalização das variáveis, Milligan [54].

² Lembra-se que, a aplicação de um algoritmo de *C.H.A.* a uma matriz de dados gerada aleatória e uniformemente, portanto sem estrutura, conduz a uma hierarquia.

identificar a melhor partição do conjunto dado, ou o melhor número de classes pode ser uma tarefa delicada.

Perante tais factos, parece natural questionar se a estrutura hierárquica obtida estará de acordo com a estrutura dos dados, ou até pôr em dúvida a existência de estrutura nos dados. Dado um método de *C.H.A.* parece ser importante medir a sua eficiência e a sua estabilidade, ou ainda o valor real de cada uma das classes obtidas.

Estas e outras questões têm, nas últimas décadas, levado estatísticos a reconhecer a importância do tema “Validação em Classificação”. No que se segue apresentam-se algumas definições e referem-se algumas contribuições ou pontos de vista relevantes para este assunto.

2.4.2 Diferentes Abordagens

A abordagem a este tema pode ser feita com diferentes níveis de formalização, dependendo da natureza dos dados e da investigação pretendida.

Numa abordagem mais formal, assume-se que os dados observados constituem uma amostra de uma população multivariada estruturada. Este facto insere os modelos classificatórios num contexto probabilístico, o que permite usar ferramentas da estatística inferencial. Alguns testes para a ausência de estrutura nos dados, a adequação da classificação obtida aos dados, ou para a determinação do melhor número de classes são possíveis. Para uma perspectiva geral desta abordagem referem-se os trabalhos de Bock [12] e [13].

Uma abordagem menos formal recorre a instrumentos empíricos, descritivos ou exploratórios para analisar o valor da classificação obtida. O número de trabalhos inseridos nesta abordagem é bastante superior ao da abordagem anterior.

No tema “Validação em Classificação” podemos considerar essencialmente dois tipos de problemas:

- Obtida uma classificação, saber da sua qualidade e fiabilidade.

- Obtidas várias classificações decidir por uma dessas classificações ou por uma nova, determinada a partir dessas.

O segundo tipo de problema inclui a comparação de árvores de classificação e a teoria de consenso. Uma prática comum em Classificação é aplicar vários critérios ao mesmo conjunto de dados e em seguida comparar as estruturas obtidas. Quanto maior for a concordância entre elas mais elas são consideradas válidas. Este procedimento, não permitindo avaliar, de uma forma precisa, a qualidade da estrutura obtida, é bastante intuitivo e está relacionado com os problemas aqui referidos.

Dadas várias classificações sobre o mesmo conjunto de dados, usando métodos diferentes ou usando variáveis diferentes, a teoria de consenso define uma única e nova classificação, dita consensual das iniciais, que capta os aspectos estruturais comuns às várias classificações a comparar. A quantificação da estrutura comum, é feita por um coeficiente, denominado índice de consenso, que pode ser usado como uma medida de semelhança entre essas classificações.

Vários trabalhos têm surgido propondo metodologias de construção de árvores consensuais, Shao & Sokal [67], Day [19], Lapointe & Cucumel [43], e índices para medir o consenso entre estruturas classificatórias, Rohlf [63], Faith & Belbin [21], Shao & Rohlf, [66], Shao & Sokal [67]. Para uma discussão sobre este assunto refere-se Barthélemy et al [7] e Gordon [32].

Para a comparação de dendrogramas¹ obtidos independentemente² referem-se os trabalhos de Lapointe & Legendre [44] e [46]. O método proposto, assume o aspecto de um teste estatístico, em que a hipótese nula exprime que os dois dendrogramas em comparação não são mais semelhantes do que dois dendrogramas quaisquer gerados aleatoriamente. A semelhança entre dendrogramas é calculada utilizando uma forma normalizada do Índice de Consenso Intermédio de Faith & Belbin, Faith & Belbin [21].

Dos vários aspectos referidos sobre validação, interessa desenvolver as noções que, dada uma classificação hierárquica, permitem avaliar a sua qualidade e fiabilidade. Também aqui a validação assume diferentes vertentes, de acordo com o objectivo em causa:

¹ Esta comparação tem em conta os valores dos níveis de fusão entre os elementos.

² Este método de testar a semelhança de duas classificações, só pode ser usado quando as duas classificações a comparar estão associadas a diferentes conjuntos de dados. A aplicação de métodos distintos a um mesmo conjunto de dados, dá origem a dendrogramas não independentes.

- Validar uma classe individual
- Validar uma partição
- Validar uma hierarquia.

Pode considerar-se, Jain & Dubes [37], que em validação classificatória estão disponíveis três tipos de medidas, ou critérios:

- Os critérios de validação externa, quando existe informação útil à priori sobre os dados. Estes critérios medem o grau de concordância entre a informação obtida e a informação prévia.
- Os critérios de validação interna, quando não existe informação à priori, ou existe e não é usada. Estes critérios medem o afastamento entre os dados e a estrutura obtida por um algoritmo classificatório.
- Os critérios de validação relativa, que comparam várias classificações sobre o mesmo conjunto de elementos.

Cada nó interno de um diagrama em árvore define uma classe de objectos. Sendo um objectivo importante em classificação a substituição de um dendrograma por uma árvore reduzida¹, é importante que esta seja constituída pelos nós internos que correspondam a verdadeiras classes.

Para a validação de uma classe de uma classificação são habitualmente referidas três noções diferentes: a homogeneidade, o isolamento e a estabilidade. É difícil dar um modelo geral, para validar uma classe, que seja aplicável a todos os métodos classificatórios, porque diferentes métodos podem produzir classes com características bem diferentes. No entanto é natural aceitar uma classe como válida se ela apresenta uma boa homogeneidade e um bom isolamento das outras classes da partição.

Para a validação de uma classe referem-se os trabalhos de Gordon [29] e de Bel Mufti [10].

¹ Quando o número de elementos classificados é grande, é útil construir uma árvore, que contenha apenas os níveis correspondentes à formação das classes mais significativas. Esta árvore tem o nome de árvore reduzida.

Gordon [29] propõe uma metodologia de validação de uma classe, tendo por base um modelo nulo¹ para os dados e a estatística U de Mann & Whitney. Esta metodologia visa decidir se uma classe resultante de um procedimento de classificação tem propriedades significativamente diferentes das obtidas a partir do modelo nulo.

Bel Mufti [10] propõe uma metodologia de validação de uma classe gerada por um método de partição, propondo novos índices de validade que se baseiam na avaliação da estabilidade de uma classe por reamostragem.

Para um resumo dos trabalhos sobre validação de partições ver Hubert [36], capítulo 7, e Gordon [32].

Quanto à validade de classificações hierárquicas, o número de trabalhos é bastante inferior, dada a sua complexidade. Um resumo pode ser visto em Gordon [32]. De notar que a validação de uma hierarquia aparece, por vezes, misturada com a validação de partições, pelo facto de uma hierarquia ser uma sucessão de partições.

Muitos dos trabalhos nesta área recorrem à noção de estabilidade. Em seguida apresentam-se algumas utilizações da noção de estabilidade, bem como alguns trabalhos que a usaram.

Uma maneira de analisar a estabilidade de uma classificação é:

- Comparar a classificação original com a classificação obtida de uma versão modificada dos dados. Os dados podem ser modificados por:
 - . Adição de termo ou erro à matriz de dados crua.
 - . Adição de termo ou erro à matriz de dissemelhanças
 - . Apagar um ou mais objectos da matriz de dados inicial
 - . Acrescentar novos objectos ou novas variáveis
 - . Omissão de uma classe
 - . Construir uma nova matriz com os mesmos objectos e um conjunto de variáveis completamente novas.

¹ Para a definição dos vários modelos nulos ver Bock [13], Gordon [32] ou Bel Mufti [10].

A estabilidade da classificação mede-se à custa do desvio¹ entre a hierarquia obtida a partir dos dados iniciais e a hierarquia dos dados perturbados. Quanto menor for esse desvio mais estável é a classificação inicial.

- Considerar uma partição em duas partes do conjunto de objectos, e comparar a classificação de cada subconjunto com a classificação do conjunto todo.
- Alguns trabalhos vão no sentido de identificar os nós ou classes importantes em termos de estabilidade, usando critérios baseados em homogeneidade interna e ou isolamento externo das classes; ou obtendo um dendrograma com um número limitado de nós internos, usando directamente um algoritmo de optimização ou simplificando um dendrograma através de uma classificação com restrições ao conjunto de nós internos.
- Alguns factores que podem afectar a estabilidade interna de uma classificação são a existência de erros de medida, a má escolha dos coeficientes, a codificação preliminar de algumas variáveis e a existência de “observações aberrantes”.

Nesta linha de procedimento pode referir-se alguns trabalhos:

- Em Baker [5] é feito um estudo de sensibilidade de dendrogramas produzidos pelos critérios *SL* e *CL*. Partindo de três tipos de estruturas, binária, cadeia e aleatória, foram introduzidas perturbações e comparadas as hierarquias antes e depois de perturbação, usando o coeficiente de Goodman & Kruskal².
- Em vários trabalhos Milligan usou a noção de estabilidade. Em Milligan [52] foram realizados estudos de sensibilidade sobre vários métodos classificatórios. Partindo de classes artificiais coesas internamente e bem separadas ou isoladas externamente, introduziu sobre elas vários tipos de perturbações, procedendo depois à respectiva comparação. Concluiu que os métodos têm sensibilidades diferentes para diferentes tipos de perturbação. Partindo do mesmo tipo de procedimento de simulação de dados, em Milligan [53] analisou o comportamento de 30 medidas de critério interno.

¹ O cálculo desse desvio será tratado no capítulo 5 deste trabalho.

² Que será definido no capítulo 5.

- Em Sriram & Lewis [69] são definidas as noções de modelo *ótimo internamente* e de modelo *ótimo externamente*. Usa várias técnicas de classificação hierárquica, um algoritmo de programação matemática e duas heurísticas. Os dados baseiam-se em ultramétricas geradas pelo algoritmo de De Soete et al [20]. Como medida de optimalidade interna utiliza o quadrado do coeficiente de correlação linear de Pearson, entre os dados e a estrutura obtida. Para índice de optimalidade externa usa o *índice de Rand corrigido* para partições e o coeficiente de Goodman & Kruskal para ultramétricas. Este trabalho mostra que há uma correlação positiva acentuada entre os critérios de optimalidade interna e externa, mas que nem sempre os dois critérios conduzem à mesma decisão.

No capítulo 6 será apresentada uma metodologia que permite avaliar a qualidade e a robustez da hierarquia produzida, bem como medir o impacto que a aplicação de um dado método tem sobre diferentes tipos de estruturas de dados.

REFERÊNCIAS

- [1] Anderberg, R. (1973): *Cluster Analysis for Applications*, Academic Press, New York, London.
- [2] Bacelar-Nicolau, H. (1972): *Analyse d'un Algorithme de Classification Automatique*, Thèse de Doctorat de 3^{ème} Cycle, ISUP, Paris VI.
- [3] Bacelar-Nicolau, H. (1980): *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória*, Tese de Doutoramento, Lisboa.
- [4] Bacelar-Nicolau, H., Nicolau, F. C. (1994): Hierarchical Clustering Models and Robustness, *Actes des XXVI Journées de Statistique*, A.S.U., Neuchatel, pp. 96-99.
- [5] Baker, F. B. (1974): Stability of Two Hierarchical Grouping Techniques. Case I: Sensitivity to Data Errors, *Journal of the American Statistical Association*, 69, 440-445.
- [6] Barbut, M. e Monjardet, B (1970): *Ordre et Classification, Algèbre et Combinatoire, Tome 1, Méthodes Mathématiques des Sciences de L'Homme*, Classiques Hachette.
- [7] Barthélemy, J.-P., Leclerc, B. e Monjardet, B. (1986): On the Use of Ordered Sets in Problems of Comparison and Consensus Classifications, *Journal of Classification*, 3, 187-224.
- [8] Barthélemy, J.-P. e Guénoche, A. (1988): *Les Arbres et les Représentations des Proximités, Methode + Programmes*, Masson, Paris.
- [9] Batagelj, V. (1981): Note on Ultrametric Hierarchical Clustering Algorithms, *Psychometrika*, 46, 351-352.
- [10] Bel Mufti, G. (1998): *Validation d'une Classe par Estimation de sa Stabilité*, Thèse de Doctorat, Université Paris IX-Dauphine, Paris.
- [11] Benzecri (1984): *L'Analyse des Données*, Dunod, Paris.
- [12] Bock, H. H. (1985): On Some Significance Tests in Cluster Analysis, *Journal of Classification*, 2, 77-108.
- [13] Bock, H. H. (1996): Probability Models and Hypotheses Testing in Partitioning Cluster Analysis, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, pp 377-453.
- [14] Boorman, S. A. e Olivier, D. C. (1973): Metrics on Spaces of Finite Trees, *Journal of Mathematical Psychology*, 10, 26-59.
- [15] Celeux, G., Diday, E., Govaert, G., Lechevallier, Y. e Ralambondrainy, H. (1989): *Classification Automatique des Données*, Dunod informatique, Paris.

- [16] Chen, Z. e Van Ness, J. W. (1996): Space-Conserving Agglomerative Algorithms, *Journal of Classification*, 13, 157-168.
- [17] Constantinescu, P. (1966): The Classification of a Set of Elements with Respect to a Set of Properties, *Computer Journal*, 8, 352-357.
- [18] Cormack, R. M. (1971): A Review of Classification, *Journal of the Royal Statistical Society Association*, A, 134, 321-367.
- [19] Day, W. H. E. e Faith, D. P. (1986): A Model in Partial Orders for Comparing Objects by Dualistic Measures, *Mathematical Biosciences*, 78, 179-182.
- [20] De Soete, G., DeSarbo, W. S., Furnas, G. W. e Carroll, J. D. (1984): The Estimation of Ultrametric and Path Length Trees from Rectangular Proximity Data. *Psychometrika*, 49, 289-310.
- [21] Faith, D. P. e Belbin, L. (1986): Comparison of Classifications Using Measures Intermediate Between Metric Dissimilarity and Consensus Similarity, *Journal of Classification*, 3, 257-280.
- [22] Felsenstein, J. (1978): The Number of Evolutionary Trees, *Systematic Zoology*, 27, 27-33.
- [23] Foulds, L. R. e Robinson, R. W. (1981): *Enumeration of Binary Phylogenetic Trees*, Lecture Notes in Mathematics, Springer-Verlag, New York, 884, pp.173-186.
- [24] Frank, O. e Svensson, K. (1981): On Probability Distributions of Single-Linkage Dendograms, *Journal of Statistical Computation and Simulation*, 12, 121-131.
- [25] Furnas, G. W. (1984): The Generation of Random, Binary Unordered Trees, *Journal of Classification*, 1, 187-233.
- [26] Gordon, A. D. (1981): *Classification*. Chapman & Hall, London.
- [27] Gordon, A.D. (1987): A Review of Hierarchical Classification, *Journal of the Royal Statistical Society*, A, 150, 119-137.
- [28] Gordon, A. D. (1990): Constructing Dissimilarity Measures, *Journal of Classification*, 7, 257-269.
- [29] Gordon, A. D. (1994): Identifying Genuine Clusters in a Classification, *Computational Statistics & Data Analysis*, 18, 561-581.
- [30] Gordon, A. D. (1996): Hierarchical Classification, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, pp 65-121.
- [31] Gordon, A. D. (1998): Partitions of Partitions, *Journal of Classification*, 15, 265-285.
- [32] Gordon, A. D. (1999): *Classification*, 2nd ed., Chapman & Hall, London.

- [33] Gower, J. C. (1985): Measures of Similarity, Dissimilarity, and Distance, in *Encyclopedia of Statistical Sciences* (eds. S. Kotz, N. L. Johnson and C.B. Read), Wiley, New York, vol.5, pp. 397-405.
- [34] Hartigan, J. A. (1967): Representation of Similarity Matrices by Trees, *Journal of the American Statistical Association*, 62, 1140-1158.
- [35] Hartigan, J. A. (1975): *Clustering Algorithms*, Wiley, New York.
- [36] Hubert, L. J. (1987): *Assignment Methods in Combinatorial Data Analysis*, Marcel Dekker, New York.
- [37] Jain, A. K. e Dubes, R. C. (1988): *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ.
- [38] Jambu, M. e Lebeaux, M. O. (1978): *Classification Automatique pour l'Analyse des Données, Tome I: Méthodes et Algorithmes*, Dunod, Paris.
- [39] Jardine, N. e Sibson, R. (1971): *Mathematical Taxonomy*, Wiley, London.
- [40] Knuth, D. E. (1973): *The Art of Computer Programming*, Vol. 1, 2nd ed., Reading, MA: Addison-Wesley.
- [41] Lance, G. N. e Williams, W. T. (1966): A Generalised Sorting Strategy for Computer Classifications, *Nature*, 212, 218.
- [42] Lance, G. N. e Williams, W. T. (1967): A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems, *Computer Journal*, 9, 373-380.
- [43] Lapointe, F.-J. e Cucumel, G. (1997): The Average Consensus Procedure : Combination of Weighted Trees Containing Identical or Overlapping Sets of Taxa, *Systematic Biology*, 46, 306-312.
- [44] Lapointe, F.- J. e Legendre, P. (1990): A Statistical Framework to Test the Consensus of two Nested Classifications, *Systematic Zoology*, 39, 1-13.
- [45] Lapointe, F.-J. e Legendre, P. (1991): The Generation of Random Ultrametric Matrices Representing Dendrograms, *Journal of Classification*, 8, 177-200.
- [46] Lapointe, F.-J. e Legendre, P. (1995): A Comparaison Tests for Dendrograms : A Comparative Evaluation, *Journal of Classification*, 12, 265-282.
- [47] Leclerc, B. (1982): *Description, Évaluation et Comparaison des Hiérarchiques de Parties*, Rapport, Centre d'Analyse et de Mathématique Sociales.
- [48] Lerman, I. C. (1970): Sur l'Analyse des Données à une Classification Automatique, *Mathématiques et Sciences Humaines*, 32, 5-15.
- [49] Lerman, I. C. (1970): *Les Bases de la Classification Automatique*, Gauthier-Villars, Paris.
- [50] Lerman, I. C. (1981): *Classification et Analyse Ordinale des Données*, Dunod, Paris.

- [51] Milligan, G. W. (1979): Ultrametric Hierarchical Clustering Algorithms, *Psychometrika*, 44, 343-346.
- [52] Milligan, G. W. (1980): An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, *Psychometrika*, 45, 325-342.
- [53] Milligan, G. W. (1981): A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis, *Psychometrika*, 46, 187-199.
- [54] Milligan, G. W. (1996): Clustering Validation: Results and Implications for Applied Analyses, in *Clustering and Classification* (eds. P. Arabie, L. J. Hubert, G. De Soete), World Scientific, Singapore, pp. 341-375.
- [55] Murtagh, F. (1983): A Probability Theory of Hierarchic Clustering Using Random Dendograms, *Journal of Statistics and Computer Simulation*, 18, 145-157.
- [56] Murtagh, F. (1984): Counting Dendrograms: A Survey, *Discrete Applied Mathematics*, 7, 191-199.
- [57] Nicolau, F. C. (1980): *Critérios de Análise Classificatória Hierárquica Baseados na Função Distribuição*, Tese de Doutoramento, Faculdade de Ciências, Universidade de Lisboa.
- [58] Nicolau, F. C. (1981): Análise Classificatória e Função de Distribuição, *Actas de VIII Jornadas Luso-Espanholas de Matemática*, Vol. II, Coimbra, pp. 369-377.
- [59] Nicolau, F. C. (1981): Novos Aspectos da Análise Classificatória, *Actas do II Colóquio de Estatística e Investigação Operacional*, Fundação – Covilhã, pp. 323-341.
- [60] Nicolau, F. C. (1983): Cluster Analysis and Distribution Function, *Methods of Operations Research*, 45, 431-433.
- [61] Oden, N. L. e Shao, K. T. (1984): An Algorithm to Equiprobably Generate all Directed Trees with K Labeled Terminal Nodes and Unlabeled Interior Nodes, *Bulletin of Mathematical Biology*, 46, 379-387.
- [62] Quiroz, A. J. (1989): Fast Random Generation of Binary, t-ary and Other Types of Trees, *Journal of Classification*, 6, 223-231.
- [63] Rohlf, F. J. (1982): Consensus Indices for Comparing Classifications, *Mathematical Biosciences*, 59, 131-144.
- [64] Rohlf, F. J. (1983): Numbering Binary Trees with Labeled Terminal Vertices, *Bulletin of Mathematical Biology*, 45, 33-40.
- [65] Saporta, G. (1990): *Probabilités Analyse des Données et Statistique*, Editions Technip, Paris.
- [66] Shao, K. e Rohlf, F. J. (1983): Sampling Distribution of Consensus Indices when all Bifurcating Trees are Equally Likely, *Numerical Taxonomy*, 132-137.

-
- [67] Shao, K. e Sokal, R. R. (1986): Significance Tests of Consensus Indices, *Systematic Zoology*, 35, 582-590.
- [68] Sibson, R. (1972): Order Invariants Methods for Data Analysis, *Journal of the Royal Statistical Society, B*, 34, 311-349.
- [69] Sriram, N. e Lewis, S. (1993): Constructing Optimal Ultrametrics, *Journal of Classification*, 10, 241-268.
- [70] Van Ness, J., W. (1999): Recent Results in Clustering Admissibility, in *Applied Stochastic Models and Data Analysis* (eds. H. Bacelar-Nicolau, F. C. Nicolau, J. Janssen), Instituto Nacional de Estatística, pp. 19-29.
- [71] Ward, J. H. (1963): Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58, 236-244.

CAPÍTULO 3
NOVOS MÉTODOS DE
TENDÊNCIA CENTRAL EM C.H.A.

CAPÍTULO 3

NOVOS MÉTODOS DE TENDÊNCIA CENTRAL EM C.H.A.

ÍNDICE

INTRODUÇÃO

3.1 MÉTODOS DIRECTOS – ABORDAGEM CLÁSSICA

3.1.1 Critério *AMG*

3.1.2 Critério *AMed*

3.1.3 Critério *AC*

3.2 MÉTODOS *VL* – ABORDAGEM VALIDADE DA LIGAÇÃO

3.2.1 Critério *AVMG*

3.2.2 Critério *AVMed*

3.2.2.1 Resultados Preliminares

3.2.2.2 Dimensão Amostral Ímpar

3.2.2.2 Dimensão Amostral Par

3.2.3 Critério *AVC*

REFERÊNCIAS

CAPÍTULO 3

NOVOS MÉTODOS DE TENDÊNCIA CENTRAL EM C.H.A.

INTRODUÇÃO

C.H.A. é uma metodologia que se baseia numa dupla escolha inicial: a medida de comparação (semelhança/dissemelhança) entre elementos e a sua extensão à medida de comparação entre classes.

Em Nicolau [13], foi proposto e estudado o método *AVM*, que utiliza a metodologia *VL* (Validade da Ligação), já referida em 2.2.2.2., que recorre à estatística média aritmética para definir a função de comparação entre partes.

São agora aqui apresentados novos métodos de classificação hierárquica ascendente, com uma característica comum entre si e entre o método *AVM*: a função de comparação entre partes do conjunto a classificar tem por base uma estatística amostral de tendência central. As estatísticas de tendência central utilizadas serão a mediana, a média geométrica e o centro.

Este capítulo encontra-se dividido em duas partes. Na primeira secção, a que se associou o nome Métodos Directos, serão apresentados os métodos *AMG*, *AMed* e *AC*, respectivamente Agregação pela Média Geométrica, Agregação pela Mediana e Agregação pelo Centro. Nestes métodos a função de comparação entre partes é dada pelo valor da estatística de tendência central, correspondente, calculada para a amostra que é constituída pelos valores da função de comparação entre os elementos das partes a comparar.

A segunda secção deste capítulo é dedicada aos novos critérios segundo a metodologia *VL*, *AVMG* (Agregação pela Validade da Média Geométrica), *AVMed* (Agregação pela Validade da Mediana) e *AVC* (Agregação pela Validade do Centro), que estão no prolongamento dos trabalhos iniciados por Lerman [9], desenvolvidos por Bacelar-Nicolau [1] e [2] e Nicolau

[13]. Uma parte deste trabalho encontra-se em Nicolau & Sousa [14]. Como já foi referido a metodologia *VL*, recorre a noções probabilísticas para a definição das funções de comparação.

Uma versão preliminar dos métodos *AMG*, *AC* e *AVMG* foram integrados num programa experimental de classificação, adaptado por F. Nicolau, de um programa em linguagem Basic de Roux [18], em uso corrente no ensino de Análise de Dados, no início da década de 90, na Universidade de Aveiro.

No que se segue será usada a notação introduzida no capítulo 2. Assim o conjunto a classificar, de m elementos, será designado por E , a função de comparação entre elementos de E denotar-se-á por γ e a função de comparação entre partes por Γ .

Os métodos aqui apresentados foram programados e integrados na sequência de programas de classificação *CLUSTER* já existente, constituindo a versão *CLUSTER7*. Pretendia-se que esta sequência de programas tivesse um funcionamento autónomo. Esta opção explica o não recurso a bibliotecas de programas, como por exemplo a *NAG*.

3.1 MÉTODOS DIRECTOS – ABORDAGEM CLÁSSICA

3.1.1 Critério *AMG*

Dadas duas classes, A e B , respectivamente com n_A e n_B elementos de E , o método aqui proposto considera para função de comparação entre A e B , a média geométrica dos valores do conjunto $\{\gamma(a,b): (a,b) \in A \times B\}$, isto é:

$$\Gamma_{MG}(A,B) = \sqrt[n_A n_B]{\prod_{(a,b) \in A \times B} \gamma(a,b)} \quad (3.1)$$

Após reunir as classes A e B é necessário actualizar os valores da função de comparação entre partes, que têm por argumento $A \cup B$. Em 2.2.2.1 foi referida a existência de fórmulas de recorrência, (2.14) e (2.15), para esta actualização.

Dada uma classe C , de cardinal n_C , a função de comparação da nova classe, $A \cup B$, com C , é dada por:

$$\Gamma_{MG}(A \cup B, C) = \left[\prod_{a \in (A \cup B), c \in C} \gamma(a,c) \right]^{\frac{1}{(n_A + n_B)n_C}}. \quad (3.2)$$

Aplicando logaritmos neperianos a (3.2) obtém-se:

$$\begin{aligned} \ln[\Gamma_{MG}(A \cup B, C)] &= \frac{1}{(n_A + n_B)n_C} \sum_{a \in (A \cup B), c \in C} \ln \gamma(a,c) \\ &= \frac{1}{(n_A + n_B)n_C} [n_A n_C \ln \Gamma_{MG}(A, C) + n_B n_C \ln \Gamma_{MG}(B, C)] \\ &= \frac{n_A}{n_A + n_B} \ln \Gamma_{MG}(A, C) + \frac{n_B}{n_A + n_B} \ln \Gamma_{MG}(B, C). \end{aligned} \quad (3.3)$$

Donde

$$\Gamma_{MG}(A \cup B, C) = \exp\left(\frac{n_A}{n_A + n_B} \ln \Gamma_{MG}(A, C) + \frac{n_B}{n_A + n_B} \ln \Gamma_{MG}(B, C)\right). \quad (3.4)$$

Transformando os valores da função de comparação entre elementos nos seus logaritmos, a actualização da função de comparação entre partes, no critério *AMG*, assume o aspecto de uma média aritmética.

O algoritmo associado ao critério *AMG*, adapta-se facilmente à fórmula de Lance & Williams:

i. Calcular

$$\gamma'(x, y) = \ln \gamma(x, y), \text{ com } (x, y) \in F, \quad (3.5)$$

em que F é o conjunto dos pares de elementos distintos de E .

ii. Usar a fórmula de recorrência de Lance & Williams, (2.14), para o caso da média geométrica, para os valores $\gamma'(x, y)$, isto é com:

$$\alpha_1 = \frac{n_A}{n_A + n_B}, \alpha_2 = \frac{n_B}{n_A + n_B} \text{ e } \beta = \gamma = 0. \quad (3.6)$$

iii. Atender a que a função associada ao índice de nível, definida em 2.2.1, é dada por

$$h'(i, j) = \exp(h(i, j)). \quad (3.7)$$

Verifica-se que este critério não dá origem a inversões nos valores de índice de nível. O facto de, na etapa k , as classes A e B serem as escolhidas para a reunião, garante que:

$$\Gamma'_{MG}(A, B) < \Gamma'_{MG}(A, C) \quad (3.8)$$

e

$$\Gamma'_{MG}(A, B) < \Gamma'_{MG}(B, C)^1. \quad (3.9)$$

¹ A notação Γ' é usada para lembrar que são os valores logaritmizados.

Aplicando as desigualdades (3.8) e (3.9) em (3.3) pode escrever-se:

$$\Gamma'_{MG}(A \cup B, C) > \frac{n_A}{n_A + n_B} \Gamma'_{MG}(A, B) + \frac{n_B}{n_A + n_B} \Gamma'_{MG}(A, B) = \Gamma'_{MG}(A, B), \quad (3.10)$$

mostrando a não existência de inversões.

Alternativamente, bastava atender ao facto da função logaritmo ser crescente e ver que os parâmetros da fórmula de Lance & Williams, definidos por (3.6), verificam as condições necessárias e suficientes para a não existência de inversões dadas por (2.16) a (2.18).

Atendendo à condição (2.19) pode ainda concluir-se que o critério *AMG* é conservativo segundo Chen & Van Ness [3].

3.1.2 Critério *AMed*

No critério *AMed* a função de comparação entre *A* e *B*, é definida pela estatística de ordem mediana das semelhanças entre os pares de elementos das duas classes, representa-se por Γ_{Med} e é dada por:

$$\Gamma_{Med}(A, B) = \text{Mediana}\{\gamma(a, b) : (a, b) \in A \times B\} \quad (3.11)$$

No critério *AMed* a actualização, em cada etapa do algoritmo, da função de comparação entre partes, não pode ser feita pela fórmula de recorrência de Lance & Williams, e é dada por:

$$\Gamma_{Med}(A \cup B, C) = \text{Mediana}\{\gamma(a, c) : (a, c) \in (A \cup B) \times C\}. \quad (3.12)$$

A função $\Gamma_{Med}(A \cup B, C)$ não pode ser calculada exclusivamente a partir de $\Gamma_{Med}(A, C)$ e de $\Gamma_{Med}(B, C)$, contudo mostra-se em seguida que:

$$\min\{\Gamma_{Med}(A, C), \Gamma_{Med}(B, C)\} \leq \Gamma_{Med}(A \cup B, C) \leq \max\{\Gamma_{Med}(A, C), \Gamma_{Med}(B, C)\}. \quad (3.13)$$

Na etapa *k*, sendo *A* e *B* as classes escolhidas para a reunião, pode afirmar-se que:

$$\Gamma_{Med}(A, B) < \Gamma_{Med}(A, C) \quad (3.14)$$

e

$$\Gamma_{Med}(A, B) < \Gamma_{Med}(B, C). \quad (3.15)$$

Admita-se¹ que

$$\Gamma_{Med}(A, C) < \Gamma_{Med}(B, C). \quad (3.16)$$

Nestas condições sabe-se que:

- i. Até ao valor $\Gamma_{Med}(A, C)$ há $n_A n_C / 2$ elementos da ligação A com C e um número de elementos da ligação B com C não superior a $n_B n_C / 2$. O número de elementos, da ligação $(A \cup B)$ com C , inferiores a $\Gamma_{Med}(A, C)$ é não superior a $(n_A + n_B) n_C / 2$. Isto permite concluir que:

$$\Gamma_{Med}(A \cup B, C) \geq \Gamma_{Med}(A, C). \quad (3.17)$$

- ii. Analogamente há $n_B n_C / 2$ elementos, da ligação das classes B com C , superiores a $\Gamma_{Med}(B, C)$ e um número de elementos da ligação A com C não superior a $n_A n_C / 2$. O número de elementos da ligação $(A \cup B)$ com C , superiores a $\Gamma_{Med}(B, C)$, é não superior a $(n_A + n_B) n_C / 2$. Isto permite concluir que:

$$\Gamma_{Med}(A \cup B, C) \leq \Gamma_{Med}(B, C). \quad (3.18)$$

De (3.17) e (3.18) resulta que

$$\Gamma_{Med}(A, C) \leq \Gamma_{Med}(A \cup B, C) \leq \Gamma_{Med}(B, C), \quad (3.19)$$

provando (3.13).

¹ A situação contrária era análoga.

Associado a este critério surge o problema de usar um algoritmo eficiente para a determinação da mediana de um conjunto de valores. A propriedade (3.13) permite simplificar significativamente o cálculo da mediana, quando as classes envolvidas têm cardinais elevados e quando se utilizam métodos de ordenação de conjuntos de valores.

A programação deste método não usou algoritmos de ordenação para a determinação da mediana amostral, antes recorreu à ideia do subprograma “MDIAN2” de Press et al. [16], adaptando-o.

Dado um vector X de valores reais, de dimensão n , este processo calcula a mediana, X_{Med} , sem modificar o vector X e sem ordenar os seus valores, e tem por base a resolução iterativa da equação implícita,

$$x_{Med} = \frac{\sum_{j=1}^n \frac{x_j}{|x_j - x_{Med}|}}{\sum_{j=1}^n \frac{1}{|x_j - x_{Med}|}}. \quad (3.20)$$

Foi realizado um breve estudo sobre o programa, controlando o número de passos para a convergência e o valor limite. Detectou-se um erro no subprograma “MDIAN2”, quando ocorrem vários empates na zona da mediana. A nossa nova versão “MDIAN3”, corrige este aspecto.

O critério $AMed$ também não dá origem a inversões. Para verificar esta propriedade basta atender a (3.14) e a (3.19). Este critério verifica ainda a condição (2.19), pelo que $AMed$ é, tal como o critério AMg , conservativo segundo Chen & Van Ness [3].

3.1.3 Critério AC

O critério AC , aqui apresentado, baseia-se na estatística centro amostral. A função de comparação entre duas partes, A e B de E , é dada por:

$$\Gamma_c(A, B) = \frac{\max(\gamma(a, b)) + \min(\gamma(a, b))}{2}, \text{ com } (a, b) \in A \times B. \quad (3.21)$$

No critério AC a actualização, em cada etapa do algoritmo, da função de comparação entre partes, é dada por:

$$\begin{aligned}\Gamma_c(A \cup B, C) &= \frac{\max\{\gamma(a, c)\} + \min\{\gamma(a, c)\}}{2}, \text{ com } (a, c) \in (A \cup B) \times C \\ &= \frac{1}{2} \max\{\max\{\gamma(a, c) : (a, c) \in A \times C\}, \max\{\gamma(b, c) : (b, c) \in B \times C\}\} + \\ &\quad + \frac{1}{2} \min\{\min\{\gamma(a, c) : (a, c) \in A \times C\}, \min\{\gamma(b, c) : (b, c) \in B \times C\}\} \quad (3.22)\end{aligned}$$

O algoritmo do critério AC não pode ser incluído na fórmula de recorrência de Lance & Williams, (2.14), já que o novo centro não pode ser calculado unicamente à custa dos centros anteriores.

Atendendo em (3.22) a que $\Gamma_c(A, C) > \Gamma_c(A, B)$ e $\Gamma_c(B, C) > \Gamma_c(A, B)$ e alguma manipulação algébrica, facilmente se conclui que neste critério não há inversões nos valores da função de comparação entre classes.

Por outro lado o critério AC verifica a condição (2.19) e é portanto conservativo segundo Chen & Van Ness [3].

3.2 MÉTODOS VL – ABORDAGEM VALIDADE DA LIGAÇÃO

Nesta secção são apresentados três novos métodos para calcular a semelhança entre partes de E , associando a cada critério uma subsecção e considerando para função de comparação entre elementos a semelhança VL . Tal como na secção anterior, estes métodos têm de comum o facto da função de comparação entre partes se basear em estatísticas de tendência central; a média geométrica, a mediana e o centro.

Como já foi referido aquando da exposição da metodologia VL , os valores $\gamma(x, y)$ são obtidos a partir da função de distribuição da função de comparação entre elementos. Como consequência, aceitando algumas hipóteses subjacentes, os $\gamma(x, y)$ serão vistos como concretizações de variáveis aleatórias independentes e com distribuição $U(0, 1)$.

Conscientes da incorrecção estatística, nesta secção e para não sobrecarregar a notação, irá usar-se a letra γ , indexada pela sigla do método em causa, para a variável aleatória e para a sua concretização.

3.2.1 O critério *AVMG*

O método aqui proposto, insere-se na metodologia *VL* e recorre à estatística média geométrica. Considera para função de comparação entre classes a função de distribuição da média geométrica dos valores do conjunto $\{\gamma(a,b):(a,b) \in A \times B\}$.

Seja γ_{MG} a estatística média geométrica de $n_A n_B$ variáveis aleatórias,

$$\gamma_{MG} = \sqrt[n_A n_B]{\prod_{(a,b) \in A \times B} \gamma(a,b)}, \quad (3.23)$$

da qual pretendemos deduzir a função de distribuição.

Aceitando as hipóteses de referência já introduzidas no ponto 2.2.2.2, as variáveis aleatórias $\gamma(a,b)$ são independentes e uniformemente distribuídas em $[0,1]$, permitindo afirmar que $-\ln \gamma(a,b)$ segue distribuição exponencial de parâmetro 1, $\text{Exp}(1)$.

É bem sabido que o logaritmo da média geométrica de um conjunto de valores positivos é a média aritmética dos logaritmos desses valores. Assim, aplicando logaritmos neperianos a (3.23) obtém-se:

$$-n_A n_B \ln \gamma_{MG} = \sum_{(a,b) \in A \times B} -\ln \gamma(a,b). \quad (3.24)$$

A soma de $n_A n_B$ variáveis aleatórias independentes com distribuição $\text{Exp}(1)$ é uma variável aleatória que segue uma lei gama de parâmetros $n_A n_B$ e 1, isto é:

$$-n_A n_B \ln \gamma_{MG} \sim G(n_A n_B, 1). \quad (3.25)$$

Como é conhecido, na situação particular do primeiro parâmetro da lei gama, $n_A n_B$, ser um inteiro e o segundo parâmetro ser a unidade, a distribuição assume um aspecto bastante mais simples e muito frequente, estando-lhe associada uma lei qui-quadrado.

Pode concluir-se que:

$$-2n_A n_B \ln \gamma_{MG} \sim \chi^2(2n_A n_B). \quad (3.26)$$

Seja $\gamma_{MG}(A, B)$ a realização da estatística γ_{MG} para o par de classes (A, B) . A semelhança entre as classes A e B , usando o método *AVMG* é dada por:

$$\begin{aligned} \Gamma_{VMG}(A, B) &= P(\gamma_{MG} \leq \gamma_{MG}(A, B)) \\ &= P(-2n_A n_B \ln \gamma_{MG} \geq -2n_A n_B \ln \gamma_{MG}(A, B)) \\ &= 1 - F(-2n_A n_B \ln \gamma_{MG}(A, B)), \end{aligned} \quad (3.27)$$

sendo F a função de distribuição da $\chi^2(2n_A n_B)$.

Para a programação do método *AVMG*, foi necessário arranjar um algoritmo eficiente, para calcular a função de distribuição da qui-quadrado, que envolve o cálculo numérico do integral da função gama incompleta.

No que se segue descreve-se sucintamente as etapas realizadas, as dificuldades encontradas e as opções feitas em termos de trabalho global.

Seja X uma variável aleatória com distribuição $\chi^2(r)$, o objectivo é encontrar um algoritmo eficiente para calcular numericamente a função distribuição de X .

Na resolução deste problema, foi inicialmente construído um programa que usava algoritmos publicados na revista *Applied Statistical* :

- Algoritmo AS 239 de Shea [19] para o cálculo do integral gama incompleto.

- Algoritmo AS 245 de Macleod [11] para o cálculo do logaritmo da função gama.
- Algoritmo AS 66 de Hill [7] para o cálculo da função distribuição da normal reduzida quando o n° de graus de liberdade é superior a 1000.

Para a comparação dos resultados obtidos, para diferentes valores de graus de liberdade e para diferentes valores de abcissas, recorreu-se a uma tabela de valores percentis para a distribuição qui-quadrado. A tabela usada foi a de [15], por se considerar a mais apropriada, atendendo à precisão e à informação aí contidas. Essa comparação levantou algumas questões quanto à utilização deste conjunto de algoritmos.

Da execução e análise do programa pode concluir-se que, para um número de graus de liberdade inferior a 30, o cálculo da função de distribuição qui-quadrado não é satisfatório. A zona de pior comportamento ocorre, para um número de graus de liberdade compreendido entre 17 e 21 e em torno do valor mediano, chegando o pôr em causa a primeira casa decimal.

Em alternativa, elaborou-se um programa adaptando algoritmos extraídos de Press et al. [16], nomeadamente:

- Algoritmo “*gammp*” para o cálculo da função gama incompleta.
- Algoritmo “*gser*” auxiliar de “*gammp*”.
- Algoritmo “*gcf*” auxiliar de “*gammp*”.
- Algoritmo “*gammln*” para o cálculo do logaritmo da função gama.

Para garantir a precisão da metodologia aqui adoptada, realizaram-se as seguintes experiências numéricas:

- Para $r = 1, \dots, 30$ e para todos os valores de abcissa que constavam da tabela já referida, foram calculados os respectivos valores da função de distribuição e a diferença entre estes e os valores tabelados. Os resultados mostram que as diferenças encontradas:

- São sempre inferiores a 5×10^{-4} , garantindo igualdade entre as 3 primeiras casas decimais;
 - Tendem a diminuir quando r aumenta.
 - São tendencialmente menores para as abcissas mais baixas.
- Para $r = 30, \dots, 100$ e r múltiplo de 10, e para os 13 valores de abcissa tabelados, foram calculados o valor da função de distribuição da lei qui-quadrado, o valor aproximado por recurso à distribuição normal, e as respectivas diferenças entre estes e entre estes e o valor tabelado. Para detalhes sobre as diversas opções aqui tomadas, e respectiva qualidade, ver por exemplo Jerrold, [8]. Foram assim analisadas e comparadas, relativamente à tabela, a precisão da função de distribuição calculada e a da aproximação, de uso comum:

$$\sqrt{2X} - \sqrt{2r-1} \cong N(0,1). \quad (3.28)$$

No conjunto das experiências realizadas, quanto à comparação entre os valores calculados por recurso à função de distribuição da qui-quadrado e os tabelados verificaram-se as características referidas no ponto anterior. No que respeita às diferenças entre os valores obtidos a partir da lei qui-quadrado e da aproximada podemos afirmar que:

- A diferença decresce quando r aumenta.
 - Fixado o valor de r , a diferença é máxima em torno da mediana.
 - Na mediana as diferenças encontradas foram sempre inferiores 8×10^{-3} e para as outras abcissas é sempre inferior a 4×10^{-3} .
- Para valores de r compreendidos entre 100 e 500, com intervalos de 20 calcularam-se o valor da função de distribuição da qui-quadrado e o obtido usando a aproximação (3.28)¹, bem com as respectivas diferenças. Constatou-se que:

¹ Apesar de, para valores de r superiores a 100, ser aceitável usar outra aproximação à normal reduzida.

- As diferenças decrescem quando r aumenta.
- Fixado o valor de r , a diferença é máxima em torno da mediana, diminuindo significativamente quando as abcissas se afastam do centro.
- Na mediana as diferenças encontradas foram sempre inferiores 5×10^{-3} .

Para o cálculo numérico da função de distribuição da lei qui-quadrado, na biblioteca de programas *CLUSTER7*, e após a análise dos resultados das experiências realizadas e aqui apresentadas, optou-se por usar o segundo processo, que incluía os algoritmos extraídos de Press et al. [16], e por não efectuar aproximações à normal.

3.2.2 O critério *AVMed*

A estatística de ordem mediana das semelhanças VL entre pares de elementos de duas classes A e B , representada por γ_{VMed} , é dada por:

$$\gamma_{VMed} = \text{Mediana}\{\gamma(a,b): (a,b) \in A \times B\} \quad (3.29)$$

A definição do critério *AVMed*, precisa da distribuição da estatística ordinal mediana amostral, cuja expressão depende do facto da dimensão da amostra ser par ou ímpar.

Correndo embora o risco de percorrer caminhos já usados por outros, apresentam-se de seguida alguns resultados, necessários à dedução da função de distribuição da estatística mediana. O percurso seguido permite, simultaneamente, organizar o método de cálculo, facilitando os algoritmos, e mostrar uma componente didáctica interessante.

3.2.2.1 Resultados Preliminares

Considerem-se as variáveis aleatórias $X_1, X_2, \dots, X_n$ contínuas, independentes e com a mesma função de distribuição $F(x)$, e sejam $X_{(1;n)}, X_{(2;n)}, \dots, X_{(n;n)}$ as correspondentes variáveis

aleatórias ordinais¹.

Genericamente $f_{(k;n)}$ e $F_{(k;n)}$ representam, respectivamente, a função densidade de probabilidade e a função de distribuição da estatística de ordem k , $X_{(k;n)}$, com $k = 1, \dots, n$, e $f_{(k,r;n)}$ a função densidade de probabilidade conjunta das estatísticas de ordem k e r , $X_{(k;n)}$ e $X_{(r;n)}$.

Resultado 3.1 – (Fizk [5], p. 376)

A função densidade de probabilidade de $X_{(k;n)}$ é dada por:

$$f_{(k;n)}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} [1 - F(x)]^{n-k} f(x), \quad (3.30)$$

onde $f(x)$ e $F(x)$ são a função densidade de probabilidade e a função de distribuição de cada variável X_i .

Resultado 3.2 – (David [4], p. 10)

A função densidade de probabilidade conjunta das estatísticas de ordem k e r , $X_{(k;n)}$ e $X_{(r;n)}$, $f_{(k,r;n)}$, é dada por:

$$f_{(k,r;n)}(x, y) = \frac{n!}{(k-1)!(r-k-1)!(n-r)!} [F(x)]^{k-1} [F(y) - F(x)]^{r-k-1} [1 - F(y)]^{n-r} f(x)f(y),$$

com $k < r$ e $x < y$. (3.31)

¹ Dada uma amostra de dimensão n , admite-se que $x_{(1;n)} < x_{(2;n)} < \dots < x_{(n;n)}$. De facto, sendo os X_i variáveis aleatórias contínuas, aceita-se que a probabilidade de haver dois valores iguais na amostra é zero, podendo assim usar o sinal de menor estrito, conforme é usual.

Resultado 3.3

Para todos os valores de m e n inteiros positivos e $x \in]0,0.5[$, verifica-se que:

$$\int_0^x t^m (1-2x+t)^n dt = \sum_{i=0}^n (-1)^i \frac{m!n!}{(m+1+i)!(n-i)!} x^{m+1+i} (1-x)^{n-i}. \quad (3.32)$$

Resultado 3.4

Para todos os valores de m e n inteiros positivos e $x \in]0.5,1[$, verifica-se que:

$$\begin{aligned} \int_{2x-1}^x t^m (1-2x+t)^n dt &= \\ &= \sum_{i=0}^n (-1)^i \frac{m!n!}{(m+1+i)!(n-i)!} x^{m+1+i} (1-x)^{n-i} + (-1)^{n+1} \frac{m!n!}{(m+n+1)!} (2x-1)^{m+n+1}. \end{aligned} \quad (3.33)$$

Os resultados 3.3 e 3.4, para pequenos valores de m e n , surgem naturalmente primitivando, sucessivamente por partes, os respectivos integrais dos resultados. A demonstração geral dos mesmos faz-se de forma semelhante e simples, recorrendo ao princípio de indução finita. A título de exemplo apresenta-se no Anexo 1 a demonstração do resultado 3.4.

Resultado 3.5 – (Fizk [5], p. 383)

Seja X uma variável aleatória contínua com função densidade de probabilidade f , em que $f(x) > 0$, para $x = a_\lambda$, sendo a_λ um quantil da população em causa. Então o quantil de uma amostra simples, de dimensão n , tem distribuição assintótica normal de média a_λ e desvio

$$\text{padrão } \frac{\sqrt{\lambda(1-\lambda)/n}}{f(a_\lambda)}.$$

3.2.2.2 Dimensão Amostral Ímpar

Recorde-se que, as variáveis aleatórias amostrais são provenientes de uma distribuição $U(0,1)$. A aplicação do resultado 3.1 a esta distribuição dá a (3.30) uma expressão particularmente simples:

$$f_{(k;n)}(x) = \begin{cases} \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k} & \text{se } 0 \leq x \leq 1 \\ 0 & \text{se } x < 0 \vee x > 1 \end{cases} \quad (3.34)$$

Dado que n e k são inteiros positivos, (3.34) é a função densidade de probabilidade de uma variável aleatória com lei beta, reduzida, com parâmetros k e $n-k+1$, isto é $X_{(k;n)} = \text{Beta}(k, n-k+1)$.

Regressando ao ambiente de *C.H.A.*, quando o número de pares $(a,b) \in A \times B$, $n_A n_B$, é ímpar, concluiu-se que:

$$\gamma_{VMed} \sim \text{Beta}\left(\frac{n_A n_B + 1}{2}, \frac{n_A n_B + 1}{2}\right), \quad (3.35)$$

sendo n_A e n_B os cardinais das classes A e B , respectivamente.

Assim, quando ambas as classes A e B têm cardinal ímpar, a semelhança Γ_{VMed} entre as duas classes, com o critério $AVMed$ é dada pelo valor da função de distribuição da lei beta, com os parâmetros expressos em (3.35), no ponto $\gamma_{VMed}(A, B)$.

A programação do critério $AVMed$, passa por construir um algoritmo para a determinação da função de distribuição da lei beta, a qual necessita de alguns cuidados numéricos. No que se segue indica-se o processo para esse cálculo.

De (3.34) pode escrever-se:

$$F_{(k;n)}(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^x t^{k-1} (1-t)^{n-k}, \text{ se } 0 \leq x \leq 1$$

$$= \frac{1}{B(k, n-k+1)} B_x(k, n-k+1), \quad (3.36)$$

onde B e B_x são as funções beta¹, respectivamente, nos pontos 1 e x .

O objectivo é arranjar um algoritmo eficiente para calcular (3.36). Considere-se um caso geral e recorra-se à função beta incompleta, $I_x(m, n)$, dada por:

$$I_x(m, n) = \frac{B_x(m, n)}{B(m, n)} = \frac{(m+n-1)!}{(m-1)!(n-1)!} \int_0^x t^{m-1} (1-t)^{n-1} dt. \quad (3.37)$$

Para calcular a função de distribuição da lei beta, foi adaptado o algoritmo recursivo apresentado por Milheiro & Calheiros [12].

□ Cálculo de $B(m, n)$:

- $b_0 = \frac{1}{m}$
- $b_i = \frac{i}{m+i} b_{i-1}$, com $i = 1, \dots, n-1$,

sendo $b_{n-1} = B(m, n)$. (3.38)

□ Cálculo de $B_x(m, n)$:

- $B_x(m, n) = \sum_{i=0}^{n-1} \frac{(m-1)!(n-1)!}{(m+i)!(n-i-1)!} x^{m+i} (1-x)^{n-i-1}$
- $c_0 = \frac{1}{m} x^m (1-x)^{n-1}$

¹Lembra-se que as funções beta de parâmetros m e n inteiros são dadas por $B_x(m, n) = \int_0^x t^{m-1} (1-t)^{n-1} dt$ e

$$B(m, n) = \int_0^1 t^{m-1} (1-t)^{n-1} dt = \frac{(m-1)!(n-1)!}{(m+n-1)!}.$$

$$\bullet \quad c_i = \frac{n-i}{m+i} \frac{x}{1-x} c_{i-1}, \text{ com } i = 1, \dots, n-1, \quad (3.39)$$

$$\text{vindo } \sum_{i=0}^{n-1} c_i = B_x(m, n).$$

A posse de um algoritmo eficiente para o cálculo da função de distribuição exacta da mediana, para dimensões amostrais ímpares, permitiu analisar a aproximação sugerida pelo resultado 3.5.

Desse resultado conclui-se que a estatística mediana amostral, associada a uma população $U(0,1)$, tem distribuição assintótica normal com média 0.5 e variância $0.5^2 / n$.

Para valores de dimensão amostral inferiores a 150 e para valores de abcissas múltiplos de 0.05, calculou-se o valor da função de distribuição exacta, dada pela lei beta, e o valor da função de distribuição aproximada, dada pela lei normal. Deste conjunto de experiências concluiu-se que:

- Só para valores de n superiores a 30 são garantidas duas casas decimais correctas, usando a aproximação.
- Para todos os valores de n considerados há valores de abcissas, para os quais, a diferença entre os respectivos valores da função de distribuição é superior a 5×10^{-4} .
- A aproximação é pior para valores de abcissas pequenos.

Optou-se, por estas razões, por não implementar a aproximação associada ao resultado 3.5, preferindo (3.36) para o cálculo da função de distribuição da mediana.

3.2.2.3 Dimensão Amostral Par

Sendo par a dimensão amostral, n , $n = 2k$ com k inteiro, é necessário deduzir a distribuição

da variável aleatória $\frac{X_{(k;n)} + X_{(k+1;n)}}{2}$.

Aplicando o Resultado 3.2 às estatísticas $X_{(k;n)}$ e $X_{(k+1;n)}$, e considerando a amostra proveniente de uma população $U(0,1)$, obtém-se a função densidade conjunta das duas variáveis ordinais:

$$f_{(k,k+1;n)}(x_k, x_{k+1}) = \frac{(2k)!}{[(k-1)!]^2} x_k^{k-1} (1-x_{k+1})^{k-1}, \text{ para } 0 \leq x_k < x_{k+1} \leq 1. \quad (3.40)$$

Considere-se a transformação do par aleatório $(X_{(k;n)}, X_{(k+1;n)})$ para o par (Y_1, Y_2) definido

por $Y_1 = X_{(k;n)}$ e $Y_2 = \frac{X_{(k;n)} + X_{(k+1;n)}}{2}$. A densidade conjunta de $g(y_1, y_2)$ é:

$$g(y_1, y_2) = 2 \frac{(2k)!}{[(k-1)!]^2} y_1^{k-1} (1-2y_2+y_1)^{k-1}, \text{ se } (y_1, y_2) \in D, \quad (3.41)$$

$$\text{com } D = \left\{ (y_1, y_2) \in \mathbb{R}^2 : 0 \leq y_1 < y_2 \leq \frac{1+y_1}{2} \leq 1 \right\}.$$

Determine-se agora a função densidade de probabilidade marginal de Y_2 , g_{Y_2} , que será uma função definida por expressões distintas e não nulas em dois ramos $0 \leq y_2 \leq 0.5$ e $0.5 < y_2 \leq 1$.

1º ramo - $0 \leq y_2 \leq 0.5$

$$\begin{aligned} g_{Y_2}(y_2) &= \int_0^{y_2} g(y_1, y_2) dy_1 = \\ &= 2 \frac{(2k)!}{[(k-1)!]^2} \int_0^{y_2} y_1^{k-1} (1-2y_2+y_1)^{k-1} dy_1. \end{aligned} \quad (3.42)$$

A aplicação do Resultado 3.3 a (3.42), para $m = n = k-1$ e $x = y_2$, permite escrever:

$$g_{Y_2}(y_2) = 2 \frac{(2k)!}{[(k-1)!]^2} \sum_{i=0}^{k-1} (-1)^i \frac{[(k-1)!]^2}{(k+i)!(k-i-1)!} y_2^{k+i} (1-y_2)^{k-i-1}, \quad (3.43)$$

ou ainda,

$$g_{Y_2}(y_2) = 2 \sum_{i=0}^{k-1} (-1)^i \frac{(2k)!}{(k+i)!(k-i-1)!} y_2^{k+i} (1-y_2)^{k-i-1}. \quad (3.44)$$

2º ramo - $0.5 < y_2 \leq 1$

$$\begin{aligned} g_{Y_2}(y_2) &= \int_{2y_2-1}^{y_2} g(y_1, y_2) dy_1 = \\ &= 2 \frac{(2k)!}{[(k-1)!]^2} \int_{2y_2-1}^{y_2} y_1^{k-1} (1-2y_2+y_1)^{k-1} dy_1. \end{aligned} \quad (3.45)$$

Associando o Resultado 3.4 a (3.45), para $m = n = k-1$ e $x = y_2$, obtém-se:

$$\begin{aligned} g_{Y_2}(y_2) &= 2 \frac{(2k)!}{[(k-1)!]^2} \left\{ \sum_{i=0}^{k-1} (-1)^i \frac{[(k-1)!]^2}{(k+i)!(k-i-1)!} y_2^{k+i} (1-y_2)^{k-i-1} + (-1)^k \frac{[(k-1)!]^2}{(2k-1)!} (2y_2-1)^{2k-1} \right\} \\ &= 2 \sum_{i=0}^{k-1} (-1)^i \frac{(2k)!}{(k+i)!(k-i-1)!} y_2^{k+i} (1-y_2)^{k-i-1} + 4k(-1)^k (2y_2-1)^{2k-1}. \end{aligned} \quad (3.46)$$

Recorrendo à função beta e às expressões (3.44) e (3.46) pode escrever-se:

$$g_{Y_2}(y_2) = \begin{cases} \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} y_2^{k+i} (1-y_2)^{k-i-1} & \text{se } 0 \leq y_2 \leq 0.5 \\ \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} y_2^{k+i} (1-y_2)^{k-i-1} + 4k(-1)^k (2y_2-1)^{2k-1} & \text{se } 0.5 < y_2 \leq 1 \\ 0 & \text{se } y_2 < 0 \vee y_2 > 1 \end{cases} \quad (3.47)$$

Para a definição do critério *AVMed* é necessário obter a expressão da função distribuição da estatística mediana, Y_2 , representada por G_{Y_2} .

1º ramo - $0 \leq y_2 \leq 0.5$

$$\begin{aligned}
 G_{Y_2}(y_2) &= \int_0^{y_2} g_{Y_2}(t) dt = \\
 &= \int_0^{y_2} \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} t^{k+i} (1-t)^{k-i-1} dt \\
 &= \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} \int_0^{y_2} t^{k+i} (1-t)^{k-i-1} dt \\
 &= \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} B_{y_2}(k+i+1, k-i) \\
 &= 2 \sum_{i=0}^{k-1} (-1)^i I_{y_2}(k+i+1, k-i). \tag{3.48}
 \end{aligned}$$

2º ramo - $0.5 < y_2 \leq 1$

$$\begin{aligned}
 G_{Y_2}(y_2) &= G_{Y_2}(0.5) + \int_{0.5}^{y_2} g_{Y_2}(t) dt \\
 &= G_{Y_2}(0.5) + \int_{0.5}^{y_2} \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} t^{k+i} (1-t)^{k-i-1} dt + \int_{0.5}^{y_2} 4k(-1)^k (2t-1)^{2k-1} dt \\
 &= G_{Y_2}(0.5) + \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} \int_{0.5}^{y_2} t^{k+i} (1-t)^{k-i-1} dt + (-1)^k (2y_2-1)^{2k} \\
 &= G_{Y_2}(0.5) + \sum_{i=0}^{k-1} (-1)^i \frac{2}{B(k+i+1, k-i)} \left[\int_0^{y_2} t^{k+i} (1-t)^{k-i-1} dt - \int_0^{0.5} t^{k+i} (1-t)^{k-i-1} dt \right] + \\
 &\quad + (-1)^k (2y_2-1)^{2k}
 \end{aligned}$$

$$\begin{aligned}
 &= 2 \sum_{i=0}^{k-1} (-1)^i I_{0.5}(k+i+1, k-i) + 2 \sum_{i=0}^{k-1} (-1)^i I_{y_2}(k+i+1, k-i) - 2 \sum_{i=0}^{k-1} (-1)^i I_{0.5}(k+i+1, k-i) + \\
 &\quad + (-1)^k (2y_2 - 1)^{2k} \\
 &= 2 \sum_{i=0}^{k-1} (-1)^i I_{y_1}(k+i+1, k-i) + (-1)^k (2y_2 - 1)^{2k} \tag{3.49}
 \end{aligned}$$

Atendendo a (3.48) e (3.49), a função distribuição da estatística mediana quando a dimensão amostral é par, assume a expressão seguinte:

$$G_{Y_2}(y_2) = \begin{cases} 0 & \text{se } y_2 < 0 \\ 2 \sum_{i=0}^{k-1} (-1)^i I_{y_2}(k+i+1, k-i) & \text{se } 0 \leq y_2 \leq 0.5 \\ 2 \sum_{i=0}^{k-1} (-1)^i I_{y_2}(k+i+1, k-i) + (-1)^k (2y_2 - 1)^{2k} & \text{se } 0.5 < y_2 \leq 1 \\ 1 & \text{se } y_2 > 1 \end{cases} \tag{3.50}$$

Pode agora concluir-se que, quando $n_A n_B$ é um número par, a semelhança Γ_{VMed} entre duas classes A e B , com o critério $AVMed$ é definida pelo valor da função de distribuição expressa em (3.50), no ponto $\gamma_{VMed}(A, B)$.

Na programação da função de distribuição dada por (3.50), usou-se o algoritmo de cálculo da função beta incompleta, referido no ponto anterior. O somatório das expressões de (3.50) fica bastante simplificado atendendo ao facto das parcelas consecutivas se relacionarem. Considerem-se duas parcelas consecutivas, p_{i+1} e p_i , do somatório da expressão (3.50), por exemplo para valores de $0 \leq y_2 \leq 0.5$,

$$p_{i+1} = (-1)^{i+1} I_{y_2}(k+i+2, k-i-1). \tag{3.51}$$

Primitivando por partes (3.51) e alguma manipulação algébrica permite concluir que:

$$p_{i+1} = (-1)^i \frac{1}{k+i+1} \frac{1}{B(k+i+1, k-i)} y_2^{k+i+1} (1-y_2)^{k-i-1} - p_i. \tag{3.52}$$

3.2.3 O critério AVC

O critério aqui apresentado baseia-se na estatística centro amostral, dada por:

$$\gamma_c = \frac{\max(\gamma(a,b)) + \min(\gamma(a,b))}{2}, \text{ com } (a,b) \in A \times B. \quad (3.53)$$

Na dedução da função distribuição de γ_c utiliza-se o seguinte resultado.

Resultado 3.6 - (Read [17], p. 481)

A função densidade de probabilidade da estatística centro amostral, C , associada a uma população contínua com funções densidade e distribuição $f(x)$ e $F(x)$, respectivamente, é dada por:

$$g(c) = 2n(n-1) \int_{-\infty}^c [F(2c-t) - F(t)]^{n-2} f(t) f(2c-t) dt. \quad (3.54)$$

Aplicando o resultado 3.6 e considerando as $n_A n_B$ variáveis aleatórias $\gamma(a,b)$ independentes e $U(0,1)$, facilmente se conclui que a função de distribuição de γ_C assume o aspecto:

$$\Gamma_{VC}(A,B) = \begin{cases} 0 & \text{se } \gamma_C(A,B) \leq 0 \\ 2^{n_A n_B - 1} \gamma_C(A,B)^{n_A n_B} & \text{se } 0 < \gamma_C(A,B) \leq 0.5 \\ 1 - 2^{n_A n_B - 1} (1 - \gamma_C(A,B))^{n_A n_B} & \text{se } 0.5 < \gamma_C(A,B) \leq 1 \\ 1 & \text{se } \gamma_C(A,B) > 1 \end{cases} \quad (3.55)$$

traduzindo a semelhança entre as classes A e B , respectivamente de cardinais n_A e n_B pelo critério AVC.

Para os critérios da metodologia VL aqui apresentados, não é possível analisar analiticamente o seu comportamento quanto à existência de inversões. No entanto a sua utilização prática

revelou que as inversões são possíveis. Como se fez notar anteriormente isso não significa uma fraqueza dos métodos, antes algo intrínseco da sua natureza, já que o espaço de contexto deixa de ser métrico para ser probabilístico.

Da experiência de trabalho com critérios que seguem a abordagem VL^1 , verificou-se que enquanto o método está a evoluir de forma positiva ocorrem frequentemente inversões, o que pode ser interpretado como um acréscimo de probabilidade de ocorrência da nova partição.

¹ De notar que os métodos da família VL não originam inversões.

REFERÊNCIAS

- [1] Bacelar-Nicolau, H. (1972): *Analyse d'un Algorithme de Classification Automatique*, Thèse de Doctorat de 3^{ème} Cycle, ISUP, Paris VI.
- [2] Bacelar-Nicolau, H. (1980): *Contribuições ao Estudo dos Coeficientes de Comparação em Análise Classificatória*, Tese de Doutoramento, Lisboa.
- [3] Chen, Z. e Van Ness, J. W. (1996): Space-Conserving Agglomerative Algorithms, *Journal of Classification*, 13, 157-168.
- [4] David, H. A. (1981): *Order Statistics*, 2nd edition, John Wiley & Sons, Inc.
- [5] Fizek, M. (1963): *Probability Theory and Mathematical Statistics*. John Wiley & Sons, Inc.
- [6] Gower, J. C. (1985): Measures of Similarity, Dissimilarity, and Distance, in *Encyclopedia of Statistical Sciences* (eds. S. Kotz, N. L. Johnson and C.B. Read), Wiley & Sons, Inc, vol.5, pp. 397-405.
- [7] Hill, I. D. (1973): Algorithm AS 66 - The Normal Integral, *Applied Statistics*, 22, 424-427.
- [8] Jerrold, H. Z. (1978): Approximations for the Percentage Points of the Chi-Squared Distribution, *Applied Statistics*, 27, 280-290.
- [9] Lerman, I.C. (1970): Sur l'Analyse des Données à une Classification Automatique. *Mathématiques et Sciences Humaines*, 32, 5-15.
- [10] Lerman, I. C. (1970): *Les Bases de la Classification Automatique*, Gauthier-Villars, Paris.
- [11] Macleod, A. J. (1989): Algorithm AS 245 – A robust and Reliable Algorithm for the Logarithm of the Gamma Function, *Applied Statistics*, 38, 397-423.
- [12] Milheiro, P. , Calheiros, F. (1988): Relatório de Progresso Interno.
- [13] Nicolau, F. C. (1980): *Critérios de Análise Classificatória Hierárquica Baseados na Função de Distribuição* - Tese de Doutoramento, Faculdade de Ciências, Universidade de Lisboa.
- [14] Nicolau, F. C., Sousa, F. (1995): Métodos de Classificação Baseados em Estatísticas de Tendência Central, *Actas do III Congresso da Sociedade Portuguesa de Estatística*, Guimarães, pp. 583 – 595.
- [15] Pearson, E. S., Hartley, H. O. (1972): *Biometrika Tables for Statisticians*, Vol. 2, Cambridge, University Press.

- [16] Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1986): *Numerical Recipes – The Art of Scientific Computing*, Cambridge University Press.
- [17] Read, C. B. (1985) : Midranges, in *Encyclopedia of Statistical Sciences* (eds. S. Kotz, N. L. Johnson and C.B. Read), Wiley& Sons, Inc, vol.5, pp. 481-483.
- [18] Roux, M. (1985): *Algorithmes de Classification*, Masson.
- [19] Shea, B. L. (1988): Algorithm AS 239 - Chi-Squared and Incomplete Gamma Integral, *Applied Statistics*, 37, 466-473.

CAPÍTULO 4
GERAÇÃO ALEATÓRIA DE
ÁRVORES DE CLASSIFICAÇÃO

CAPÍTULO 4

GERAÇÃO ALEATÓRIA DE ÁRVORES DE CLASSIFICAÇÃO

ÍNDICE

INTRODUÇÃO

4.1 METODOLOGIAS PARA A GERAÇÃO DE ÁRVORES

4.1.1 Geração de Árvores com Parâmetro de Forma

4.1.2 Geração Uniforme de Árvores

4.2 ANÁLISE DAS DIFERENTES METODOLOGIAS

4.2.1 Algumas Considerações

4.2.2 Comportamento Teórico das Diferentes Metodologias

4.2.3 Algumas Simulações

REFERÊNCIAS

CAPÍTULO 4

GERAÇÃO ALEATÓRIA DE ÁRVORES DE CLASSIFICAÇÃO

INTRODUÇÃO

Como já foi referido em 2.3, o conceito de árvore é vasto, mas este capítulo trata de um caso particular, os dendrogramas. Um dendrograma, resultado mais comum de uma *C.H.A.*, foi definido como uma árvore em que importam três aspectos: a topologia, a etiquetagem terminal e os níveis de fusão. Assume-se que os dendrogramas são binários.

Em 2.3.2 foi justificada a necessidade de gerar dendrogramas aleatórios e foi referida uma metodologia, método de *Permutação Dupla*, Lapointe & Legendre [3] e [4], para a geração de dendrogramas aleatórios.

Neste capítulo são considerados dois novos processos de geração de dendrogramas. Ao primeiro processo, deu-se o nome de *Geração de Dendrogramas com Parâmetro de Forma*, dado que a topologia (forma) gerada depende de um parâmetro p , com $0 < p \leq 0.5$. O segundo processo, *Geração Uniforme de Dendrogramas*, gera dendrogramas aleatória e uniformemente, no conjunto de todos os dendrogramas existentes com um dado número de nós terminais. A descrição da metodologia destes processos é feita na primeira secção.

Note-se que nos processos propostos, as árvores geradas são dendrogramas, isto é, os três aspectos que caracterizam um dendrograma são gerados aleatoriamente. Note-se ainda que as árvores são geradas partindo da raiz para as folhas terminais, isto é adopta-se um procedimento “top-down”.

No que se segue referem-se os níveis de fusão, contudo estes podem ser substituídos por índices de níveis. Esta substituição tem alguns problemas inerentes, que já foram referidos em 2.3.2, implica conhecer a distribuição dos índices de níveis, ou usar os valores dos índices de

um dado dendrograma. A distribuição dos índices de nível depende da distribuição das funções de comparação usadas na construção dos respectivos dendrogramas. Foi já referido no capítulo 2, que existe uma grande variedade dessas funções, o que dificulta muito este procedimento.

Na segunda secção será feita uma análise, teórica e recorrendo à simulação, dos dois processos de geração aqui propostos, usando o método de Permutação Dupla para a comparação dos resultados.

4.1 METODOLOGIAS PARA A GERAÇÃO DE ÁRVORES

Cada nó de uma árvore será constituído por um conjunto de elementos, elementos terminais das respectivas sub-árvores, e por um conjunto de nós, níveis ou índices de nível de fusão dos nós das sub-árvores correspondentes. Estas metodologias de geração actuam sobre cada nó da árvore, dividindo aleatoriamente os conjuntos dos elementos e dos nós.

O processo tem início na raiz, onde o conjunto dos elementos tem cardinal m e o conjunto dos nós cardinal $m-1$. A informação de cada nó é decomposta em dois ramos, actuando sucessivamente nos ramos esquerdos, por escolha, até que a decomposição trivial ocorra, passando depois ao ramo da direita e procedendo de igual forma. Este procedimento pára quando o ramo mais à direita está decomposto.

A figura seguinte é um esquema deste procedimento, comum às duas metodologias aqui apresentadas.

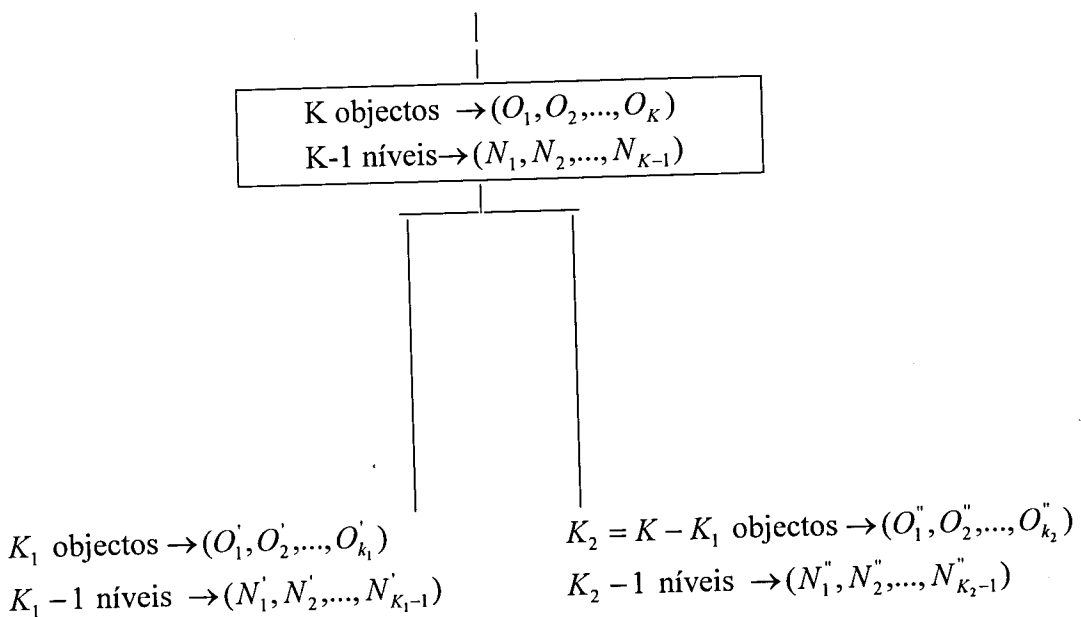


Fig. 4.1 – Esquema de procedimento de actuação num nó

A diferença destes dois processos de geração de dendrogramas reside na maneira como se efectua a divisão aleatória dos elementos em cada nó e dos respectivos níveis de agregação.

4.1.1 Geração de Árvores com Parâmetro de Forma

Neste processo de geração de dendrogramas é introduzido um coeficiente p , chamado parâmetro de forma, que, para cada geração, será uma constante do intervalo $]0,0.5]$, permitindo gerar árvores predominantemente de determinado tipo prefixado.

A divisão do conjunto de elementos obtém-se atendendo a que, em cada nó cada elemento tem probabilidade p de ir para o ramo da esquerda e $1-p$ de ir para o ramo da direita¹. Este coeficiente é a probabilidade de, na ramificação dos elementos de cada nó da árvore, cada elemento tem de pertencer ao ramo da esquerda.

Note-se que neste processo pode acontecer que o cardinal de um dos subconjuntos de elementos, o da direita ou o da esquerda, seja zero. Neste caso o procedimento de divisão de elementos repete-se até que esta situação não tenha lugar.

Associado a cada nó há um conjunto de níveis de agregação cujos elementos, após atribuição do maior valor ao nível do nó em causa, serão divididos pelos dois ramos de acordo com o número de elementos de cada subconjunto. A divisão dos níveis pelos dois ramos é feita de forma aleatória (com um procedimento análogo ao usado para os elementos).

Na geração de árvores por este processo, a tendência para gerar árvores do tipo cadeia aumenta, quando o valor do parâmetro p diminui. Quando p se aproxima de 0.5 as árvores geradas são predominantemente equilibradas. Mais à frente serão apresentadas tabelas que quantificam estas noções.

Neste processo de geração dá-se alguma importância a um valor particular de p , o 0.5 . Uma breve reflexão sobre a metodologia associada a este processo revela-nos que, para $p = 0.5$, as árvores do tipo equilibrado serão privilegiadas² relativamente às restantes árvores. De facto,

¹ É gerado um número aleatório uniforme em $(0,1)$, se for inferior a p pertence ao ramo da esquerda, caso contrário ao da direita.

² Note-se que o termo privilegiado está a ser usado em relação à proporção real do tipo equilibrado e não em relação à probabilidade de ramificação.

num dado nó com K elementos o número de elementos que irá para o ramo da direita será uma realização próxima¹ de uma variável aleatória binomial de parâmetros n e 0.5 .

4.1.2 Geração Uniforme de Árvores

As árvores são geradas aleatória e uniformemente no conjunto de todos os dendrogramas possíveis, fixado o número de elementos terminais, isto é uma geração uniforme no sentido de Furnas [1].

Em cada nó da árvore com um número de elementos superior a 3, gera-se aleatória e uniformemente o número de elementos que irá para o ramo da esquerda (direita). Assim, de acordo com a notação da Fig. 4.1, seja K o número de elementos que constituem um nó. É gerado um número, K_1 , uniforme do conjunto $\{1, 2, \dots, K-1\}$ e é feita uma partição aleatória dos K elementos em duas classes, uma com K_1 e outra com $K_2 = K - K_1$ elementos. Dos $K-1$ valores de níveis é atribuído o maior ao nó cuja ramificação está a ser efectuada, dos restantes $K-2$ é feita uma partição aleatória do tipo $(K_1 - 1, K_2 - 1)$.

Este procedimento é realizado em todos os nós com um número de elementos superior a 3. Num nó com um número de elementos não superior a 3 o tipo de divisão é único, sendo apenas escolhidas aleatoriamente as etiquetas.

4.2 ANÁLISE DAS DIFERENTES METODOLOGIAS

4.2.1 Algumas Considerações

Neste trabalho foram considerados três geradores de números pseudo-aleatórios uniformes no intervalo $]0,1[$, com características, à partida, diferentes. O gerador do tipo congruencial multiplicativo, Knuth [2], a que se chamou "*RANI*", o gerador incorporado no computador usado, *Pentium III* e no compilador do *Fortran Powerstation* versão 4.0 e ainda o gerador da Biblioteca *NAG*, o que levou à execução dos programas numa máquina *UNIX*.

¹ Não é exactamente uma binomial já que as ocorrências 0 e K não são possíveis.

Antes de iniciar os trabalhos de geração de árvores foi feita uma bateria de testes aos três geradores, não se tendo registado nenhuma situação que conduzissem à não utilização ou à utilização preferencial de algum deles. Por esta razão e atendendo ao facto de se poder controlar o gerador “RANI”, optou-se por usar este nas simulações seguintes.

O factor dupla precisão foi analisado e também aqui os resultados não pareciam depender da utilização de precisão simples ou dupla. Desta forma decidiu-se efectuar as simulações em precisão simples.

As árvores geradas pelos processos aqui considerados, são aleatórias quanto à topologia, etiquetagem e ordens dos níveis de fusão, o que engloba, obviamente, qualquer outro processo de geração em que só um ou dois destes aspectos são considerados.

Seja m o número de nós terminais de cada dendrograma. Para cada valor de m , o número de tipos topológicos de árvores, isto é o número de topologias distintas que é possível definir nos dendrogramas de m elementos terminais, é dado por (2.27)¹. O cálculo desse número, está estreitamente ligado à determinação do número de partições, não triviais, quanto ao número de elementos, que é possível efectuar num conjunto de cardinal m .

Para valores pequenos de m podem identificar-se esses tipos topológicos e recorrendo a cálculos de análise combinatória e de probabilidades pode determinar-se, do conjunto total de dendrogramas distintos, para m fixo, as proporções de dendrogramas de cada tipo.

Naturalmente que o trabalho deste procedimento tem um crescimento exponencial em m , restando-nos considerar valores de m pequenos² ou construir um processo indutivo. No que se segue consideram-se valores de m não superiores a sete.

As representações esquemáticas seguintes ilustram as sequências possíveis de agregação numa árvore binária. Fixado um valor de m , seguindo o esquema correspondente, é fácil identificar os diferentes tipos topológicos. É ainda possível perceber como surgem as proporções de dendrogramas de cada tipo.

¹ Para a ordem de grandeza destes valores ver quadro 2.2.

² Note-se que para $m=7$ já temos 56700 dendrogramas e 11 tipos topológicos.

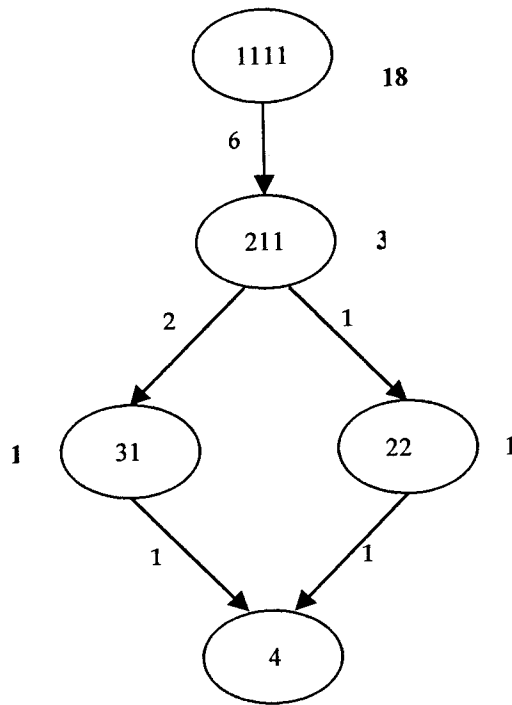


Fig. 4.2 – Representação esquemática dos dendrogramas para $m = 4$

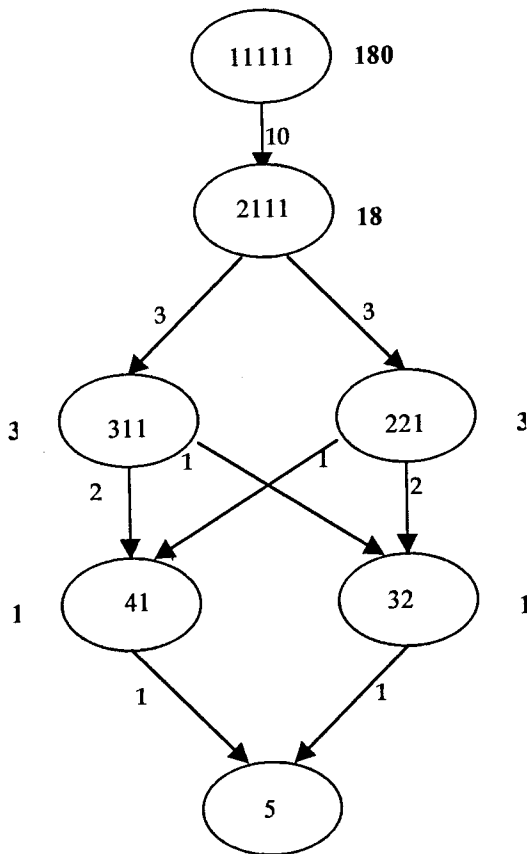


Fig. 4.3 – Representação esquemática dos dendrogramas para $m = 5$

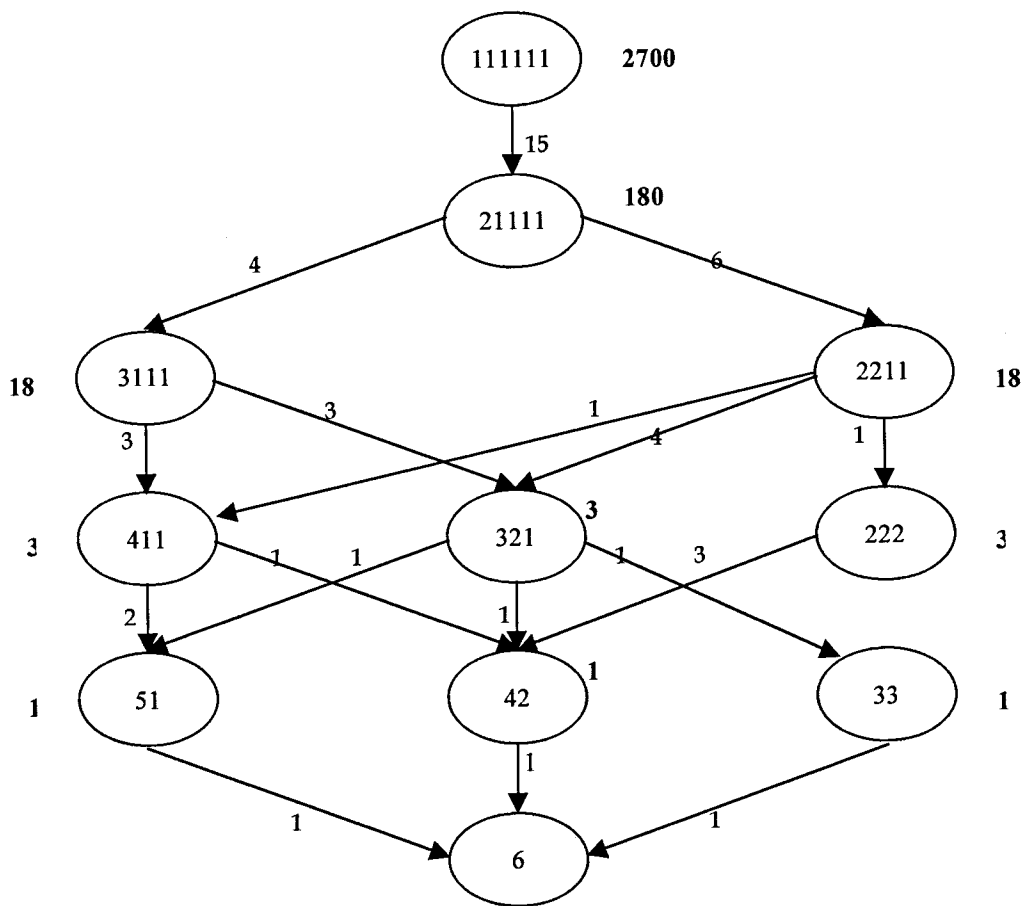


Fig. 4.4 – Representação esquemática dos dendrogramas para $m = 6$

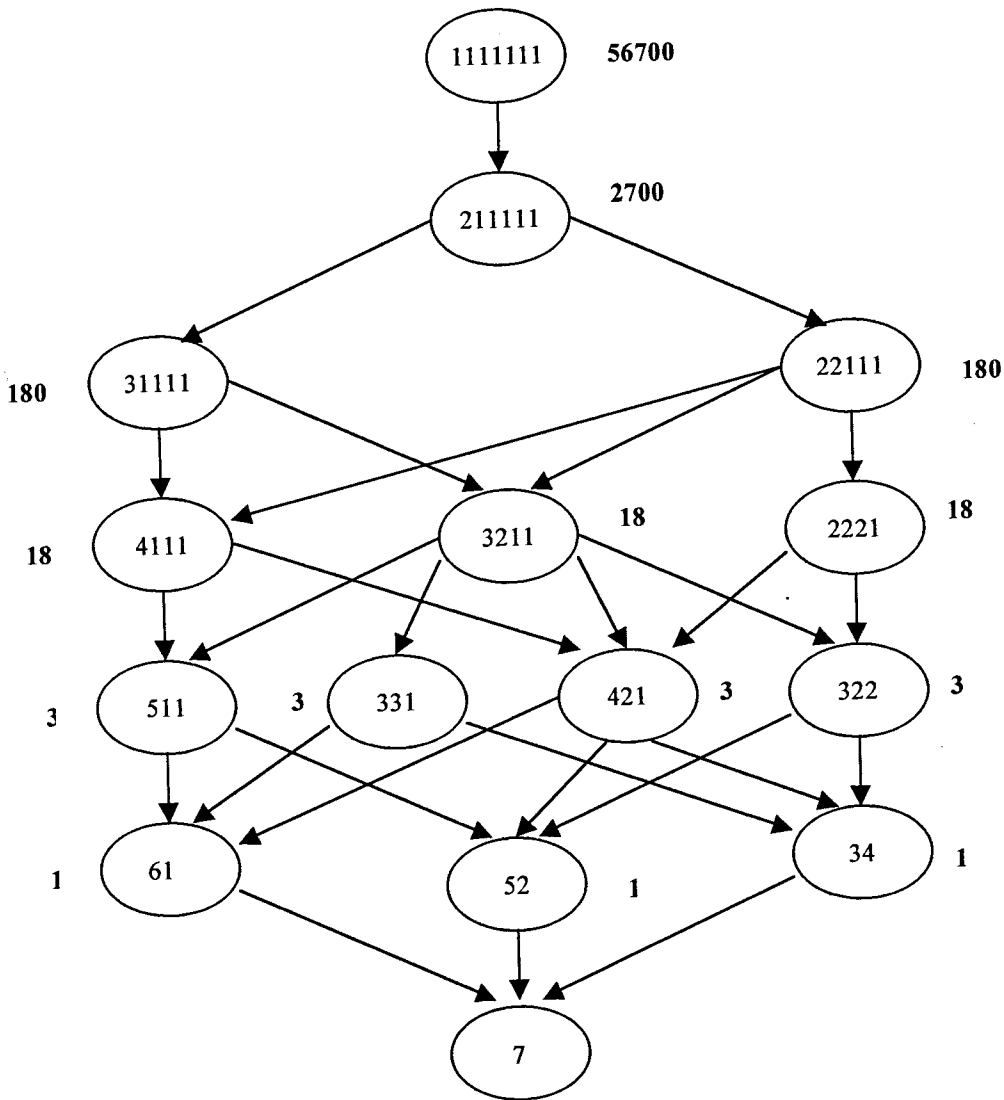


Fig. 4.5 – Representação esquemática dos dendrogramas para $m = 7$

Nas figuras seguintes, para cada valor de m compreendido entre 4 e 7, identificam-se os respectivos tipos topológicos.

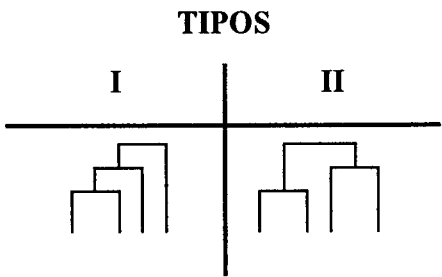


Fig. 4.6 – Tipos de dendrogramas para $m = 4$

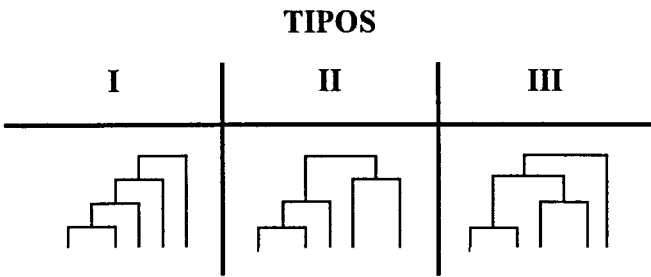


Fig. 4.7 – Tipos de dendrogramas para $m = 5$

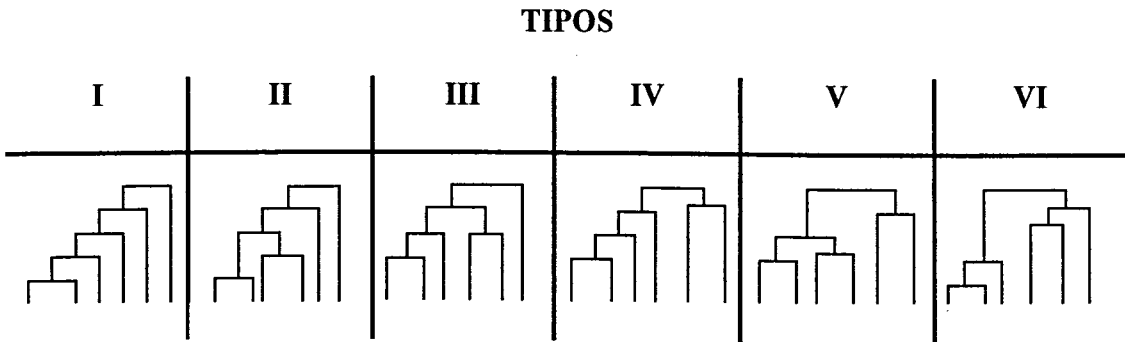


Fig. 4.8 – Tipos de dendrogramas para $m = 6$

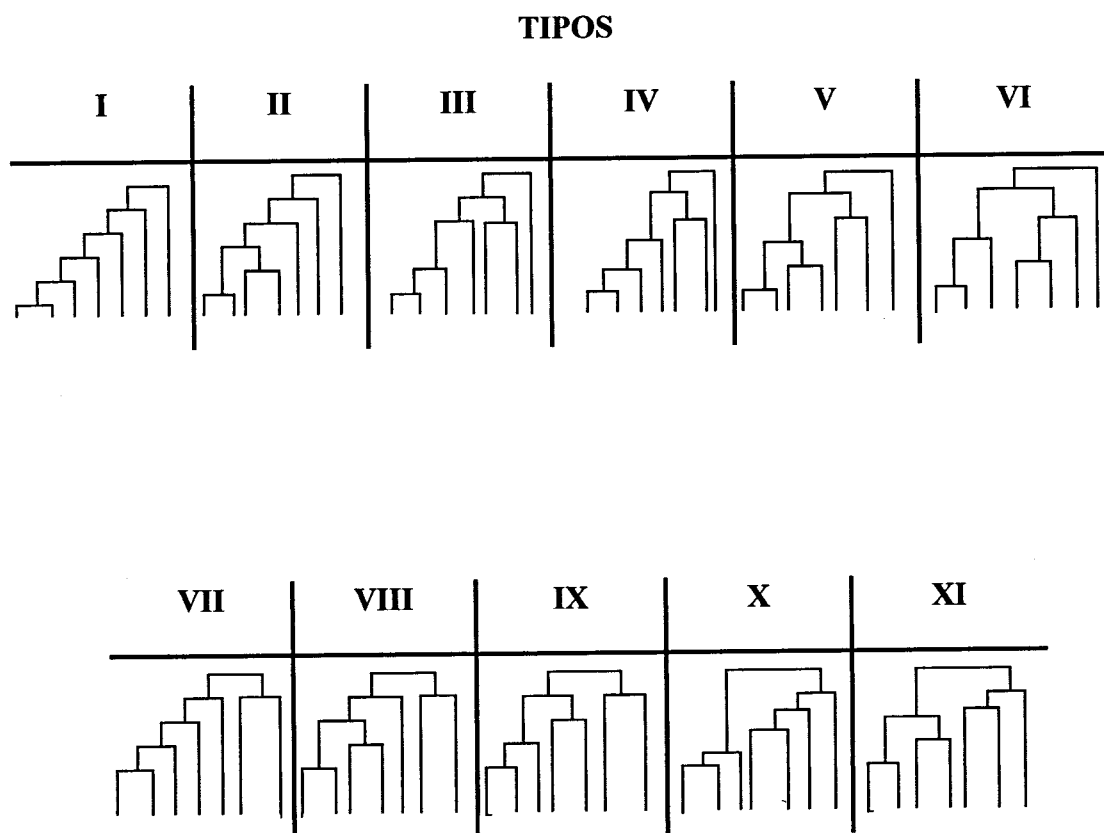


Fig. 4.9 – Tipos de dendrogramas para $m = 7$

Pretende-se analisar as metodologias de geração de dendrogramas expostas. Para tal, para cada valor de m , associado à geração aleatória de um dendrograma de m folhas, é feita uma identificação do seu tipo topológico. O dendrograma gerado é guardado recorrendo à “Notação Polaca”. Para os diferentes valores de m considerados, foi feita uma análise exaustiva das representações polacas possíveis, classificando-as nos vários tipos topológicos. Examinando a representação polaca da árvore gerada é identificado o seu tipo.

Na primeira metade deste século o logicista polaco Jan Lukasiewicz, apresentou vários trabalhos, sobre a utilização de uma notação original para a escrita de expressões matemáticas, inteiramente livres de parêntesis. Estes trabalhos surgiram inicialmente em alemão e só algumas décadas depois foram apresentados trabalhos sobre esse assunto em

inglês, Lukasiewicz [5] e [6]. O uso desta notação foi alargada à álgebra e a outros domínios e ficou conhecida como “*Notação Polaca*”. Este tipo de representação consiste basicamente em colocar, de forma sistemática, os operadores antes (ou depois) dos respectivos operandos, conduzindo à não necessidade de parêntesis, já que o número de operandos associados a cada operador é um número fixo preestabelecido.

Pode referir-se duas formas de notação polaca, a prefixo, na qual os operadores precedem os operandos, muito utilizada em linguagens informáticas, e a sufixo, em que os operadores surgem depois dos operandos, usada em expressões intermédias em muitos compiladores e nas calculadoras Hewlett-Packard. As propriedades de facilidade de cálculo e representação única, de uma expressão sem recurso a parêntesis, ou a outra notação, justificam o interesse na utilização da notação polaca em aplicações bem diversas que recorrem à programação computacional.

Como foi referido, a notação polaca será usada para representar uma árvore hierárquica. No que se segue é explicado o modo como a notação polaca é adaptada a este caso particular.

Os operadores são os índices de nível associados ao carácter ”-“ e os operandos os índices de elementos. Quanto à forma de representação polaca é do tipo prefixo, já que os operadores precedem os operandos. A representação da árvore em notação polaca será constituída pelos índices dos elementos classificados, colocados pela ordem esquerda - direita em que surgem na árvore e pelos níveis de agregação precedidos de um sinal “-“, acoplados aos respectivos pares de reunião. A introdução do sinal “-“ faz-se para permitir a distinção entre índice de nível e índice de elemento.

De notar que se está a admitir que todas as árvores são binárias. Contudo este facto não é restritivo, já que uma solução para árvores não binárias será usar repetidamente o mesmo índice de nível.

Em seguida apresenta-se uma árvore e a sua correspondente representação polaca, pretendendo ilustrar o que se acaba de referir.

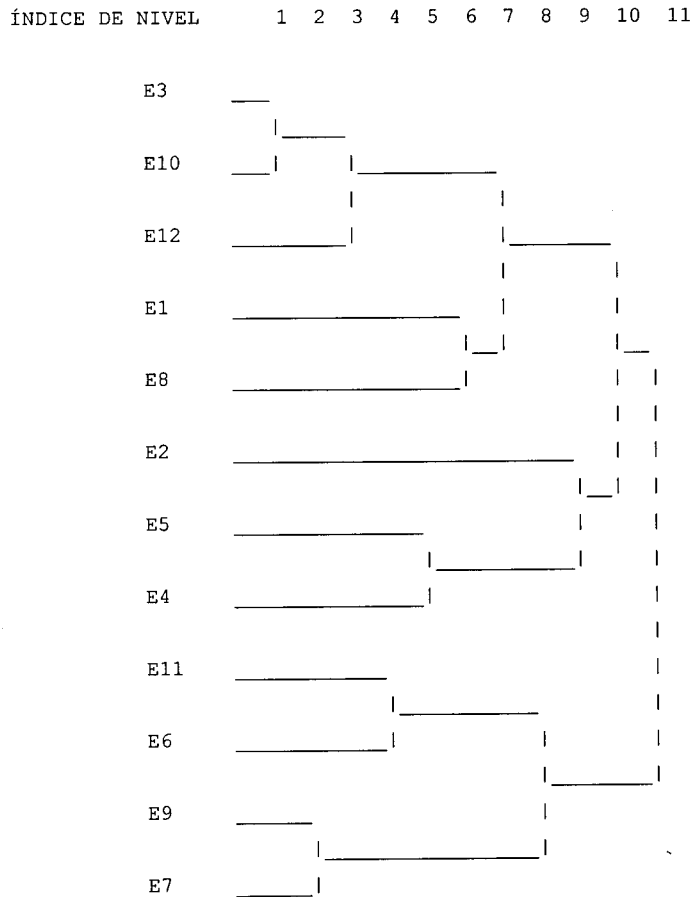


Fig. 4.10 – Representação da Árvore

-11 -10 -7 -3 -1 3 10 12 -6 1 8 -9 2 -5 5 4 -8 -4 11 6 -2 9 7

Fig. 4.11 – Representação Polaca da Árvore

A uma reunião binária estão sempre associados um operador, índice de nível, e dois operandos, elementos simples ou classes, constituindo uma sequência de um elemento negativo e duas quantidades com valor lógico positivo (sem sinal, por omissão). Note-se que a sequência de um operador e dois operandos, correspondendo a uma classe, constitui ela própria um elemento de valor lógico positivo.

Em seguida apresentam-se as sucessivas sequências que figuram na representação polaca anterior, quando esta é percorrida da esquerda para a direita.

-1 3 10

-3 (-1 3 10) 12

-6 1 8

-7 (-3 (-1 3 10) 12) (-6 1 8)

-5 5 4

-9 2 (-5 5 4)

-10 (-7 (-3 (-1 3 10) 12) (-6 1 8)) (-9 2 (-5 5 4))

-4 11 6

-2 9 7

-8 (-4 11 6) (-2 9 7)

-11 (-10 (-7 (-3 (-1 3 10) 12) (-6 1 8)) (-9 2 (-5 5 4))) (-8 (-4 11 6) (-2 9 7)).

Facilmente se verifica que existe uma correspondência biunívoca entre as árvores binárias e as representações polacas associadas. Note-se ainda que podia ser definida uma notação polaca do tipo sufixo, tudo decorreria de forma equivalente.

4.2.2 Comportamento Teórico das Diferentes Metodologias

A situação geral no processo de *Geração com Parâmetro de Forma* é ter um nó com K elementos, podendo cada elemento ser atribuído ao ramo da esquerda, com probabilidade p , ou da direita com probabilidade $1-p$. Seja X uma variável aleatória que representa o número de elementos atribuídos ao ramo da esquerda. Esta variável aleatória segue uma lei binomial de parâmetros K e p .

Por imposição do programa as situações $X = 0$ ou $X = K$ não são admitidas. Considere-se o acontecimento A associado a uma situação admitida pelo processo de geração. Pode escrever-se:

$$p_i = P(X = i | A) = P(X = i \cap A) / P(A), \text{ para } i = 1, 2, \dots, K-1. \quad (4.1)$$

Par ilustrar considere-se $K = 4$, donde $X = Bi(4, p)$ e então:

$$P(A) = 2p(1-p)[2(1-p)^2 + 3p(1-p) + 2p^2], \quad (4.2)$$

$$p_1 = p_3 = \frac{4p(1-p)^3}{P(A)} \quad \text{e} \quad p_2 = \frac{6p^2(1-p)^2}{P(A)} \quad (4.3)$$

De acordo com o exposto atrás, o tipo topológico I corresponde à situação de três elementos para um ramo e um elemento para o outro. Assim a probabilidade de uma árvore gerada por este processo ser do tipo I é dada por $p_1 + p_3$, enquanto que de ser do tipo II é dada por p_2 .

Para valores de m superiores a quatro o raciocínio de base é o mesmo, complicando-se bastante o cálculo das probabilidades, porque o número de nós intervenientes aumenta.

As fórmulas de cálculo do valor das probabilidades, dos vários tipos topológicos, para a geração de árvores com parâmetro de forma, foram deduzidas para valores de m inferiores a oito. Nos quadros 4.1 a 4.4, são apresentados esses valores de probabilidades para alguns valores de p .

Em seguida deduzem-se os valores de probabilidades, dos vários tipos topológicos dos dendrogramas, gerados pelo processo de *Geração Uniforme*.

Dado um nó com K elementos, $K > 3$, é gerado aleatória e uniformemente um número do conjunto $\{1, 2, \dots, K-1\}$, seja u , que será o número de elementos atribuído ao ramo da esquerda. Esta informação é guardada num par ordenado $(u, K-u)$. Note-se que as situações $(u, K-u)$ e $(K-u, u)$, quando $u \neq K-u$, são simétricas, portanto semelhantes topologicamente.

Seja p_i a probabilidade da decomposição $(i, K-i)$ dos K elementos. A expressão geral de p_i depende de K ser par ou ímpar.

Se K par, tem-se:

$$p_i = \frac{2}{K-1}, \text{ para } i = 1, \dots, \frac{K}{2} - 1$$

e

$$p_{K/2} = \frac{1}{K-1}.$$

(4.4)

Se K ímpar, tem-se:

$$p_i = \frac{2}{K-1}, \text{ para } i = 1, \dots, \frac{K-1}{2}.$$

(4.5)

Também nesta metodologia de geração de dendrogramas, foram deduzidos os valores das probabilidades para cada tipo topológico, quando o número de elementos terminais é menor que oito. Esses valores constam dos quadros 4.1 a 4.4.

Nos mesmos quadros encontram-se as probabilidades de cada tipo para a geração de árvores com o procedimento de *Permutação Dupla*.

	PROBABILIDADES POR TIPOS	
	I	II
Real	2/3 (.67)	1/3 (.33)
Parâmetro de Forma, p=.50	8/14 (.57)	6/14 (.43)
Parâmetro de Forma, p=.40	(.591)	(.409)
Parâmetro de Forma, p=.30	(.648)	(.352)
Parâmetro de Forma, p=.25	(.690)	(.310)
Parâmetro de Forma, p=.20	(.739)	(.261)
Parâmetro de Forma, p=.15	(.796)	(.204)
Parâmetro de Forma, p=.10	(.859)	(.141)
Parâmetro de Forma, p=.05	(.927)	(.073)
Parâmetro de Forma, p=.025	(.963)	(.037)
Geração Uniforme	2/3 (.67)	1/3 (.33)
Permutação Dupla	2/3 (.67)	1/3 (.33)

Quadro 4.1 – Valores de probabilidade para $m = 4$

4.2.3 Algumas Simulações

	PROBABILIDADES POR TIPOS		
	I	II	III
Real	2/6 (.33)	3/6 (.50)	1/6 (.17)
Parâmetro de Forma, p=.50	4/21 (.19)	14/21(.67)	3/21 (.14)
Parâmetro de Forma, p=.40	(.218)	(.632)	(.151)
Parâmetro de Forma, p=.30	(.304)	(.532)	(.165)
Parâmetro de Forma, p=.25	(.371)	(.462)	(.167)
Parâmetro de Forma, p=.20	(.458)	(.381)	(.161)
Parâmetro de Forma, p=.15	(.563)	(.292)	(.145)
Parâmetro de Forma, p=.10	(.689)	(.198)	(.113)
Parâmetro de Forma, p=.05	(.835)	(.100)	(.066)
Parâmetro de Forma, p=.025	(.915)	(.050)	(.035)
Geração Uniforme	2/6 (.33)	3/6 (.50)	1/6 (.17)
Permutação Dupla	2/6 (.33)	3/6 (.50)	1/6 (.17)

Quadro 4.2 – Valores de probabilidade para $m = 5$

	PROBABILIDADES POR TIPOS					
	I	II	III	IV	V	VI
Real	2/15 (.1333)	1/15 (.0667)	3/15 (.2000)	4/15 (.2667)	2/15 (.1333)	3/15 (.2000)
Parâmetro de Forma, p=.50	16/434 (.0369)	12/434 (.0276)	56/434 (.129)	120/434 (.2764)	90/434 (.2074)	140/434 (.3226)
Parâmetro de Forma, p=.40	.051255	.035485	.148697	.279678	.193623	.291262
Parâmetro de Forma, p=.30	.107664	.058473	.188587	.282020	.153166	.210090
Parâmetro de Forma, p=.25	.162839	.073278	.202386	.276600	.124470	.160428
Parâmetro de Forma, p=.20	.244814	.086405	.203827	.261594	.092327	.111034
Parâmetro de Forma, p=.15	.361425	.092782	.187567	.232090	.059580	.066556
Parâmetro de Forma, p=.10	.520905	.085759	.149588	.182573	.030058	.031117
Parâmetro de Forma, p=.05	.731314	.057576	.087399	.107181	.008438	.008091
Parâmetro de Forma, p=.025	.857955	.032977	.046859	.057928	.002227	.002055
Geração Uniforme	2/15 (.1333)	1/15 (.0667)	3/15 (.2000)	4/15 (.2667)	2/15 (.1333)	3/15 (.2000)
Permutação Dupla	2/15 (.1333)	1/15 (.0667)	3/15 (.2000)	4/15 (.2667)	2/15 (.1333)	3/15 (.2000)

Quadro 4.3 – Valores de probabilidade para $m = 6$



PROBABILIDADES POR TIPOS											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Real	4/90 (.044)	2/90 (.022)	6/90 (.667)	8/90 (.089)	4/90 (.044)	6/90 (.667)	10/90 (.111)	5/90 (.056)	15/90 (.167)	20/90 (.222)	10/90 (.111)
Parâmetro de Forma, p=.50	8/1953 (.004)	8/1953 (.003)	8/1953 (.014)	8/1953 (.031)	8/1953 (.023)	8/1953 (.036)	8/1953 (.064)	8/1953 (.048)	8/1953 (.222)	8/1953 (.318)	8/1953 (.238)
Parâmetro de Forma, p=.40	.007809	.005406	.022655	.042610	.029499	.044375	.075985	.052605	.220440	.294636	.203979
Parâmetro de Forma, p=.30	.029413	.015974	.051521	.077046	.041844	.057395	.113362	.061567	.198568	.228960	.124349
Parâmetro de Forma, p=.25	.058776	.026449	.073051	.099838	.044927	.057906	.138433	.062295	.172053	.183636	.082636
Parâmetro de Forma, p=.20	.113802	.040165	.094749	.121603	.042919	.051614	.161857	.057126	.134759	.134083	.047323
Parâmetro de Forma, p=.15	.210696	.054088	.109344	.135299	.034733	.038799	.174732	.044856	.090680	.084962	.021811
Parâmetro de Forma, p=.10	.371446	.061153	.106668	.130189	.021434	.022189	.163944	.026991	.047080	.041994	.006914
Parâmetro de Forma, p=.05	.623724	.049105	.074541	.091413	.007197	.006901	.112402	.008849	.013433	.011527	.000908
Parâmetro de Forma, p=.025	.794185	.030525	.043377	.053622	.002061	.001902	.065145	.002504	.003558	.003005	.000116
Geração Uniforme	4/90 (.044)	2/90 (.022)	6/90 (.667)	8/90 (.089)	4/90 (.044)	6/90 (.667)	10/90 (.111)	5/90 (.056)	15/90 (.167)	20/90 (.222)	10/90 (.111)
Permutação Dupla	4/90 (.044)	2/90 (.022)	6/90 (.667)	8/90 (.089)	4/90 (.044)	6/90 (.667)	10/90 (.111)	5/90 (.056)	15/90 (.167)	20/90 (.222)	10/90 (.111)

Quadro 4.4 – Valores de probabilidade para $m = 7$

4.2.3 Algumas Simulações

Com o objectivo de analisar o comportamento dos métodos de geração de árvores expostos em 4.1, foram realizados alguns estudos de simulação para pequenos valores de número de elementos terminais, m .

Para cada processo, para um gerador e para um valor de m fixo, compreendido entre 4 e 7, foi feita geração aleatória de árvores de acordo com o seguinte esquema:

- i. Gerar *NARV* árvores, analisando o tipo topológico de cada uma.
- ii. Testar a hipótese, recorrendo a um teste de ajustamento de qui-quadrado, da distribuição observada das árvores geradas pelos vários tipos, seguir uma lei multinomial em que os parâmetros são os valores de probabilidades (quadros 4.1 a 4.4) de cada tipo topológico, para o respectivo processo.
- iii. Repetir i. e ii *NREP* vezes.
- iv. Registrar o número de vezes, nas *NREP* repetições, que o teste de ajustamento conduziu à rejeição da hipótese nula.

Na decisão do teste estatístico foi sempre usado um nível de significância de 5%. Este procedimento foi realizado para os três geradores de números pseudo-aleatórios, contudo os resultados não parecem depender do tipo de gerador usado e por esta razão apresentam-se resultados só para “*RANI*”.

Em seguida apresentam-se alguns quadros¹, numa tentativa de resumir os resultados obtidos para as metodologias “*Geração Uniforme*” e “*Permutação Dupla*”. Seria interessante fazer o mesmo tipo de experiências para a “*Geração com Parâmetro de Forma*”, no entanto, por falta de tempo e porque o objectivo inicial do trabalho era comparar o comportamento dos dois métodos, que teoricamente têm o mesmo desempenho, isso não foi feito. No capítulo 6 a “*Geração com Parâmetro de Forma*” revelou-se de grande utilidade e foi analisado o seu comportamento em simulação mas não da forma aqui apresentada para os outros métodos.

¹ Nos quadros seguintes ni representa o número de vezes que a hipótese nula foi rejeitada, com dimensões amostrais de *NARV*, em *NREP* repetições.

NREP	NARV = 30		NARV = 60		NARV = 102		NARV = 180		NARV = 300	
	ni	%	ni	%	ni	%	ni	%	ni	%
50	0	0.	1	2.	2	4.	2	4.	2	4.
100	1	1.	4	4.	2	2.	5	5.	4	4.
150	3	2.	6	4.	5	3.33	6	4.	7	4.67
250	5	2.	11	4.40	9	3.60	14	5.60	13	5.20
450	11	2.44	18	4.	19	4.22	26	5.78	18	4.
750	18	2.44	28	3.73	34	4.53	43	5.73	27	3.60
1000	25	2.50	33	3.30	46	4.60	59	5.90	37	3.70
5000	140	2.80	194	3.88	226	4.52	246	4.92	218	4.36
10000	297	2.97	360	3.60	469	4.69	501	5.01	426	4.26
50000	1552	3.10	1944	3.89	2266	4.53	2397	4.79	2120	4.24
100000	3202	3.20	3881	3.88	4552	4.55	4829	4.83	4303	4.30
1000000	31274	3.13	38571	3.86	45589	4.56	47486	4.75	42888	4.29

Quadro 4.5 – Geração Uniforme, $m = 4$

NREP	NARV = 30		NARV = 60		NARV = 102		NARV = 180		NARV = 300	
	ni	%	ni	%	ni	%	ni	%	ni	%
50	3	6.	1	2.	1	2	5	10.	0	0.
100	6	6.	3	3.	3	3	7	7.	4	4.
150	10	6.67	6	4.	7	4.67	9	6.	6	4.
250	17	6.8	9	3.6	15	6	16	6.4	9	3.6
450	27	6	24	5.33	24	5.33	20	4.44	22	4.89
750	47	6.27	40	5.33	41	5.47	36	4.8	36	4.8
1000	58	5.8	51	5.1	55	5.5	50	5.	49	4.9
5000	266	5.32	269	5.38	277	5.54	227	4.54	228	4.56
10000	535	5.35	547	5.47	532	5.32	471	4.71	487	4.87
50000	2876	5.75	2668	5.34	2549	5.10	2425	4.85	2576	5.15
100000	5787	5.79	5280	5.28	4946	4.95	4732	4.73	5065	5.07
1000000	56305	5.63	53521	5.35	50171	5.02	48002	4.8	50312	5.03

Quadro 4.6 - Geração Uniforme, $m = 5$

Da análise dos quadros anteriores parece poder concluir-se que o comportamento das duas metodologias de geração de dendrogramas é muito semelhante. Por vezes, para os valores mais baixos de *NREP*, registam-se diferenças que serão devidas ao reduzido número de repetições e ao funcionamento do gerador. Quando *NREP* aumenta estes efeitos desaparecem.

<i>NREP</i>	<i>NARV</i> = 60		<i>NARV</i> = 90		<i>NARV</i> = 105		<i>NARV</i> = 180		<i>NARV</i> = 270	
	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%
50	2	4.	2	4.	2	4.	1	2.	3	6.
100	4	4.	5	5.	5	5.	3	3.	7	7.
150	9	6.	7	4.67	6	4.	6	4.	9	6.
250	13	5.2	11	4.40	9	3.6	11	4.44	14	5.6
450	25	5.56	19	4.22	17	3.78	16	3.56	23	5.11
750	41	5.47	38	5.07	29	3.87	39	5.2	35	4.67
1000	51	5.1	48	4.80	45	4.5	54	5.4	45	4.5
5000	254	5.08	245	4.90	203	4.06	219	4.38	248	4.96
10000	494	4.94	501	5.01	483	4.83	484	4.84	490	4.90
50000	2428	4.86	2510	5.02	2476	4.95	2481	4.96	2503	5.01
100000	4802	4.80	4774	4.77	4918	4.92	5053	5.05	5053	5.05
1000000	48072	4.81	48749	4.87	49377	4.94	49575	4.96	50153	5.02

Quadro 4.7 - Geração Uniforme, *m* = 6

<i>NREP</i>	<i>NARV</i> = 90		<i>NARV</i> = 180		<i>NARV</i> = 270		<i>NARV</i> = 360		<i>NARV</i> = 450	
	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%
50	1	2.	1	2.	2	4.	1	2.	1	2.
100	6	6.	3	3.	4	4.	7	7.	5	5.
150	8	5.33	10	6.67	9	6.	8	5.33	8	5.33
250	13	5.20	14	5.6	13	5.2	16	6.4	14	5.6
450	23	5.11	26	5.78	20	4.44	21	4.67	24	5.33
750	39	5.20	44	5.87	32	4.27	32	4.27	37	4.93
1000	51	5.10	51	5.1	52	5.2	41	4.1	45	4.5
5000	262	5.24	249	4.98	254	5.08	235	4.7	245	4.9
10000	530	5.30	505	5.05	506	5.06	501	5.01	474	4.74
50000	2598	5.20	2627	5.25	2575	5.15	2463	4.93	2463	4.93
100000	5085	5.09	5166	5.17	5093	5.09	5049	5.05	5022	5.02
1000000	51298	5.13	50393	5.04	50507	5.05	50263	5.03	49933	4.99

Quadro 4.8 - Geração Uniforme, *m* = 7

<i>NREP</i>	<i>NARV</i> = 30		<i>NARV</i> = 60		<i>NARV</i> = 102		<i>NARV</i> = 180		<i>NARV</i> = 300	
	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%
50	7	14.	5	2.	1	2.	0	0.	1	2.
100	10	10.	7	3.	2	2.	2	2.	3	3.
150	14	9.33	7	4.	4	2.67	2	1.33	8	5.33
250	18	7.23	11	3.6	9	3.6	6	2.4	13	5.2
450	32	7.11	20	5.33	19	4.22	11	2.44	22	4.89
750	50	6.67	40	5.33	31	4.13	21	2.8	39	5.2
1000	64	6.4	51	5.1	43	4.3	34	3.4	54	5.4
5000	268	5.36	264	5.38	244	4.88	225	4.5	262	5.24
10000	559	5.59	540	5.47	491	4.91	469	4.69	524	5.24
50000	2764	5.53	2606	5.34	2564	5.13	2318	4.63	2516	5.03
100000	5530	5.53	5322	5.28	5046	5.05	4784	4.78	5083	5.08
1000000	56015	5.6	53469	5.35	50210	5.02	47944	4.79	50508	5.05

Quadro 4.9 – Permutação Dupla, *m* = 5

<i>NREP</i>	<i>NARV</i> = 60		<i>NARV</i> = 90		<i>NARV</i> = 105		<i>NARV</i> = 180		<i>NARV</i> = 270	
	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%	<i>ni</i>	%
50	2	4.	1	2.	3	6.	2	4.	3	6.
100	3	3.	2	2.	6	6.	3	3.	3	3.
150	6	4.	5	3.33	10	6.67	5	3.33	5	3.33
250	13	5.2	10	4.	13	5.2	10	4.	9	3.6
450	25	5.5	20	4.45	22	4.89	24	5.33	17	3.78
750	42	5.6	34	4.53	36	4.8	37	4.93	33	4.4
1000	56	5.6	45	4.5	49	4.9	49	4.9	46	4.6
5000	266	5.32	258	5.16	253	5.06	251	5.02	239	4.78
10000	513	5.13	526	5.26	515	5.15	487	4.87	483	4.83
50000	2494	4.99	2474	4.95	2519	5.04	2414	4.83	2468	4.94
100000	4911	4.91	4961	4.96	5015	5.02	4999	5.	4951	4.95
1000000	48081	4.81	49331	4.93	50072	5.01	49336	4.93	49530	4.95

Quadro 4.10 - Permutação Dupla, *m* = 6

Da análise dos quadros anteriores parece poder concluir-se que o comportamento das duas metodologias de geração de dendrogramas é muito semelhante. Por vezes, para os valores mais baixos de *NREP*, registam-se diferenças que serão devidas ao reduzido número de repetições e ao funcionamento do gerador. Quando *NREP* aumenta estes efeitos desaparecem.

Sendo os resultados dos dois processos bastante semelhantes entende-se contudo que o de *Geração Uniforme* tem a seu favor a naturalidade e simplicidade conceptuais.

Note-se ainda que a ideia da metodologia de geração de dendrogramas, comum aos dois processos aqui propostos, permite, através da *Geração com Parâmetro de Forma*, gerar árvores com topologias predominantemente de um tipo previamente fixado.

REFERÊNCIAS

- [1] Furnas, G. W. (1984): The Generation of Random, Binary Unordered Trees, *Journal of Classification*, 1, 187-233.
- [2] Knuth, D. E. (1969): *The Art of Computer Programming, Vol 2 - Seminumerical Algorithms*, Reading, MA: Addison-Wesley.
- [3] Lapointe, F. J., Legendre, P. (1990): A Statistical Framework to Test the Consensus of two Nested Classifications, *Systematic Zoology*, 39, 1-13.
- [4] Lapointe, F.-J., Legendre, P. (1991): The Generation of Random Ultrametric Matrices Representing Dendograms, *Journal of Classification*, 8, 177-200.
- [5] Lukasiewicz, J. (1954): Arithmetic and Modal Logic. *Journal Comput. Systems*, 1, 213-219.
- [6] Lukasiewicz, J. (1963): *Elements of Mathematical Logic*, Pergamon Press, Warszawa, PWN-Polish Scientific Publishers, XI.

CAPÍTULO 5
COMPARAÇÃO ORDINAL DE
ESTRUTURAS CLASSIFICATÓRIAS

CAPÍTULO 5

COMPARAÇÃO ORDINAL DE ESTRUTURAS CLASSIFICATÓRIAS

ÍNDICE

INTRODUÇÃO

5.1 COMPARAÇÃO DE ORDENAÇÕES

5.1.1 Coeficientes de Comparação

5.1.2 Comparação Analítica dos Coeficientes

5.1.3 Análise do Comportamento Assintótico dos Coeficientes

5.1.4 A Distribuição Exacta da Estatística C

5.2 COMPARAÇÃO DE PREORDENAÇÕES

5.2.1 Coeficientes de Comparação

5.3 COMPARAÇÃO DE UMA ORDENAÇÃO E UMA PREORDENAÇÃO

5.3.1 Análise dos Vários Coeficientes

5.3.2 A Distribuição Exacta da Estatística C

5.3.2.1 Preordenação Associada a uma Cadeia Pura

5.3.2.2 Preordenação Associada a uma Árvore Equilibrada

REFERÊNCIAS

CAPÍTULO 5

COMPARAÇÃO ORDINAL DE ESTRUTURAS CLASSIFICATÓRIAS

INTRODUÇÃO

Um problema corrente em classificação é a comparação de estruturas classificatórias. Neste capítulo analisam-se as situações de maior interesse para o contexto deste trabalho.

Na comparação de estruturas classificatórias podem ser usados dois princípios bem diferentes que recorrem a níveis de informação distintos: a abordagem métrica e a abordagem ordinal. Na abordagem métrica são usados os valores reais associados às estruturas a comparar, enquanto que na abordagem ordinal apenas são utilizadas as ordenações desses valores numéricos. Na primeira situação a informação usada é mais rica, contudo, sendo ela consequência das escolhas necessárias dos índices, é importante pensar de que forma ela é afectada por mudança de índices. A abordagem ordinal sendo menos informativa é mais robusta. São frequentes as situações em que a utilização de índices diferentes conduz naturalmente a valores diferentes, mas a ordenações iguais.

Este capítulo, dedicado à abordagem ordinal, tem uma estrutura por secções, de acordo com o tipo de estruturas a comparar; na primeira secção trata da comparação de duas ordenações, na segunda de duas preordenações e na terceira de uma ordenação com uma preordenação.

Nas secções deste capítulo há necessidade de trabalhar com os conjuntos de dois elementos de E , designado por F , e ainda com os pares de elementos distintos de F . Embora na introdução se use a notação correcta, para a representação destes pares e a partir de 5.16, para aligeirar as expressões e facilitar a leitura, usar-se-á a notação $((a,b),(c,d))$ em vez de $(\{a,b\},\{c,d\})$. Assim o par (a,b) denota na realidade o conjunto $\{a,b\}$.

Esta opção tem ainda a vantagem da aproximação à notação usada por grande parte dos especialistas nesta área. A título de exemplo, a notação correcta para 5.16 seria,

$$gr(O) = \{(\{x, y\}, \{z, t\}) \in F \times F : \{x, y\} \neq \{z, t\} \text{ e } \{x, y\} < \{z, t\} \text{ para } O\}.$$

Trata-se de uma solução de compromisso, em que a precisão de notação é sacrificada à legibilidade e simplificação de leitura.

De acordo com a notação introduzida no capítulo 2, seja E o conjunto a classificar, cujo cardinal é m , e F o conjunto dos pares de elementos distintos de E ,

$$F = \{\{x, y\} : x, y \in E \wedge x \neq y\}, \quad (5.1)$$

donde,

$$card(F) = M = \frac{m(m-1)}{2}, \quad (5.2)$$

em que $card$ é a função cardinal de um conjunto.

A necessidade de comparar estruturas em *C.H.A.* pode surgir de diferentes formas:

- Comparar duas funções de comparação entre pares de elementos de E .
- Comparar uma função de comparação entre pares de elementos de E e uma hierarquia sobre E .
- Comparar duas hierarquias sobre a mesma população.
- Comparar duas partições de E .

A abordagem métrica é possível para a resolução dos três primeiros destes problemas, já que o último não tem uma resolução métrica possível. Esta abordagem métrica é simples e não depende do tipo de estrutura em comparação, mas tem de ter óbvias salvaguardas face a um problema de escala.

A escolha de uma função de comparação, γ , sobre os elementos de E , dá origem a um vector de $\mathfrak{R}^M$.

Como resultado de uma C.H.A., após ter escolhido a função de comparação entre elementos e a função de comparação entre partes de E , obtém-se uma hierarquia indiciada de partições. O índice de níveis permite definir uma distância ultramétrica d sobre E dada por, $d(x, y) = d_i$, em que i é o menor nível para o qual x e y pertencem á mesma classe. Deste modo associado a uma hierarquia indiciada há também um vector de $\mathfrak{R}^M$.

Os dois aspectos anteriores fazem com que a comparação de duas estruturas se faça a partir da comparação de dois vectores reais de igual dimensão. Introduzindo uma métrica em $\mathfrak{R}^M$, podemos medir a distância entre os dois vectores, quer eles estejam associados a duas funções de comparação entre elementos sobre o mesmo conjunto E , quer estejam associados a duas hierarquias dos elementos de E , ou ainda a uma função de comparação e a uma hierarquia dos elementos de E .

O problema da comparação de duas estruturas assume assim um aspecto muito simples, o do cálculo da distância entre dois vectores, sejam X e Y de $\mathfrak{R}^M$. Muitas são as distâncias que podem ser definidas, de uma forma geral pode referir-se a distância de Minkowsky, que engloba algumas das mais usadas, cuja expressão é dada por:

$$D(X, Y) = \left[\sum_{i=1}^M \left| \frac{X_i}{\|X\|_D} - \frac{Y_i}{\|Y\|_D} \right|^\lambda \right]^{1/\lambda} \quad (5.3)$$

No cálculo da distância os vectores são normalizados, para que a distância permaneça independente do factor escala de medida dos dois vectores intervenientes.

Algumas distâncias bem conhecidas são obtidas de (5.3) para valores particulares de λ :

- Distância city-block, para $\lambda = 1$
- Distância euclidiana, para $\lambda = 2$
- Distância do máximo, para $\lambda = \infty$

Uma alternativa para quantificar a semelhança entre dois vectores reais de igual dimensão é o *coeficiente de correlação empírico*. Dados dois vectores X e Y de $\mathbb{R}^M$, o referido coeficiente para os pares (X_i, Y_i) com $i = 1, \dots, M$, é dado por:

$$R = \frac{\frac{1}{M} \sum_{i=1}^M (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{M} \sum_{i=1}^M (X_i - \bar{X})^2} \sqrt{\frac{1}{M} \sum_{i=1}^M (Y_i - \bar{Y})^2}}. \quad (5.4)$$

No caso em que X é o vector associado a uma função de comparação de pares de elementos e Y é o vector associado a uma hierarquia sobre os mesmos elementos, este coeficiente foi chamado por Sokal [10] e Sokal & Sneath [11] de *coeficiente de correlação cofenético*.

Foi já visto que a escolha de um coeficiente de comparação, γ , sobre os elementos de E , define sobre F uma preordem total, $\leq$. Isto significa que, se γ for uma função do tipo semelhança, temos:

$$\forall (\{x, y\}, \{z, t\}) \in F \times F : \{x, y\} \leq \{z, t\} \Leftrightarrow \gamma(x, y) \geq \gamma(z, t). \quad (5.5)$$

Esta preordem sobre F é a preordenação sobre E associada a γ , que será representada por ω_γ . Esta preordenação por ser muito fina é frequentemente tratada como uma ordenação, O_γ . Neste trabalho antes de proceder à comparação de estruturas associadas a funções de comparação, é feita uma análise do tipo de estrutura em causa.

Uma hierarquia H para os elementos de E dá sempre origem a uma preordem ω_H sobre F .

Seja $P(m; m_1, m_2, \dots, m_k)$ o conjunto das partições de E do tipo $(m_1, m_2, \dots, m_k)$, ou seja, partições com k classes de cardinais $m_1, m_2, \dots, m_k$ em que $m_1 + m_2 + \dots + m_k = m$. Dada uma partição P deste tipo, ela define em F uma partição em duas classes: $S(P)$ e $R(P)$, respectivamente conjunto dos pares separados e conjunto dos pares reunidos, isto é:

$R(P)$ - conjunto dos pares de elementos de E , em que as duas componentes estão na mesma classe de P .

$S(P)$ - conjunto dos pares de elementos de E , em que as duas componentes estão em duas classes distintas de P .

Para uma partição P de E do tipo $(m_1, m_2, \dots, m_k)$ facilmente se conclui que:

$$card(R(P)) = \sum_{1 \leq i \leq k} \frac{m_i(m_i - 1)}{2} \quad (5.6)$$

e

$$card(S(P)) = \sum_{i < j} m_i m_j, \quad (5.7)$$

sendo $card(R(P)) + card(S(P)) = card(F) = M$.

Uma partição P define em $F \times F$ uma preordenação sobre E através do produto $S(P) \times R(P)$ ou $R(P) \times S(P)$.

Pelo exposto facilmente se conclui que os diferentes tipos de comparações, envolvendo funções de comparação, hierarquias ou partições, podem ser efectuadas pela comparação de duas ordenações, duas preordenações ou uma ordenação e uma preordenação.

5.1 COMPARAÇÃO DE ORDENAÇÕES

Sejam O_1 e O_2 duas ordenações sobre E , a comparar. Definam-se as variáveis aleatórias ordinais U e V dadas por:

$U_k =$ "Número de ordem do k -ésimo elemento de O_1 "

$V_k =$ "Número de ordem do k -ésimo elemento de O_2 ",

com $k = 1, \dots, M$ e considere-se a variável aleatória A função de U e V :

$$A_{ij} = \text{sgn}(U_j - U_i) \text{sgn}(V_j - V_i), \quad (5.8)$$

com $1 \leq i < j \leq M$, em que sgn representa a função "sinal de".

Se $A_{ij} = 1$, os pares (U_i, V_i) e (U_j, V_j) dizem-se concordantes.

Se $A_{ij} = -1$, os pares (U_i, V_i) e (U_j, V_j) dizem-se discordantes.

Se $A_{ij} = 0$ nos pares (U_i, V_i) e (U_j, V_j) ocorre empate¹.

Considerem-se ainda as variáveis aleatórias C e D :

$C =$ "Número de concordâncias entre as variáveis aleatórias U e V "

$=$ "Número de vezes que $A_{ij} = 1$, para $1 \leq i < j \leq M$ "

$D =$ "Número de discordâncias entre as variáveis aleatórias U e V "

$=$ "Número de vezes que $A_{ij} = -1$, para $1 \leq i < j \leq M$ ".

Existem vários coeficientes estatísticos para medir o grau de semelhança ou de concordância entre duas ordenações. Neste ponto podemos referir alguns trabalhos clássicos em que estes assuntos são desenvolvidos teoricamente, Kendall [6] e Kendall & Stuart [4].

5.1.1 Coeficientes de Comparação

O coeficiente de correlação ordinal de Spearman para ordenações, R_S , mede a sua concordância e é dado por:

$$R_S = \frac{\frac{1}{M} \sum_{k=1}^M (U_k - \bar{U})(V_k - \bar{V})}{\sqrt{\frac{1}{M} \sum_{i=1}^M (U_k - \bar{U})^2} \sqrt{\frac{1}{M} \sum_{i=1}^M (V_k - \bar{V})^2}} \quad (5.9)$$

Facilmente se percebe que este coeficiente R_S é o coeficiente de correlação empírico de (5.4) para as variáveis aleatórias ordinais U e V .

Alguma manipulação algébrica permite dar a (5.9) um aspecto mais simples:

$$R_S = \frac{12}{M(M^2 - 1)} \sum_{1 \leq k \leq M} U_k V_k - \frac{3(M+1)}{M-1}, \quad (5.10)$$

ou ainda

$$R_S = 1 - \frac{6S(d^2)}{M(M^2 - 1)}, \quad (5.11)$$

em que

$$S(d^2) = \sum_{1 \leq k \leq M} (U_k - V_k)^2. \quad (5.12)$$

A distribuição exacta de R_S é difícil de deduzir, está tabelada para valores de $M \leq 10$, é simétrica e converge, embora lentamente para a distribuição normal. Em Soromenho [12], por exemplo, é demonstrado que a estatística R_S tem distribuição assintótica $N\left(0; \frac{1}{M-1}\right)$. A aproximação à distribuição normal só é razoável para valores de $M > 20$ e é aconselhável fazer

¹ No caso de duas ordenações temos que $A_{ij} \neq 0, \forall 1 \leq i < j \leq M$.

correção de continuidade, retirando uma unidade, a $S(d^2)^1$. Para valores de $11 \leq M \leq 20$ é necessário arranjar um processo intermédio para calcular a distribuição.

No cálculo do coeficiente Tau de Kendall, que será representado por T_K , intervêm apenas as posições relativas dos valores de U e V ,

$$T_K = \frac{C - D}{\frac{1}{2}M(M-1)}. \quad (5.13)$$

Em Kendall [5] é apresentado o estudo assintótico de T_K e demonstrado que, admitindo independência entre as duas ordenações O_1 e O_2 , T_K segue uma lei assintótica normal com

$$E(T_K) = 0 \quad (5.14)$$

e

$$V(T_K) = 2(2M+5)/9M(M-1). \quad (5.15)$$

A aproximação de T_K à lei normal é habitualmente usada para $M \geq 10$, podendo ser acompanhada da respectiva correção de continuidade. O numerador de (5.13) é uma variável aleatória discreta em que os valores consecutivos do domínio estão separados de duas unidades. A correção de continuidade é de uma unidade em $C - D$, associando a massa probabilística de um dado valor t a um intervalo $]t-1, t+1[$.

De uma forma geral a uma ordenação O pode associar-se um grafo¹, $gr(O)$, dado por:

$$gr(O) = \{((x, y), (z, t)) \in F \times F : (x, y) \neq (z, t) \text{ e } (x, y) < (z, t) \text{ para } O\}. \quad (5.16)$$

O grafo assim obtido está definido em $F \times F - \Delta$, em que Δ é o conjunto de pares de elementos iguais de F , isto é:

$$\Delta = \{((x, y), (x, y)) : (x, y) \in F\}. \quad (5.17)$$

¹ Para valores de $S(d^2)$ grandes a correção de continuidade pode ser omitida.

Com esta definição tem-se que $\text{card}(\text{gr}(O)) = M(M-1)/2$.

Dadas duas ordenações, O_1 e O_2 sobre o mesmo conjunto, E , uma medida natural da concordância entre O_1 e O_2 é $T_{\text{Ler}} = \text{card}(\text{gr}(O_1) \cap \text{gr}(O_2))$, que representa o número de pares ordenados da mesma forma por O_1 e O_2 . Lerman [7] (pág. 123 e seguintes) demonstrou que este coeficiente tem distribuição assintótica normal com:

$$E(T_{\text{Ler}}) = \frac{M(M-1)}{4} \quad (5.18)$$

e

$$V(T_{\text{Ler}}) = \frac{1}{72} M(M-1)(2M+5). \quad (5.19)$$

Na demonstração do resultado apresentado, Lerman usa a abordagem de Haigh [3], recorrendo à função geradora de momentos da soma de variáveis aleatórias independentes e ao teorema de Lindeberg-Feller de convergência para a normal.

Em Soromenho [12], é apresentada uma demonstração alternativa do mesmo resultado, sendo o cálculo das expressões do valor médio e da variância de T_{Ler} aí proposto mais simples que os propostos por Lerman. Para mostrar a convergência para a lei normal recorre ao teorema de Lindberg-Lévy.

No ponto 5.1.4 será deduzida a lei exacta da estatística C que coincide com T_{Ler} .

Em Nicolau [9] são propostos novos coeficientes para medir a concordância entre duas ordenações, com princípios de base um pouco diferentes e por essa razão não são aqui apresentados.

5.1.2 Comparação Analítica dos Vários Coeficientes

Os coeficientes R_S e T_K assumem valores no intervalo $[-1,1]$ e estamos já familiarizados com a interpretação, em termos de grau e tipo de associação, dos seus valores amostrais. O

¹ Definido em 2.1.4.

mesmo não acontece com o coeficiente de Lerman anteriormente apresentado, cuja variação é não limitada¹.

Pelo exposto no ponto anterior sabe-se que os vários coeficientes de comparação de ordenações aqui referidos têm distribuição assintótica normal. Segue-se uma comparação dos vários coeficientes, após respectiva centragem na média e redução para variância unitária.

Considerem-se os novos coeficientes:

$$R_S^* = \frac{R_S - 0}{\sqrt{\frac{1}{M-1}}} = \sqrt{M-1} R_S \quad (5.20)$$

$$T_K^* = \frac{T_K - 0}{\sqrt{\frac{2(2M+5)}{9M(M-1)}}} = \sqrt{\frac{9M(M-1)}{2(2M+5)}} T_K \quad (5.21)$$

$$T_{Ler}^* = \frac{T_{Ler} - \frac{M(M-1)}{4}}{\sqrt{\frac{1}{72} M(M-1)(2M+5)}} = \frac{12T_{Ler} - 3M(M-1)}{\sqrt{2M(M-1)(2M+5)}} \quad (5.22)$$

A variável C do numerador de (5.13) coincide com T_{Ler} e, pelo facto da comparação ser entre ordenações, sabe-se que $C + D = M(M-1)/2$, o que permite escrever (5.13) na forma,

$$T_K = \frac{4C}{M(M-1)} - 1 \quad (5.23)$$

Tendo em consideração estes aspectos e denotando o número de pares concordantes, entre os M , por C pode escrever-se:

¹ Esse coeficiente pode assumir valores muito elevados quando M aumenta.

$$T_K^* = \frac{12C - 3M(M-1)}{\sqrt{2M(M-1)(2M+5)}} \quad (5.24)$$

$$T_{Ler}^* = \frac{12C - 3M(M-1)}{\sqrt{2M(M-1)(2M+5)}}, \quad (5.25)$$

permitindo concluir que T_K^* e T_{Ler}^* coincidem.

5.1.3 Comportamento Assintótico dos Coeficientes

No ponto anterior definiram-se os coeficientes de comparação de ordenações R_S^* , T_K^* e T_{Ler}^* e referiu-se que todos têm distribuição assintótica $N(0,1)$. Neste ponto serão analisadas as convergências para essas distribuições, recorrendo à simulação de pares de ordenações e consequente cálculo dos diferentes coeficientes.

Nas simulações aqui referidas M representa o cardinal de F , conjunto onde as ordens estão definidas, e $Nord$ representa o número de pares de ordenações geradas. Foram realizadas simulações para valores de M e $Nord$ de acordo com o seguinte:

- $M = \frac{m(m-1)}{2}$, para $m = 4, \dots, 20(1)$ e $m = 20, \dots, 50(5)$, num total de 23 valores.
- $Nord$ entre 20 e 2500 num total de 16 valores.

Os quadros 5.1 e 5.2 apresentam os valores de médias e de variâncias, dos vários coeficientes de comparação de ordenações após transformação, obtidos em algumas das simulações realizadas.

A análise desses quadros revela um comportamento idêntico de R_S^* e de T_K^* ou T_{Ler}^* , e de uma forma geral pode afirmar-se que, quando M aumenta e para valores suficientemente grandes de $Nord$, há uma tendência para os valores médios se aproximarem de zero e as variâncias da unidade.

NORD	M = 6		M = 10		M = 21		M = 36		M = 55		M = 105		M = 595		M = 1225	
	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.
20	-.057	.855	-.515	1.196	-.192	1.087	-.184	.985	-.180	1.047	-.059	.766	-.360	.593	.013	.660
30	-.055	.969	-.234	1.236	-.190	1.034	.041	1.165	-.223	.971	.016	.901	-.319	.852	-.031	.653
50	-.125	.963	-.228	1.158	-.033	1.197	.027	.968	-.131	.920	.089	1.035	-.074	1.018	-.023	.608
100	-.148	.992	-.052	1.160	-.033	1.060	.011	.993	-.033	1.172	.001	.930	.017	.823	-.021	.617
150	-.122	1.004	.018	1.100	.029	1.122	.082	1.114	-.031	1.101	.075	.894	.045	.858	-.119	.754
200	-.024	1.048	-.015	1.051	.024	1.115	.137	1.134	-.022	1.039	.075	.953	.043	.818	-.165	.831
500	-.037	.969	-.043	.981	.016	1.028	.074	1.086	-.047	.956	.013	.949	-.010	.924	-.046	.967
1000	-.013	.982	.035	1.018	.030	1.028	.043	1.027	.007	.942	-.006	.967	-.036	.984	-.024	.982
2500	.021	.983	.006	.986	.003	.992	.009	1.034	.011	.958	.012	.955	-.014	1.025	-.010	.995

Quadro 5.1 – Valores médios e variâncias obtidos para R_S^*

NORD	M = 6		M = 10		M = 21		M = 36		M = 55		M = 105		M = 595		M = 1225	
	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.	Med.	Var.
20	-.038	.923	-.492	1.283	-.178	1.122	-.193	.957	-.162	1.074	-.038	.740	-.359	.600	.011	.660
30	-.038	1.032	-.221	1.282	-.181	1.013	.041	1.196	-.214	.986	.034	.883	-.319	.868	-.033	.658
50	-.098	1.014	-.204	1.213	-.010	1.218	.027	1.018	-.118	.947	.092	1.003	-.069	1.022	-.025	.613
100	-.120	1.018	-.029	1.220	.022	1.082	.017	1.002	-.034	1.170	.004	.921	.017	.827	-.023	.623
150	-.100	1.055	.023	1.136	.033	1.128	.080	1.111	-.040	1.099	.075	.885	.046	.860	-.121	.760
200	-.011	1.100	-.011	1.066	.029	1.116	.139	1.123	-.027	1.044	.077	.941	.043	.817	-.166	.834
500	-.040	.997	-.053	.994	.008	1.018	.081	1.082	-.047	.954	.011	.947	-.010	.923	-.045	.967
1000	-.018	.997	.031	1.026	.024	1.031	.048	1.029	.007	.949	-.009	.965	-.036	.982	-.024	.980
2500	.016	.986	.006	.997	.002	.996	.011	1.039	.011	.961	.010	.954	-.014	1.025	-.009	.994

Quadro 5.2 – Valores médios e variâncias obtidos para $T_K^* \equiv T_{Ler}$

Com o objectivo de quantificar o afastamento entre os valores obtidos na simulação e os esperados tendo por base os resultados assintóticos, para os coeficientes R_S^* , T_K^* e T_{Ler}^* , referidos no ponto anterior, elaborou-se um procedimento que passamos a descrever.

Genericamente seja X a variável aleatória “Valor do coeficiente de comparação entre duas ordenações de comprimento M ”. Fixado M , para cada um dos coeficientes e para cada um dos valores de $Nord$, obtém-se uma amostra para X de dimensão $Nord$. De acordo com a hipótese de distribuição assintótica que estamos a considerar para X , com valores de M suficientemente grandes, temos:

$$X_i \cong N(0,1), \text{ com } i = 1, \dots, Nord. \quad (5.26)$$

Admitindo que as variáveis aleatórias X_i são independentes facilmente se conclui que, em média, $(1 - \alpha) \times 100\%$ dos valores observados da média amostral, $\bar{X}$, pertencem ao intervalo

$$\left] -\frac{c}{\sqrt{Nord}}, \frac{c}{\sqrt{Nord}} \right], \quad (5.27)$$

onde $c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ e em que Φ é a função distribuição da $N(0,1)$.

Analogamente podemos concluir que, em média, $(1 - \alpha) \times 100\%$ dos valores observados da variância amostral, S^2 , pertencem ao intervalo

$$\left] c_1 \frac{Nord + 1}{Nord^2}, c_2 \frac{Nord + 1}{Nord^2} \right], \quad (5.28)$$

em que c_1 e c_2 são as abcissas onde a função distribuição da lei qui-quadrado com $Nord - 1$ graus de liberdade atinge, respectivamente, $\left(\frac{\alpha}{2}\right)$ e $\left(1 - \frac{\alpha}{2}\right)$ de probabilidade.

Dada a aproximação da distribuição qui-quadrado à normal, podemos ainda dar outra forma ao intervalo anterior, para valores de $Nord > 100$,

$$\left] \frac{1}{Nord} \left(-c\sqrt{2Nord-2} + Nord - 1 \right), \frac{1}{Nord} \left(c\sqrt{2Nord-2} + Nord - 1 \right) \right], \quad (5.29)$$

sendo $c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

Para os diferentes valores de $Nord$ e considerando $1 - \alpha = 0.95$, foram determinados os intervalos dados por (5.27) e (5.28) ou (5.29). Foram apenas consideradas as simulações em que $M > 20$, num total de 20 valores diferentes de M . Para cada valor de $Nord$ e de M foram calculadas a média e a variância amostrais correspondentes, para os coeficientes de comparação R_S^* , T_K^* e T_{Ler}^* , e analisada a pertença ou não ao respectivo intervalo de probabilidade 0.95.

No conjunto total das situações¹ verificou-se que:

- Para todos os valores de M e $Nord$, os três coeficientes tiveram o mesmo comportamento, isto é, não ocorreu nenhuma situação em que o valor da média ou da variância de um coeficiente, não pertencia ao intervalo calculado e de outro coeficiente pertencia.
- Para cada um dos coeficientes em causa apenas em 13 dos casos, correspondendo a 4%, a média não pertencia ao referido intervalo.
- Para cada um dos coeficientes, também para a variância o número de casos em que a variância não pertencia ao intervalo correspondente foi de 13².

Como síntese das simulações efectuadas para estes três coeficientes de comparação, podemos dizer que o comportamento é bastante semelhante e os resultados obtidos estão de acordo com a teoria das suas distribuições apresentada no ponto anterior.

¹ O número total de situações é igual a 320, já que foram considerados 16 valores de $Nord$ e 20 de M .

² Apesar da frequência ser a mesma na média e na variância, as situações que lhe deram origem não são as mesmas.

Com o objectivo de quantificar o afastamento entre os valores obtidos na simulação e os esperados tendo por base os resultados assintóticos, para os coeficientes R_S^* , T_K^* e T_{Ler}^* , referidos no ponto anterior, elaborou-se um procedimento que passamos a descrever.

Genericamente seja X a variável aleatória “Valor do coeficiente de comparação entre duas ordenações de comprimento M ”. Fixado M , para cada um dos coeficientes e para cada um dos valores de $Nord$, obtém-se uma amostra para X de dimensão $Nord$. De acordo com a hipótese de distribuição assintótica que estamos a considerar para X , com valores de M suficientemente grandes, temos:

$$X_i \cong N(0,1), \text{ com } i = 1, \dots, Nord. \quad (5.26)$$

Admitindo que as variáveis aleatórias X_i são independentes facilmente se conclui que, em média, $(1 - \alpha) \times 100\%$ dos valores observados da média amostral, $\bar{X}$, pertencem ao intervalo

$$\left] -\frac{c}{\sqrt{Nord}}, \frac{c}{\sqrt{Nord}} \right], \quad (5.27)$$

onde $c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ e em que Φ é a função distribuição da $N(0,1)$.

Analogamente podemos concluir que, em média, $(1 - \alpha) \times 100\%$ dos valores observados da variância amostral, S^2 , pertencem ao intervalo

$$\left] c_1 \frac{Nord + 1}{Nord^2}, c_2 \frac{Nord + 1}{Nord^2} \right], \quad (5.28)$$

em que c_1 e c_2 são as abcissas onde a função distribuição da lei qui-quadrado com $Nord-1$ graus de liberdade atinge, respectivamente, $\left(\frac{\alpha}{2}\right)$ e $\left(1 - \frac{\alpha}{2}\right)$ de probabilidade.

Dada a aproximação da distribuição qui-quadrado à normal, podemos ainda dar outra forma ao intervalo anterior, para valores de $Nord > 100$,

$$\left] \frac{1}{Nord} \left(-c\sqrt{2Nord-2} + Nord - 1 \right), \frac{1}{Nord} \left(c\sqrt{2Nord-2} + Nord - 1 \right) \right], \quad (5.29)$$

sendo $c = \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$.

Para os diferentes valores de $Nord$ e considerando $1 - \alpha = 0.95$, foram determinados os intervalos dados por (5.27) e (5.28) ou (5.29). Foram apenas consideradas as simulações em que $M > 20$, num total de 20 valores diferentes de M . Para cada valor de $Nord$ e de M foram calculadas a média e a variância amostrais correspondentes, para os coeficientes de comparação R_S^* , T_K^* e T_{Ler}^* , e analisada a pertença ou não ao respectivo intervalo de probabilidade 0.95.

No conjunto total das situações¹ verificou-se que:

- Para todos os valores de M e $Nord$, os três coeficientes tiveram o mesmo comportamento, isto é, não ocorreu nenhuma situação em que o valor da média ou da variância de um coeficiente, não pertencia ao intervalo calculado e de outro coeficiente pertencia.
- Para cada um dos coeficientes em causa apenas em 13 dos casos, correspondendo a 4%, a média não pertencia ao referido intervalo.
- Para cada um dos coeficientes, também para a variância o número de casos em que a variância não pertencia ao intervalo correspondente foi de 13².

Como síntese das simulações efectuadas para estes três coeficientes de comparação, podemos dizer que o comportamento é bastante semelhante e os resultados obtidos estão de acordo com a teoria das suas distribuições apresentada no ponto anterior.

¹ O número total de situações é igual a 320, já que foram considerados 16 valores de $Nord$ e 20 de M .

² Apesar da frequência ser a mesma na média e na variância, as situações que lhe deram origem não são as mesmas.

5.1.4 A Distribuição Exacta da Estatística C

A comparação analítica, feita em 5.1.2, entre os vários coeficientes aqui apresentados, revela a importância da estatística C como medida de comparação entre duas ordenações de igual comprimento.

Em 5.1.1 foi já referida a distribuição assintótica de C^1 , bem como as expressões genéricas da média e da variância. Em seguida deduz-se a distribuição exacta de C .

Sejam O_1 e O_2 as duas ordenações a comparar, de comprimento M , e U e V as variáveis ordinais associadas, definidas em 5.1.

Por simplificação e sem perda de generalidade, considera-se que U é a sequência natural, isto é $u = (1, 2, \dots, M)$ e que $v = (v_1, v_2, \dots, v_M)$, sendo os elementos de v uma permutação dos elementos de u .

Neste caso a variável aleatória A função de U e V definida por (5.8) assume um aspecto mais simples:

$$A_{ij} = \text{sgn}(V_j - V_i), \text{ para } 1 \leq i < j \leq M \quad (5.30)$$

Assim se $V_j > V_i$ ocorre uma concordância e quando $V_j < V_i$ ocorre uma discordância.

Considere-se a estatística $C_M = \text{"número de concordâncias entre } O_1 \text{ e } O_2\text{"}$, que coincide com o número de vezes que a variável aleatória A assume o valor 1.

Seja ainda $E_k = \text{"número de concordâncias do elemento } k \text{ de } O_2 \text{ com os elementos que estão à sua esquerda"}$, que coincide com o número de vezes que A_{ik} assume o valor 1, com $1 \leq i < k$ e $2 \leq k < M$.

Pode então escrever-se:

¹ A estatística C coincide com o coeficiente de Lerman

$$C_M = \sum_{k=2}^M E_k . \quad (5.31)$$

Algumas propriedades associadas às variáveis aleatórias C_M e E_k podem ser referidas, para as quais não serão apresentadas demonstrações, porque envolvem noções de estatística e de análise combinatória elementares.

P1 - Cada variável aleatória E_k é uniforme em $\{0,1,\dots,k-1\}$.

P2 - As $(M-1)$ variáveis E_k são independentes.

P3 - $C_M = C_{M-1} + E_M$ ¹.

P4 - O domínio da variável aleatória C_M é $\{0,1,\dots,M(M-1)/2\}$.

No que se segue usar-se-á a seguinte notação, com $k = 1,2,\dots,M$:

f_k - função de probabilidade da variável aleatória C_k

F_k - função de distribuição da variável aleatória C_k

g_k - função de probabilidade da variável aleatória E_k .

Pretende-se obter uma expressão geral para a função $f_M(c)$ para $c = 0,1,\dots,M(M-1)/2$.

Usando a propriedade *P3* referida acima, facilmente se percebe que $f_M(c)$ será dada pela soma das probabilidades dos acontecimentos em que a soma do valor da variável C_{M-1} com o valor da variável E_M é igual a c .

Recorrendo a *P1* e *P4*, que definem os domínios de C_{M-1} e de E_M , e à noção de decomposição de um inteiro em duas parcelas inteiras não negativas com variação limitada, pode deduzir-se a expressão geral de f_M .

¹ O número de concordâncias entre duas ordenações de comprimento M é igual à soma do número de concordâncias das ordenações obtidas eliminando os elementos que ocupam a posição M em O_1 e O_2 , C_{M-1} , com o número de concordâncias, à esquerda, associadas aos elementos de ordem M , E_M .

Na dedução que se segue considera-se $M \geq 4$, para simplificar as expressões de variação associadas a cada ramo. Para valores de $M < 4$ as funções de probabilidade são de obtenção imediata e são as seguintes:

c	0	1
$f_2(c)$	1/2	1/2

Quadro 5.3 – Função de probabilidade de C_2

c	0	1	2	3
$f_3(c)$	1/6	2/6	2/6	1/6

Quadro 5.4 – Função de probabilidade de C_3

A função f_M , será definida por três ramos¹:

1º ramo: $0 \leq c \leq M - 1$

$$\begin{aligned} f_M(c) &= \sum_{i=0}^c P(C_{M-1} = i, E_M = c - i) = \\ &= \sum_{i=0}^c f_{M-1}(i) g_M(c - i) = \\ &= \sum_{i=0}^c f_{M-1}(i) \frac{1}{M} = \\ &= \frac{1}{M} F_{M-1}(c). \end{aligned}$$

2º ramo: $M \leq c \leq (M - 1)(M - 2)/2$

Neste ramo ocorrem limitações no conjunto de valores que podem ser assumidos pela variável E_M . C_{M-1} pode assumir valores inteiros até $(M - 1)(M - 2)/2$ contudo o maior valor de E_M é $M - 1$.

¹ Para $M \geq 4$, C_{M-1} e E_M podem assumir valores até $M - 1$, já que $\min((M - 1)(M - 2)/2, M - 1) = M - 1$.

$$\begin{aligned}
 f_M(c) &= \sum_{i=c-(M-1)}^c P(C_{M-1} = i, E_M = c - i) = \\
 &= \sum_{i=c-(M-1)}^c f_{M-1}(i) g_M(c - i) = \\
 &= \sum_{i=c-M+1}^c f_{M-1}(i) \frac{1}{M} = \\
 &= \frac{1}{M} [F_{M-1}(c) - F_{M-1}(c - M)].
 \end{aligned}$$

3º ramo: $(M-1)(M-2)/2 + 1 \leq c \leq M(M-1)/2$

$$\begin{aligned}
 f_M(c) &= \sum_{i=c-(M-1)}^{M(M-1)/2} P(C_{M-1} = i, E_M = c - i) = \\
 &= \sum_{i=c-(M-1)}^{M(M-1)/2} f_{M-1}(i) g_M(c - i) = \\
 &= \sum_{i=c-M+1}^{M(M-1)/2} f_{M-1}(i) \frac{1}{M} = \\
 &= \frac{1}{M} [1 - F_{M-1}(c - M)].
 \end{aligned}$$

Pode então escrever-se:

$$f_M(c) = \begin{cases} \frac{1}{M} F_{M-1}(c) & \text{se } 0 \leq c \leq M-1 \\ \frac{1}{M} [F_{M-1}(c) - F_{M-1}(c - M)] & \text{se } M \leq c \leq (M-1)(M-2)/2 \\ \frac{1}{M} [1 - F_{M-1}(c - M)] & \text{se } (M-1)(M-2)/2 + 1 \leq c \leq M(M-1)/2 \end{cases} \quad (5.32)$$

A expressão genérica da função de probabilidade de C_M escreve-se à custa da função de distribuição de C_{M-1} . Para um dado valor de c o cálculo de $f_M(c)$ é feito por recorrência e usando a função de distribuição F_3 , obtida a partir de f_3 definida no quadro 5.4.

Resultado 5.1

A função de probabilidade da variável aleatória C_M , de domínio $\{0,1,...,M(M-1)/2\}$ é uma função simétrica .

Demonstração:

Para $M = 2$ ou $M = 3$ é óbvio pela análise dos quadros 5.3 e 5.4.

Nesta demonstração usar-se-á o princípio de indução matemática, provando inicialmente que o resultado é válido para $M = 4$ e em seguida que, para todo o número natural $M > 4$, se f_{M-1} for simétrica então também f_M o é.

Se $M = 4$ o domínio de C_M é $\{0,1,...,6\}$. Recorrendo a (5.32) e à função f_3 definida no quadro 5.4 determina-se f_4 :

c	0	1	2	3	4	5	6
$f_4(c)$	1/24	3/24	5/24	6/24	5/24	3/24	1/24

Quadro 5.5 – Função de probabilidade de C_4

Donde se constata que f_4 é simétrica.

Considere-se agora $M > 4$, M inteiro, e as seguintes hipótese e tese de indução:

Hipótese: A variável aleatória C_{M-1} = "Número de concordâncias entre duas ordenações de comprimento $M-1$, sobre o mesmo conjunto", de domínio $\{0,1,...,(M-1)(M-2)/2\}$, tem uma função de probabilidade simétrica, isto é:

$$f_{M-1}(c) = f_{M-1}\left(\frac{(M-1)(M-2)}{2} - c\right), \text{ para } c = 0,1,...,\text{int}\left(\frac{(M-1)(M-2)}{4}\right), \tag{5.33}$$

onde *int* representa a função "maior inteiro contido em".

Tese: A função de probabilidade da variável aleatória C_M , de domínio $\{0,1,...,M(M-1)/2\}$ e definida por (5.32) é uma função simétrica, isto é:

$$f_M(c) = f_M\left(\frac{M(M-1)}{2} - c\right), \text{ para } c = 0,1,...,\text{int}\left(\frac{M(M-1)}{4}\right). \quad (5.34)$$

1º ramo: $0 \leq c \leq M-1$

$$f_M(c) = \frac{1}{M} F_{M-1}(c) = \sum_{i=0}^c f_{M-1}(i), \quad (5.35)$$

por outro lado,

$$0 \leq c \leq M-1 \Leftrightarrow \frac{(M-1)(M-2)}{2} \leq \frac{M(M-1)}{2} - c \leq \frac{M(M-1)}{2}. \quad (5.36)$$

Donde,

$$\begin{aligned} f_M\left(\frac{M(M-1)}{2} - c\right) &= \frac{1}{M} \left[1 - F_{M-1}\left(\frac{M(M-1)}{2} - c - M\right) \right] = \\ &= \frac{1}{M} \left[f_{M-1}\left(\frac{M(M-1)}{2} - c - M + 1\right) + \dots + f_{M-1}\left(\frac{(M-1)(M-2)}{2}\right) \right] = \\ &= \frac{1}{M} \left[f_{M-1}\left(\frac{(M-1)(M-2)}{2} - c\right) + \dots + f_{M-1}\left(\frac{(M-1)(M-2)}{2}\right) \right] = \\ &= \frac{1}{M} \sum_{i=0}^c f_{M-1}\left(\frac{(M-1)(M-2)}{2} - c\right), \end{aligned} \quad (5.37)$$

usando a hipótese de simetria de f_{M-1} em (5.37) pode escrever-se:

$$f_M\left(\frac{M(M-1)}{2} - c\right) = \frac{1}{M} \sum_{i=0}^c f_{M-1}(i). \quad (5.38)$$

De (5.35), (5.36) e (5.38) pode concluir-se que:

$$f_M(c) = f_M\left(\frac{M(M-1)}{2} - c\right), \text{ para } 0 \leq c \leq M-1$$

o que prova a simetria nesse ramo da função.

$$\text{2º ramo: } M \leq c \leq \text{int}\left(\frac{(M-1)(M-2)}{4}\right)$$

$$\begin{aligned} f_M(c) &= \frac{1}{M} [F_{M-1}(c) - F_{M-1}(c-M)] = \\ &= \frac{1}{M} \sum_{i=0}^{M-1} f_{M-1}(c-i). \end{aligned} \quad (5.39)$$

Um pequeno cálculo permite verificar que, para $M > 4$:

$$\begin{aligned} M \leq c \leq \text{int}\left(\frac{(M-1)(M-2)}{4}\right) &\Leftrightarrow \\ \Leftrightarrow \frac{M(M-1)}{2} - \text{int}\left(\frac{(M-1)(M-2)}{4}\right) &\leq \frac{M(M-1)}{2} - c \leq \frac{(M-1)(M-2)}{2} - 1, \end{aligned} \quad (5.40)$$

o que permite escrever

$$\begin{aligned} f_M\left(\frac{M(M-1)}{2} - c\right) &= \frac{1}{M} \left[F_{M-1}\left(\frac{M(M-1)}{2} - c\right) - F_{M-1}\left(\frac{M(M-1)}{2} - c - M\right) \right] = \\ &= \frac{1}{M} \sum_{i=0}^{M-1} f_{M-1}\left(\frac{M(M-1)}{2} - c - i\right). \end{aligned} \quad (5.41)$$

Usando a hipótese de simetria de f_{M-1} em (5.41) tem-se:

$$\begin{aligned} & \frac{1}{M} \sum_{i=0}^{M-1} f_{M-1} \left(\frac{(M-1)(M-2)}{2} - \frac{M(M-1)}{2} + c + i \right) = \\ & = \frac{1}{M} \sum_{i=0}^{M-1} f_{M-1} (c - (M-1-i)). \end{aligned} \quad (5.42)$$

Fazendo a mudança de variável $i' = M-1-i$ e atendendo a que $0 \leq i \leq M-1 \Leftrightarrow 0 \leq i' \leq M-1$, pode escrever-se:

$$f_M \left(\frac{M(M-1)}{2} - c \right) = \frac{1}{M} \sum_{i=0}^{M-1} f_{M-1} (c - i'). \quad (5.43)$$

De (5.39), (5.40) e (5.43) pode concluir-se que:

$$f_M(c) = f_M \left(\frac{M(M-1)}{2} - c \right), \text{ para } M \leq c \leq \text{int} \left(\frac{(M-1)(M-2)}{4} \right),$$

o que prova a simetria de f_M para valores de c que verifiquem a dupla desigualdade correspondente, concluindo a demonstração pretendida.

As expressões para os parâmetros média e variância da distribuição exacta, coincidem com as expressões obtidas por Lerman [7] e Soromenho [12], já referidas no ponto 5.1.1 e expressas em (5.18) e (5.19).

A obtenção da função de probabilidade exacta de C_M permite analisar a qualidade das distribuições assintóticas propostas por Kendall e por Lerman. Em 5.1.3 mostrou-se ainda que as estatísticas T_K^* e T_{Ler}^* , obtidas por normalização das anteriores, coincidiam.

Para valores de M compreendidos entre 4 e 30, calculou-se a distribuição exacta de C_M , f_M , usando (5.32), e a respectiva distribuição aproximada de acordo com os resultados de Kendall e Lerman, $\tilde{f}_M$, isto é:

$$\tilde{f}_M(c) = \Phi\left(\frac{c + 0.5 - E(C)}{\sqrt{V(C)}}\right) - \Phi\left(\frac{c - 0.5 - E(C)}{\sqrt{V(C)}}\right), \text{ para } 0 \leq c \leq \frac{M(M-1)}{2}, \quad (5.44)$$

onde Φ denota a função de distribuição da $N(0,1)$ e $E(C)$ e $V(C)$ são dados por (5.18) e (5.19).

Da análise dos resultados obtidos podemos constatar que a diferença entre as duas funções é reduzida e decresce com M . Para cada valor de M registou-se o máximo dos valores absolutos das diferenças entre as duas funções, isto é:

$$\max_c \left| f_M(c) - \tilde{f}_M(c) \right|. \quad (5.45)$$

O valor máximo ocorre sempre na zona central da distribuição e é sempre inferior a 2×10^{-2} , obtido para $M=4$. Para ilustrar o decréscimo desses valores máximos, para os M considerados, pode referir-se que para $M=6$ esse valor é já inferior a 7×10^{-3} , para $M=10$ é inferior a 2×10^{-3} e para $M=25$ é da ordem de 2×10^{-4} .

No anexo 1 são dadas algumas tabelas com os valores obtidos nas experiências realizadas, para M compreendido entre 4 e 10.

As duas figuras seguintes são uma representação gráfica das curvas ajustadas aos valores das duas funções, f_M e $\tilde{f}_M$, para $M=4$ e $M=10$.

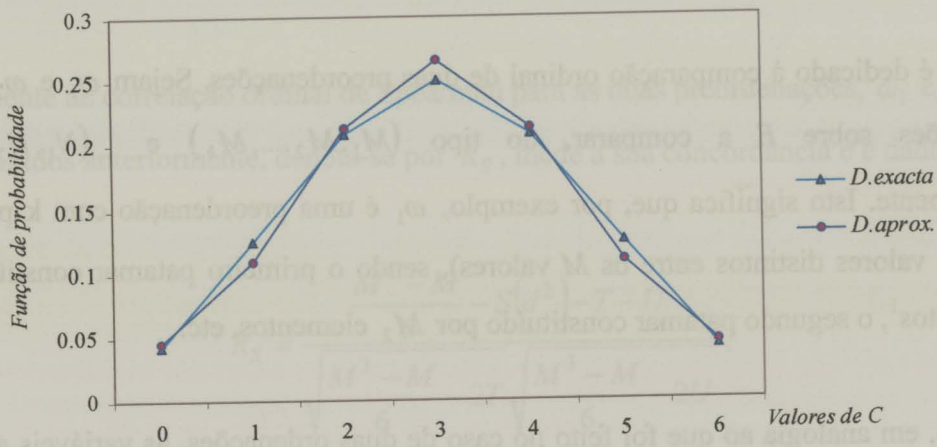


Fig. 5.1 – Distribuições exacta e aproximadas de C_M , para $M=4$

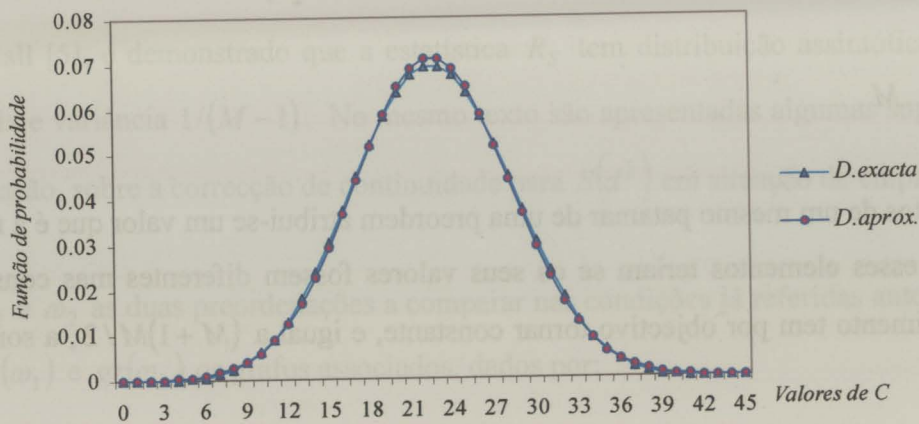


Fig. 5.2 – Distribuições exacta e aproximadas de C_M , para $M=10$

Este estudo permite concluir, que a utilização da distribuição assintótica de C_M , pode ser feita para valores de M bastante reduzidos.

5.2 COMPARAÇÃO DE PREORDENAÇÕES

Este ponto é dedicado à comparação ordinal de duas preordenações. Sejam ω_1 e ω_2 as duas preordenações sobre E a comparar, do tipo $(M_1, M_2, \dots, M_k)$ e $(N_1, N_2, \dots, N_h)$, respectivamente. Isto significa que, por exemplo, ω_1 é uma preordenação com k patamares (ocorrem k valores distintos entre os M valores), sendo o primeiro patamar constituído por M_1 elementos¹, o segundo patamar constituído por M_2 elementos, etc.

Definam-se, em analogia ao que foi feito no caso de duas ordenações, as variáveis aleatórias ordinais U e V dadas por:

$$U_k = \text{"Número de ordem do } k\text{-ésimo elemento de } \omega_1 \text{"}$$

e

$$V_k = \text{"Número de ordem do } k\text{-ésimo elemento de } \omega_2 \text{"},$$

com $k = 1, \dots, M$.

Aos elementos de um mesmo patamar de uma preordem atribui-se um valor que é a média das ordens que esses elementos teriam se os seus valores fossem diferentes mas consecutivos. Este procedimento tem por objectivo tornar constante, e igual a $(M+1)M/2$, a soma das M ordens.

De seguida apresentam-se alguns dos coeficientes possíveis para quantificar a comparação entre ω_1 e ω_2 . Por simplificação e porque não dará origem a dúvidas, optou-se por usar a mesma notação, para os coeficientes no caso de preordenações, que já havia sido usada na comparação de duas ordenações.

¹ M_1 elementos iguais, que será o maior valor, se a ordem for decrescente, dos M associados à função de comparação.

5.2.1 Coeficientes de Comparação

O coeficiente de correlação ordinal de Spearman para as duas preordenações, ω_1 e ω_2 de tipos definidos anteriormente, denota-se por R_S , mede a sua concordância e é dado por:

$$R_S = \frac{\frac{M^3 - M}{6} - S(d^2) - T - U}{\sqrt{\frac{M^3 - M}{6} - 2T} \sqrt{\frac{M^3 - M}{6} - 2U}}, \quad (5.46)$$

em que:

$$T = \frac{\sum_{1 \leq i \leq k} (M_i^3 - M_i)}{12}, \quad U = \frac{\sum_{1 \leq j \leq h} (N_j^3 - N_j)}{12} \quad \text{e} \quad S(d^2) = \sum_{1 \leq l \leq M} (U_l - V_l)^2. \quad (5.47)$$

Em Kendall [5], é demonstrado que a estatística R_S tem distribuição assintótica normal de média nula e variância $1/(M-1)$. No mesmo texto são apresentadas algumas sugestões, sem demonstração, sobre a correcção de continuidade para $S(d^2)$ em situação de empates.

Sejam ω_1 e ω_2 as duas preordenações a comparar nas condições já referidas anteriormente, e sejam $gr(\omega_1)$ e $gr(\omega_2)$ os grafos associados, dados por:

$$gr(\omega_i) = \{((x, y), (z, t)) \in F \times F - \Delta : (x, y) < (z, t) \text{ e não } (z, t) < (x, y)\}. \quad (5.48)$$

Esta definição de grafo tem como consequência que os elementos $((x, y), (x, y))$ não pertencem a $gr(\omega_i)$ e que se $((x, y), (z, t))$ pertence a $gr(\omega_i)$ então $((z, t), (x, y))$ não pode pertencer.

Facilmente se conclui que:

$$card(gr(\omega_1)) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k M_i M_j = \frac{M(M-1)}{2} - \sum_{i=1}^k \frac{M_i(M_i-1)}{2} \quad (5.49)$$

e analogamente,

$$\text{card}(gr(\omega_2)) = \sum_{i=1}^{h-1} \sum_{j=i+1}^h N_i N_j = \frac{M(M-1)}{2} - \sum_{i=1}^h \frac{N_i(N_i-1)}{2} \quad (5.50)$$

O coeficiente Tau de Kendall para preordenações, que será representado por T_K , é dado por:

$$T_K = \frac{C-D}{\sqrt{\text{card}(gr(\omega_1)) \times \text{card}(gr(\omega_2))}} = \frac{C-D}{\sqrt{\sum_{i<j} M_i M_j} \sqrt{\sum_{i<j} N_i N_j}}, \quad (5.51)$$

onde C e D representam, respectivamente, o número de concordâncias e o número de discordâncias entre ω_1 e ω_2 , no sentido em que foram definidos no ponto 5.1.

Kendall [5] demonstra que a variável aleatória $(C-D)$ tem distribuição assintótica normal com:

$$E(C-D) = 0 \quad (5.52)$$

e

$$\begin{aligned} V(C-D) = & \frac{1}{18} \left\{ M(M-1)(2M+5) - \sum_{1 \leq i \leq k} M_i(M_i-1)(2M_i+5) - \sum_{1 \leq l \leq h} N_l(N_l-1)(2N_l+5) \right\} + \\ & + \frac{1}{9M(M-1)(M-2)} \left\{ \sum_{1 \leq i \leq k} M_i(M_i-1)(M_i-2) \right\} \left\{ \sum_{1 \leq l \leq h} N_l(N_l-1)(N_l-2) \right\} + \\ & + \frac{1}{2M(M-1)} \left\{ \sum_{1 \leq i \leq k} M_i(M_i-1) \right\} \left\{ \sum_{1 \leq l \leq h} N_l(N_l-1) \right\}. \end{aligned} \quad (5.53)$$

Lerman considerou para medida de concordância entre ω_1 e ω_2 o coeficiente

$$T_{Ler} = \text{card}(gr(\omega_1) \cap gr(\omega_2))$$

$$= \{((x, y), (z, t)) \in F \times F - \Delta : (x, y) < (z, t) \text{ e não } (z, t) < (x, y) \text{ para } \omega_1 \text{ e } \omega_2\}, \quad (5.54)$$

que representa o número de pares ordenados da mesma forma, pelas duas preordenações a comparar, ω_1 e ω_2 , e onde $gr(\omega_1)$ e $gr(\omega_2)$ denotam os respectivos grafos, definidos por 5.48.

Em Lerman [8], o autor demonstra que este coeficiente tem distribuição assintótica normal e que:

$$E(T_{Ler}) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^k M_i M_j}{\sqrt{M(M-1)}} \frac{\sum_{i=1}^{h-1} \sum_{j=i+1}^h N_i N_j}{\sqrt{M(M-1)}} \quad (5.55)$$

e

$$\begin{aligned} V(T_{Ler}) = E(T_{Ler}) + & \frac{\sum_{i=1}^{k-1} M_i \sum_{j=i+1}^k M_j \left(\sum_{j=i+1}^k M_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} \frac{\sum_{i=1}^{h-1} N_i \sum_{j=i+1}^h N_j \left(\sum_{j=i+1}^h N_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} + \\ & + \frac{\sum_{i=2}^k M_i \sum_{j=1}^{i-1} M_j \left(\sum_{j=1}^{i-1} M_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} \frac{\sum_{i=2}^h N_i \sum_{j=1}^{i-1} N_j \left(\sum_{j=1}^{i-1} N_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} + \\ & + 2 \frac{\sum_{i=2}^{k-1} M_i \left(\sum_{j=i+1}^k M_j \right) \left(\sum_{l=1}^{i-1} M_l \right)}{\sqrt{M(M-1)(M-2)}} \frac{\sum_{i=2}^{h-1} N_i \left(\sum_{j=i+1}^h N_j \right) \left(\sum_{l=1}^{i-1} N_l \right)}{\sqrt{M(M-1)(M-2)}} + \Lambda \end{aligned} \quad (5.56)$$

onde $\Lambda = \theta \xi - E(T_{Ler})^2$,

$$\theta = \frac{1}{\sqrt{M(M-1)(M-2)(M-3)}} \left\{ \sum_{i=1}^{k-1} \sum_{j=i+1}^k M_i M_j \left[\left(\sum_{p=1}^{k-1} \sum_{l=i+1}^k M_p M_l \right) + M_i + M_j - 2M + 1 \right] \right\}$$

e ξ é a expressão análoga à de θ , correspondente à preordenação ω_2 .

Lerman [7], (pág. 134 e seguintes) afirma que o valor de Λ é desprezável e aconselha o cálculo aproximado de $V(T_{Ler})$ não incluindo esse termo. No quadro seguinte apresentam-se, para alguns pares de preordenações, os valores de $V(T_{Ler})$, de Λ e ainda da variância sem o termo Λ , isto é o valor aproximado de $V(T_{Ler})$ sugerido por Lerman. Este quadro mostra bem que o termo Λ não só não deve ser desprezado, como tem uma ordem de grandeza que pode ser superior à da própria variância.

M	$(M_1, M_2, \dots, M_k)$	$(N_1, N_2, \dots, N_h)$	$V(T_{Ler})$	Λ	$V(T_{Ler}) - \Lambda$
10	(2, 8)	(6, 4)	7.9289	-10.2044	18.1333
10	(2, 3, 5)	(3, 3, 4)	24.5251	-48.8582	73.3833
10	(1, 2, 3, 4)	(3, 3, 4)	25.9318	-61.5516	87.4834
10	(1, 2, 3, 4)	(1, 2, 3, 4)	27.5464	-70.6592	98.2056
10	(1, 2, 3, 4)	(2, 2, 2, 2, 2)	27.1517	-91.1817	118.3334
10	(2, 2, 2, 2, 2)	(2, 2, 2, 2, 2)	29.0300	-117.6367	146.6667
21	(1, 2, 3, 4, 5, 6)	(5, 4, 4, 4, 4)	247.4622	-997.5202	1244.9824
21	(1, 2, 3, 4, 5, 6)	(7, 7, 7)	230.9774	691.2112	922.1886

Quadro 5.6 – Valores de Λ e de $V(T_{Ler})$.

Goodman e Kruskal [1] propuseram uma medida de associação para tabelas de contingência ordenadas¹, que representaram por γ , de Goodman-Kruskal, e que neste trabalho será representado por T_{GK} , dado que a letra γ está reservada para representar coeficiente de comparação entre elementos.

Em classificação, surgem com frequência preordenações com empates bastante extensos, correspondendo a patamares de grande comprimento. Este aspecto quando muito acentuado conduz à situação extrema de tabelas de contingência, e consequentemente à utilização de T_{GK} .

Como nos coeficientes anteriores ω_1 e ω_2 são duas preordenações sobre E do tipo $(M_1, M_2, \dots, M_k)$ e $(N_1, N_2, \dots, N_h)$ respectivamente, em que alguns valores M_i ou N_j podem ser elevados.

Usando a mesma notação que foi usada no início do ponto 5.1., a variável aleatória A , definida em (5.8) pode assumir os valores -1 , 1 e 0 (zero), porque podem ocorrer empates.

O coeficiente de associação de Goodman-Kruskal toma a seguinte forma:

$$T_{GK} = \frac{C - D}{C + D}, \quad (5.57)$$

onde C e D representam, respectivamente, o número de concordâncias e o número de discordâncias entre ω_1 e ω_2 .

T_{GK} é um coeficiente que assume valores entre -1 e 1 , atingindo o valor 1 quando o número de discordâncias é nulo (e $C \neq 0$) e sendo -1 na situação oposta, isto é, quando não existem concordâncias (e $D \neq 0$).

Alguns resultados teóricos sobre este coeficiente podem ser encontrados em Goodman & Kruskal [2], entre os quais a expressão da variância de T_{GK} , um método para a calcular e um limite superior simples que a permite estimar.

Não foi definido o coeficiente correspondente a este no caso de comparação de duas ordenações, porque não existindo empates tem-se que $C + D = \frac{1}{2}M(M - 1)$, e consequentemente o coeficiente T_{GK} para ordenações coincide com o coeficiente T_K .

¹ As modalidades ou valores das duas variáveis podem ser ordenadas.

5.3 COMPARAÇÃO DE UMA ORDENAÇÃO E UMA PREORDENAÇÃO

Este ponto é dedicado à comparação ordinal de uma ordenação e uma preordenação. Sejam O e ω as ordenação e preordenação de comprimentos M , sobre E , a comparar, sendo $(M_1, M_2, \dots, M_k)$ o tipo de ω . Em analogia ao que foi feito na comparação de duas ordenações e de duas preordenações, sejam U e V as variáveis aleatórias ordinais associadas, respectivamente, a O e ω .

No que se segue, para os vários coeficientes apresentados nas secções 5.1 e 5.2, será feita uma comparação entre o coeficiente proposto para duas ordenações e o correspondente para duas preordenações. Tal procedimento tem como objectivo averiguar se, em cada caso, o coeficiente para comparar duas preordenações é uma generalização natural do coeficiente para ordenações. Se tal acontecer, a consideração de preordenações com M patamares de comprimento igual a um, deverá conduzir a expressões iguais às encontradas para ordenações.

A possibilidade de apresentar um único coeficiente geral para as três situações, torna simples a abordagem da comparação de uma ordenação com uma preordenação.

5.3.1 Análise dos Vários Coeficientes

Considere-se o coeficiente de correlação ordinal de Spearman para duas preordenações, R_S , apresentado em (5.46). No caso particular de ω_1 e ω_2 serem ordenações de comprimento M , cada uma será constituída por M patamares todos de um elemento, isto é:

$$M_1 = M_2 = \dots = M_M = 1 \quad \text{e} \quad N_1 = N_2 = \dots = N_M = 1. \quad (5.58)$$

Neste contexto T e U em (5.47) assumem o valor zero e facilmente se verifica que (5.46) coincide com o coeficiente R_S para comparar duas ordenações (por exemplo a expressão 5.11).

Pode afirmar-se que o coeficiente R_S para preordenações é uma generalização do coeficiente, apresentado em 5.1.1, para comparar duas ordenações.

Parece então natural que para comparar uma ordenação com uma preordenação se considere o coeficiente R_S geral, (5.46), e se considerem os comprimentos dos patamares adequados.

Também para o coeficiente de Kendall, se verifica que o coeficiente T_K para preordenações, definido em (5.51), é uma generalização do coeficiente correspondente para comparar duas ordenações, apresentado em (5.13). Quanto aos parâmetros valor médio e variância, o primeiro coincide nos dois casos, e fazendo as substituições adequadas em (5.53) obtém-se:

$$V(C - D) = \frac{1}{18} M(M-1)(2M+5), \quad (5.59)$$

consequentemente por (5.51) e porque $\sum_{i < j} M_i M_j = \sum_{i < j} N_i N_j = \frac{M(M-1)}{2}$ vem:

$$\begin{aligned} V(T_K) &= \left[\frac{M(M-1)}{2} \right]^2 \frac{1}{18} M(M-1)(2M+5) \\ &= \frac{2(2M+5)}{9M(M-1)}, \end{aligned} \quad (5.60)$$

coincidindo com a expressão da variância de T_K obtida para duas ordenações, (5.15).

A comparação de uma ordenação com uma preordenação tem assim uma solução simples, que é considerar o coeficiente geral.

Em analogia com o que foi feito para os dois coeficientes anteriores, interessa agora verificar que o coeficiente de Lerman para ordenações é uma particularização do mesmo coeficiente para preordenações. Sendo ω_1 e ω_2 duas ordenações de comprimento M , cada uma será constituída por M patamares, todos de um elemento, isto é $k = h = M$ e $M_i = N_j = 1$, para $i, j = 1, \dots, M$.

Atendendo a que

$$\frac{\sum_{i=1}^{M-1} \sum_{j=i+1}^M M_i M_j}{\sqrt{M(M-1)}} = \frac{\sum_{i=1}^{M-1} \sum_{j=i+1}^M N_i N_j}{\sqrt{M(M-1)}} = \frac{\sqrt{M(M-1)}}{2}, \quad (5.61)$$

facilmente se conclui que (5.55) vem igual à expressão encontrada anteriormente para média do coeficiente de Lerman para comparar ordenações, (5.18).

Substituições análogas na expressão da variância, (5.56) incluindo o termo Λ , e bastante trabalho algébrico, mostram que também a expressão da variância assim obtida, coincide com a expressão da variância para ordenações, (5.19). As linhas gerais da demonstração são:

$$\begin{aligned} & \frac{\sum_{i=1}^{M-1} M_i \sum_{j=i+1}^M M_j \left(\sum_{j=i+1}^M M_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} = \frac{\sum_{i=1}^{M-1} N_i \sum_{j=i+1}^M N_j \left(\sum_{j=i+1}^M N_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} = \\ & = \frac{\sum_{i=2}^M M_i \sum_{j=1}^{M-1} M_j \left(\sum_{j=1}^{M-1} M_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} = \frac{\sum_{i=2}^M N_i \sum_{j=1}^{M-1} N_j \left(\sum_{j=1}^{M-1} N_j - 1 \right)}{\sqrt{M(M-1)(M-2)}} = \\ & = \frac{1}{\sqrt{M(M-1)(M-2)}} \sum_{i=2}^M \left[\left(\sum_{j=1}^{i-1} M_j \right)^2 - \sum_{j=1}^{i-1} M_j \right] = \\ & = \frac{1}{\sqrt{M(M-1)(M-2)}} \left[\frac{M(M-1)(2M-1)}{6} - \frac{M(M-1)}{2} \right] = \\ & = \frac{M(M-1)}{\sqrt{M(M-1)(M-2)}} \frac{M-2}{3}. \end{aligned} \quad (5.62)$$

Por outro lado

$$\begin{aligned}
 \frac{\sum_{i=2}^{M-1} M_i \left(\sum_{j=i+1}^M M_j \right) \left(\sum_{l=1}^{i-1} M_l \right)}{\sqrt{M(M-1)(M-2)}} &= \frac{\sum_{i=2}^{M-1} N_i \left(\sum_{j=i+1}^M N_j \right) \left(\sum_{l=1}^{i-1} N_l \right)}{\sqrt{M(M-1)(M-2)}} = \\
 &= \frac{\sum_{i=2}^{M-2} i(M-i-1)}{\sqrt{M(M-1)(M-2)}} = \\
 &= \frac{M(M-1)(M-2)}{6\sqrt{M(M-1)(M-2)}}. \tag{5.63}
 \end{aligned}$$

Quanto à parcela $\Lambda = \theta\xi - E(T_{Ler})^2$ tem-se:

$$\begin{aligned}
 \theta = \xi &= \frac{1}{\sqrt{M(M-1)(M-2)(M-3)}} \left\{ \sum_{i=1}^{M-1} \sum_{j=i+1}^M M_i M_j \left[\left(\sum_{p=1}^{M-1} \sum_{l=i+1}^M M_p M_l \right) + M_i + M_j - 2M + 1 \right] \right\} = \\
 &= \frac{1}{\sqrt{M(M-1)(M-2)(M-3)}} \frac{M(M-1)}{2} \frac{M(M-1)+6-4M}{4}, \tag{5.64}
 \end{aligned}$$

donde

$$\Lambda = \frac{M(M-1)}{(M-2)(M-3)} \frac{[M(M-1)+6-4M]^2}{16} - \frac{M^2(M-1)^2}{16}. \tag{5.65}$$

Usando (5.62), (5.63) e (5.65) para substituir convenientemente em (5.56) vem:

$$V(T_{Ler}) = \frac{M(M-1)}{4} \frac{2M+5}{18}, \tag{5.66}$$

que coincide com (5.19).

Na comparação de uma ordenação com uma preordenação será usado o coeficiente T_{GK} definido em (5.57).

5.3.2 A Distribuição Exacta da Estatística C

Tal como foi já referido, a estatística número de concordâncias desempenha um papel importante nesta matéria. Em 5.1.4 foi deduzida a lei exacta de C para duas ordenações. Nas comparações que envolvam preordenações, fixar o valor de M não é suficiente para permitir estudar C , já que além das discordâncias existem também os empates, que dependem dos tipos das preordenações, isto é do comprimento dos patamares. Assim a dedução da distribuição de C para duas preordenações em geral é bastante complexa.

No que se segue vai ser exposto um procedimento para deduzir a distribuição de C , no caso de comparação de uma ordenação com uma preordenação, quando esta assume um dos dois tipos particulares seguintes: associada a uma cadeia pura ou a uma árvore do tipo equilibrada.

O tipo de uma preordenação é completamente caracterizado pelo seu comprimento e pelo número e extensão dos seus patamares. Seja ω uma preordenação de comprimento M , com k patamares, de tipo $(M_1, M_2, \dots, M_k)$. A ordenação O de comprimento M que será considerada é a ordenação “natural”, isto é, a ordenação cujo elemento de ordem j é j , para $j = 1, 2, \dots, M$.

Considere-se a estatística $C_M =$ “número de concordâncias entre a preordenação ω , com um tipo fixo, e a ordenação O ”.

5.3.2.1 Preordenação Associada à Cadeia Pura

Neste caso considera-se que a preordenação ω é do tipo $(N_1, N_2, \dots, N_h)$ em que $N_i = i$ para $i = 1, \dots, h$

Seja P_i a variável aleatória que representa o “número de concordâncias dos elementos do patamar i , com todos os elementos dos patamares inferiores, ou à esquerda de i ”, para $i = 1, \dots, h$. Naturalmente que:

$$C_M = \sum_{i=2}^h P_i. \quad (5.67)$$

O objectivo é deduzir a função de probabilidade exacta de C_M e compará-la com as funções de probabilidade aproximadas de acordo com os resultados de Kendall e de Lerman.

No que se segue usar-se-á a seguinte notação:

f_k - função de probabilidade de C_k , sendo k o comprimento da preordenação.

g_i - função de probabilidade de P_i , para $i = 1, \dots, h$.

$D(X)$ - domínio da variável aleatória X .

$\max(x)$ - maior valor do domínio de X .

Considere-se uma preordenação ω' do mesmo tipo que ω , mas só com os $h-1$ primeiros patamares, e portanto com comprimento $M' < M$. Pode escrever-se que:

$$C_M = C_{M'} + P_h \quad (5.68)$$

A função de probabilidade de C_M , f_M , pode ser obtida à custa das funções de probabilidade de $C_{M'}$, $f_{M'}$, e de P_h , g_h .

Alguma combinatoria permite concluir que:

$$\square \quad D(C_M) = \{0, 1, \dots, \max(c_M)\}, \text{ com } \max(c_M) = \max(c_{M'}) + \max(p_h). \quad (5.69)$$

$$\square \quad \text{Se } \max(c_{M'}) \leq \max(p_h):$$

$$f_M(c_M) = \begin{cases} \sum_{i=0}^{c_M} f_{M'}(i)g_h(c_M - i) & \text{se } 0 \leq c_M \leq \max(c_{M'}) \\ \sum_{i=0}^{\max(c_{M'})} f_{M'}(i)g_h(c_M - i) & \text{se } \max(c_{M'}) + 1 \leq c_M \leq \max(p_h) \\ \sum_{i=c_M - \max(p_h)}^{\max(c_{M'})} f_{M'}(i)g_h(c_M - i) & \text{se } \max(p_h) + 1 \leq c_M \leq \max(c_{M'}) + \max(p_h) \end{cases} \quad (5.70)$$

□ Se $\max(c_{M'}) > \max(p_h)$:

$$f_M(c_M) = \begin{cases} \sum_{i=0}^{c_M} f_{M'}(i)g_h(c_M - i) & \text{se } 0 \leq c_M \leq \max(p_h) \\ \sum_{i=c_M - \max(p_h)}^{c_M} f_{M'}(i)g_h(c_M - i) & \text{se } \max(p_h) + 1 \leq c_M \leq \max(c_{M'}) \\ \sum_{i=c_M - \max(p_h)}^{\max(c_{M'})} f_{M'}(i)g_h(c_M - i) & \text{se } \max(c_{M'}) + 1 \leq c_M \leq \max(p_h) + \max(c_{M'}) \end{cases} \quad (5.71)$$

Quanto à determinação da função g_h , note-se que:

$$P_h = \sum_{i=1}^h E_i, \quad (5.72)$$

em que E_i , é o número de concordâncias do elemento i , do patamar h , com os elementos que estão à sua esquerda na preordenação.

Note-se que, dentro de um patamar da preordenação, os índices dos elementos podem ser ordenados de forma crescente. Este facto permite escrever que:

$$\square D(E_1) = \{0, 1, \dots, M'\}$$

$$\square D(E_i) = \{0, 1, \dots, e_{i-1}\}, \text{ para } i = 2, \dots, h.$$

- As variáveis E_i têm distribuição uniforme nos respectivos domínios.

Atendendo a (5.72), pode afirmar-se que P_h é a soma de h variáveis aleatórias uniformes com diferentes domínios. Para pequenos valores de h , um algoritmo relativamente simples permite calcular g_h .

Quando $h = 2$ tem-se $M = 3$ e facilmente se verifica que:

c_3	0	1	2
$f_3(c_3)$	1/3	1/3	1/3

Quadro 5.7 – Função de probabilidade de C_3 (caso cadeia pura).

O conhecimento de f_3 , define completamente as expressões de recorrência (5.70) e (5.71).

Foram elaborados programas para o cálculo da função de probabilidade do número de concordâncias entre uma ordenação e uma preordenação associada a uma cadeia pura, de comprimento M , C_M . Quando M cresce o algoritmo complica-se bastante. Foram considerados valores de M não superiores a 50.

O coeficiente de Kendall para preordenações, (5.51), depende do número de concordâncias, D . Foi ainda referido que a variável aleatória $C-D$ tem distribuição assintótica normal, com parâmetros dados por (5.52) e (5.53). No caso particular de uma ordenação e uma preordenação, estas expressões simplificam-se, passando a ser:

$$E(C - D) = 0 \quad (5.73)$$

$$V(C - D) = \frac{1}{18} \left[M(M-1)(2M+5) - \sum_{1 \leq i \leq h} N_i(N_i-1)(2N_i+5) \right]. \quad (5.74)$$

Fixado M e o tipo de preordenação pode obter-se a distribuição assintótica para C . Note-se que:

$$C + D + E = M(M-1)/2, \quad (5.75)$$

sendo E o número de empates associados à preordenação,

$$E = \sum_{i=2}^h \binom{N_i}{2}. \quad (5.76)$$

De (5.73), (5.75) e (5.76) conclui-se que:

$$E(C) = \frac{1}{2} \left(\frac{M(M-1)}{2} - E \right) \quad (5.77)$$

e de (5.74) e (5.77) obtém-se:

$$V(C) = \frac{1}{72} \left[M(M-1)(2M+5) - \sum_{1 \leq i \leq h} N_i(N_i-1)(2N_i+5) \right]. \quad (5.78)$$

Fixado M e para este tipo particular de preordenação, o resultado de Kendall permite dizer que C segue uma distribuição assintótica normal com parâmetros dados por (5.77) e (5.78).

Particularizando o coeficiente de Lerman para comparar preordenações, definido em (5.54), e com distribuição assintótica normal com parâmetros dados por (5.55) e (5.56), para a situação aqui tratada, tem-se que a estatística C_M coincide com esse coeficiente.

Substituições adequadas em (5.55) e (5.56), permitem obter as expressões da média e da variância de C , no caso de uma preordenação e uma ordenação:

$$E(C) = \frac{1}{2} \sum_{i=1}^{h-1} \sum_{j=i+1}^h N_i N_j, \quad (5.79)$$

a expressão da variância não será aqui apresentada para não sobrecarregar a exposição.

No quadro 5.8, que figura na página seguinte, apresentam-se os valores da média e da variância, para a distribuição exacta e aproximadas, segundo Kendall e Lerman, da estatística C_M , para alguns valores de M .

Da análise do referido quadro verifica-se que os valores médios e as variâncias coincidem, para as três distribuições, para todos os valores de M , o que permite concluir que os resultados de distribuições assintóticas para C_M de Lerman e Kendall são equivalentes.

Em analogia ao que foi feito para a comparação de duas ordenações, calcularam-se as distribuições exacta e aproximada para alguns valores de M . Na Fig. 5.3 estão representadas essas duas distribuições, para $M = 6$.

5.3.2.2 Preordenação Associada a uma Árvore Equilibrada

Realizou-se um estudo análogo ao anterior, mas para preordenações associadas a árvores equilibradas. A notação usada mantém-se. A determinação da função de probabilidade exacta de C_M é um pouco mais complicada, já que o aumento de M provoca alterações que não são só o acréscimo de patamares.

Foram elaborados algoritmos para a determinação das funções f_k e g_i que não serão apresentadas para não sobrecarregar a exposição. O quadro 5.9 resume os valores obtidos, para as distribuições exacta e aproximadas, para este tipo de preordenação e a Fig. 5.4 representa graficamente essas duas distribuições, para $M = 6$.

Como conclusão deste estudo da comparação de uma ordenação com uma preordenação, apenas dos dois tipos considerados, verificou-se que em ambos os casos as médias e as variâncias de Kendall e de Lerman coincidem com as exactas, e que as distribuições assintóticas, propostas por esses dois autores são muito boas aproximações, mesmo para valores de M muito reduzidos.

M	MÉDIA			VARIÂNCIA		
	Exacta	Kendall	Lerman	Exacta	Kendall	Lerman
6	5.5	5.5	5.5	5.917	5.917	5.917
10	17.5	17.5	17.5	27.917	27.917	27.917
15	42.5	42.5	42.5	94.583	94.583	94.583
21	87.5	87.5	87.5	259.583	259.583	259.583
28	161.0	161.0	161.0	614.833	614.833	614.833
36	273.0	273.0	273.0	1305.500	1305.500	1305.500
45	435.0	435.0	435.0	2547.500	2547.500	2547.500

Quadro 5.8 – Valores médio e de variância de C_M (caso cadeia pura)

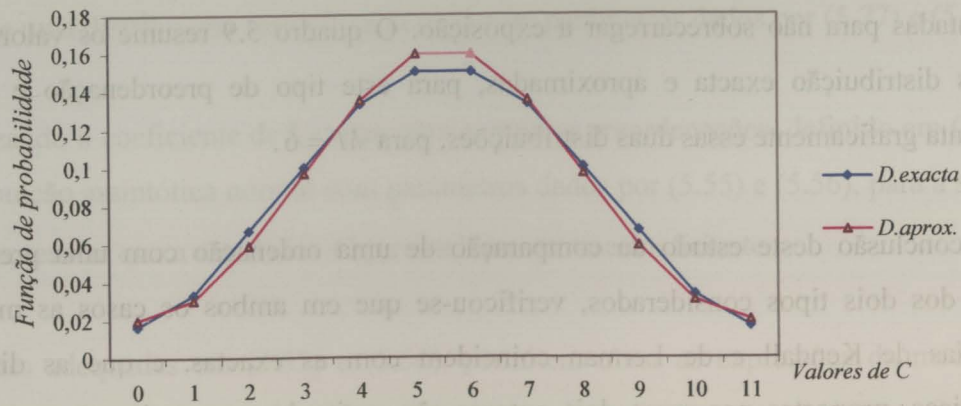


Fig. 5.3 – Distribuições exacta e aproximada de C_M (caso cadeia pura), para $M = 6$

REFERÊNCIAS

[1] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification, *Journal of the American Statistical Association*, 49, 732-764.

[2] Goodman, L. A., Kruskal, W. H. (1953): Measures of Association for Cross Classification.

[3] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[4] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[5] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[6] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[7] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[8] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[9] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[10] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

[11] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification.

M	MÉDIA			VARIÂNCIA		
	Exacta	Kendall	Lerman	Exacta	Kendall	Lerman
6	4.5	4.5	4.5	4.917	4.917	4.917
10	14.5	14.5	14.5	23.917	23.917	23.917
15	35.5	35.5	35.5	83.583	83.583	83.583
21	68.5	68.5	68.5	218.584	218.584	218.584
28	123.0	123.0	123.0	512.833	512.833	512.833

Quadro 5.9 – Valores médios e de variância de C_M (caso equilibrado)

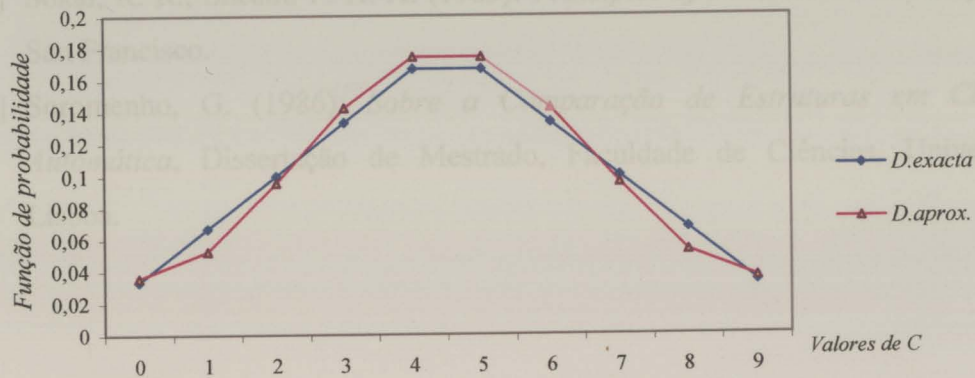


Fig. 5.4 – Distribuições exata e aproximada de C_M (caso equilibrado), para $M = 6$

REFERÊNCIAS

- [1] Goodman, L. A., Kruskal, W. H. (1954): Measures of Association for Cross Classification, *Journal of the American Statistical Association*, 49, 732-764.
- [2] Goodman, L. A., Kruskal, W. H. (1963): Measures of Association for Cross Classification, III, *Journal of the American Statistical Association*, 58, 310-364.
- [3] Haigh, J. (1971): A neat way to prove asymptotic normality, *Biometrika*, 3, 677.
- [4] Kendall, M. G., Stuart, A. (1969): *The Advanced Theory of Statistics*, Vol. 1, 3th ed., Charles Griffin & Company Limited, London.
- [5] Kendall, M. G. (1970): *Rank Correlation Methods*, 4th ed., Griffin, London.
- [6] Kendall, M. G. (1990): *Rank Correlation Methods*, 5th ed., Griffin, London.
- [7] Lerman, I. C. (1981): *Classification et Analyse Ordinale des Données*, Dunod, Paris.
- [8] Lerman, I. C. (1992): Conception et Analyse de la Forme Limite d'Une Famille de Coefficients Statistiques d'Association entre Variables Relationnelles, II, *Mathématiques Informatiques et Sciences Humaines*, 119, 75-100.
- [9] Nicolau, F. C. (1984): Problemas de Validade em Classificação Automática, *Actas do III Colóquio de Estatística e Investigação Operacional*, SPEIO, Lagos.
- [10] Sokal, R. R. (1977): Clustering and Classification: Background and Current Directions, in *Clustering and Classification* (ed. J. Van Ryzin), Academic Press, New York, pp. 1-15.
- [11] Sokal, R. R., Sneath, P. H. A. (1963): *Principles of Numerical Taxonomy*, Freeman, San Francisco.
- [12] Soromenho, G. (1986): *Sobre a Comparação de Estruturas em Classificação Automática*, Dissertação de Mestrado, Faculdade de Ciências, Universidade de Lisboa.

CAPÍTULO 6
VALIDAÇÃO ESTOCÁSTICA EM
CLASSIFICAÇÃO HIERÁRQUICA

CAPÍTULO 6

VALIDAÇÃO ESTOCÁSTICA EM CLASSIFICAÇÃO HIERÁRQUICA

ÍNDICE

INTRODUÇÃO

6.1 PROCEDIMENTO DE SIMULAÇÃO

6.2 ALGUNS RESULTADOS

6.2.1 Comparação $M_0 - H_0$

6.2.2 Comparação $M_0 - M_i$

6.2.3 Comparação $M_i - H_i$

6.2.4 Comparação $H_0 - H_i$

6.3 CONCLUSÕES

CAPÍTULO 6

VALIDAÇÃO ESTOCÁSTICA EM CLASSIFICAÇÃO HIERÁRQUICA

INTRODUÇÃO

Validação é uma área de grande interesse em Classificação. Em 2.4 apresentou-se uma introdução a este problema, bem como à sua importância. O presente capítulo pretende ser uma contribuição para este tema e apresenta uma metodologia que tem por base a simulação estocástica.

De entre os vários problemas de validação este capítulo trata da avaliação da qualidade da estrutura hierárquica, sua robustez e/ou estabilidade e fiabilidade para com a estrutura dos dados. O efeito que a aplicação de um dado critério tem sobre determinado tipo de estrutura é também aqui analisado. Um objectivo será encontrar regras práticas de escolha de critérios de C.H.A..

Na primeira secção é apresentada a sequência metodológica desenvolvida, que pode ser usada global ou parcialmente. Esta experiência recorre ao conjunto de ferramentas desenvolvidas nos capítulos anteriores, nomeadamente a geração aleatória de árvores de classificação, os critérios de C.H.A. e a comparação de estruturas classificatórias.

A secção 6.2 é dedicada à apresentação de alguns resultados das simulações efectuadas e está organizada de acordo com as comparações contempladas pelo protocolo experimental.

Finalmente, na última secção, fazem-se alguns comentários ao trabalho desenvolvido e apresentam-se as conclusões que se considera serem mais relevantes.

6.1 PROCEDIMENTO DE SIMULAÇÃO

Esta metodologia experimental tem por base três componentes:

- ❑ Geração aleatória de árvores (dendrogramas).
- ❑ Algoritmos de Classificação Hierárquica Ascendente.
- ❑ Comparação de estruturas classificatórias.

A geração de árvores será aqui usada para a obtenção de matrizes de dados. No capítulo 4, foram propostas metodologias para a geração de árvores intervindo os três aspectos; topologia, níveis de fusão e atribuição de etiquetas aos elementos terminais, isto é dendrogramas.

Como já foi referido no capítulo 2, a um dendrograma associa-se uma matriz ultramétrica. Cada uma das matrizes ultramétricas assim obtidas desempenhará, neste estudo, o papel de matriz de dados. Deste modo, é possível controlar o tipo e a intensidade de estrutura nos dados.

Para a obtenção de diferentes tipos de estruturas de dados recorre-se ao processo de *Geração com Parâmetro de Forma*. Neste estudo consideraram-se dois valores de p , $p = 0.025$, que dá origem a árvores com uma estrutura predominantemente do tipo cadeia¹, e $p = 0.5$, que gera árvores com grande tendência para o tipo equilibrado. Entendeu-se que estes dois tipos de árvores eram importantes, quer pela sua associação a métodos clássicos bem conhecidos, quer pelas suas características extremas. Foi ainda considerado o processo de *Geração Uniforme*, para comparação.

Quanto aos algoritmos de *C.H.A.* foram considerados vários critérios de agregação de classes, directos e de metodologia *VL*. Por razões de parcimónia, optou-se por apresentar na secção 6.2 apenas o estudo de simulação associado a sete critérios, quatro de abordagem clássica e três de metodologia *VL*:

¹ Ver quadros de valores teóricos de probabilidade por tipos, 4.1 a 4.4.

- *SL* – Ligação Única (Single Linkage)
- *CL* – Ligação Completa (Complete Linkage)
- *AM* – Agregação pela Média
- *AMed* – Agregação pela Mediana
- *AVM* – Agregação pela Validade da Média
- *AVMed* – Agregação pela Validade da Mediana
- *AVB* – Método da família paramétrica¹ *VL* para $\varepsilon = 1$ e $\xi = 0.5$.

A outra componente importante neste procedimento de simulação é a comparação de estruturas de classificação e este assunto foi já estudado no capítulo 5. Neste estudo comparam-se as (pre)ordenações associadas às matrizes de dados ou às hierarquias usando o coeficiente gama de Goodman-Kruskal. Duas razões essenciais justificam a escolha deste coeficiente:

- Coincide com T_K para ordenações, que goza de um bom desempenho, como foi visto em 5.1.
- Na literatura é aconselhado o seu uso quando há um grande número de empates numerosos.

Note-se que as preordenações associadas a hierarquias provocam com facilidade patamares de grandes dimensões. A título de exemplo considere-se a hierarquia representada na Fig. 2.1, o tipo de preordenação que lhe está associada é (1,2,1,1,4,1,12,14). Tratando-se de um exemplo de dimensões muito reduzidas, permite no entanto mostrar que a reunião de duas classes com vários elementos cada, dá origem a um patamar com um comprimento significativo.

No que se segue, denota-se por M_{i0} a matriz quadrada de dados, de dimensão igual ao número de elementos a classificar, simétrica e em que o elemento $M_{i0}(a,b)$ representa a semelhança ou dissemelhança entre o par de elementos (a,b) . Como já se referiu, estas matrizes de dados são as matrizes ultramétricas, associadas aos dendrogramas gerados.

¹ A família paramétrica de métodos *VL* e o critério *AVB* em particular, foram definidos em 2.2.2.2.

Cada matriz de dados é submetida a perturbações aleatórias, de acordo com protocolo da experiência que se descreve em seguida.

Notações:

- Seja X uma variável aleatória $U(0,1)$ e considere-se a transformação linear, $\delta(2X-1)$, com $\delta \in \mathbb{R}^+$. A variável transformada tem distribuição $U(-\delta, \delta)$, donde:

$$E[\delta(2X-1)] = 0 \quad (6.1)$$

e

$$V[\delta(2X-1)] = 4\delta^2 V(X) = \frac{\delta^2}{3}. \quad (6.2)$$

- Considere-se uma matriz de dados ultramétrica, M_{i0} e um valor de coeficiente de perturbação δ . M_{ij}^δ representa a matriz após a j -ésima perturbação, de coeficiente δ , da matriz de dados M_{i0} , e os seus elementos são obtidos por:

$$M_{ij}^\delta(a, b) = M_{i0}(a, b) + \delta(2x - 1), \quad (6.3)$$

para todo o par (a, b) de elementos distintos do conjunto a classificar, e em que x é uma pseudo-concretização da variável aleatória X .

- No que se segue, H_{i0} e H_{ij}^δ representam as matrizes associadas às hierarquias, obtidas pela aplicação de um algoritmo de C.H.A., respectivamente a M_{i0} e a M_{ij}^δ .

- Para os coeficientes de comparação de Goodman-Kruskal, a notação¹ usada é:

- TGK_i^1 - valor do coeficiente obtido na comparação entre M_{i0} e H_{i0} .
- $TGK_{ij}^2(\delta)$ - valor do coeficiente obtido na comparação entre M_{i0} e M_{ij}^δ .

¹ Por razões de legibilidade optou-se por usar TGK em vez de T_{GK} para denotar o coeficiente de Goodman-Kruskal.

- $TGK_{ij}^3(\delta)$ - valor do coeficiente obtido na comparação entre M_{ij}^δ e H_{ij}^δ .
- $TGK_{ij}^4(\delta)$ - valor do coeficiente obtido na comparação entre H_{i0} e H_{ij}^δ .

Experiência:

- Para cada processo de geração, ou melhor para cada tipo de estrutura de matriz de dados, e para um número de elementos terminal fixo, executa-se a seguinte sequência:

1. Gerar uma matriz de dados inicial, M_{i0} .

2. Para cada um dos critérios de agregação:

2.1. Classificar $M_{i0} \rightarrow H_{i0}$.

2.2. Comparar M_{i0} com H_{i0} , $\rightarrow TGK_i^1$.

3. Fixar δ :

3.1. Perturbar $M_{i0} \rightarrow M_{ij}^\delta$.

3.2. Comparar M_{i0} com $M_{ij}^\delta \rightarrow TGK_{ij}^2(\delta)$.

3.3. Para cada um dos critérios de agregação:

3.3.1. Classificar $M_{ij}^\delta \rightarrow H_{ij}^\delta$.

3.3.2. Comparar M_{ij}^δ com $H_{ij}^\delta \rightarrow TGK_{ij}^3(\delta)$.

3.3.3. Comparar H_{i0} com $H_{ij}^\delta \rightarrow TGK_{ij}^4(\delta)$.

• 3.4. Repetir 3.1. a 3.3. n_1 vezes (para δ fixo).

4. Repetir 3. para diferentes valores de δ .

5. Repetir 1. a 4. n_2 vezes.

Nos resultados de simulações que se apresenta de seguida considerou-se $M = 10$.

6.2 ALGUNS RESULTADOS

Apresentam-se agora os resultados de algumas simulações efectuadas, de acordo com o esquema experimental anteriormente enunciado.

Foram considerados três tipos de estruturas geradas, associados à *Geração com Parâmetro de Forma* com $p = 0.025$ e $p = 0.5$, e ainda a *Geração Uniforme*. Para cada tipo de estrutura foram consideradas 30 réplicas ($n_2 = 30$), M_{i0} com $i = 1, \dots, 30$, e para cada réplica e para os vários valores de δ , foram consideradas 20 perturbações dessas matrizes ($n_1 = 20$), dando origem a M_{ij}^δ com $j = 1, \dots, 20$.

O comportamento dos vários critérios de agregação aqui considerados, será analisado segundo várias perspectivas de acordo com a metodologia exposta no ponto anterior.

6.2.1 Comparação M_0 - H_0

Para cada um dos três tipos de estrutura foram geradas 150 árvores¹ e criadas as respectivas matrizes ultramétricas, M_{i0} , às quais foram aplicados os diferentes métodos de *C.H.A.*, dando origem a hierarquias e consequentemente a novas matrizes ultramétricas, H_{i0} . Na comparação dos pares de matrizes (M_{i0}, H_{i0}) , foi usado o coeficiente de Goodman-Kruskal, já referido anteriormente, que neste tipo de comparação se denota por TGK^1 .

A aplicação dos critérios *SL*, *CL*, *AM* e *AMed* a qualquer matriz M_{i0} , produz sempre essa mesma estrutura ultramétrica, isto é, obtém-se sempre $TGK^1 = 1$.

Quanto aos critérios da metodologia *VL* aqui considerados, *AVM*, *AVMed* e *AVB*, tentou-se analisar o seu desempenho para os diferentes tipos de estruturas considerados.

No que se segue descreve-se a informação obtida para TGK^1 , nas diferentes situações. O quadro 6.1 contem algumas medidas de localização e dispersão dos valores de TGK^1 .

¹ Nos resultados aqui expostos foram consideradas 30 réplicas e quatro valores de δ (e a não perturbação), dando origem a 150 matrizes iniciais.

	Geração Uniforme			Cadeias			Equilibradas		
	AVM	AVMed	AVB	AVM	AVMed	AVB	AVM	AVMed	AVB
Média	0.8387	0.8852	0.7246	0.9662	0.9700	0.6320	0.7915	0.8437	0.8505
Desvio-padrão	0.0943	0.0943	0.4449	0.0000	0.0000	0.0197	0.0175	0.4327	0.1926
Mínimo	0.3788	0.4674	0.1343	0.2846	0.4249	0.0000	0.3868	0.2884	0.4175
P. 25 %	0.7108	0.8575	0.6330	0.9679	0.9786	0.5539	0.6491	0.7654	0.7395
P. 50 %	0.9130	0.9404	0.7490	1.0000	1.0000	0.6444	0.8514	0.9322	0.8925
P. 75%	0.9565	0.9673	0.8732	1.0000	1.0000	0.7379	0.9479	0.9777	0.9635
Máximo	1.0000	1.0000	1.0000	1.0000	1.0000	0.9111	1.0000	1.0000	1.0000

Quadro 6.1 – Medidas descritivas dos valores de *TGK*¹

Com os mesmos dados identificou-se para cada simulação qual o método, entre os três, que melhor recuperou a estrutura inicial (maior valor de *TGK*¹). O quadro seguinte apresenta, para cada tipo de estrutura, a frequência observada de cada método¹.

	Geração Uniforme	Cadeias	Equilibradas
AVM	8.7%	0.7%	5%
AVMed	44.7%	15.3%	33.3%
AVM≡AVMed	26%	84%	20%
AVB	20.6%	0%	41.7%

Quadro 6.2 - Frequência de cada método na melhor recuperação da estrutura inicial

¹ Foi introduzida uma linha com empate de *AVM* e *AVMed* , por ser uma situação bastante frequente.

Da análise dos quadros e figuras anteriores pode concluir-se que o desempenho do critério *AVB* depende bastante da estrutura dos dados, funcionando bem quando os dados têm uma estrutura em que as classes têm aproximadamente o mesmo número de elementos e funcionando significativamente pior quando os dados têm uma estrutura do tipo cadeia. Este aspecto de funcionamento do *AVB* pode ser explicado pela tendência que o método tem para formar árvores do tipo equilibrado.

Os resultados aqui apresentados permitem concluir que em termos gerais o método *AVMed* é o que funciona melhor, seguido de perto do método *AVM*. No caso particular das estruturas em cadeia os métodos *AVMed* e *AVM* têm um comportamento muito bom na recuperação da estrutura ultramétrica inicial.

6.2.2 Comparação $M_0 - M_i$

Como já foi referido, para cada tipo de geração de estrutura inicial, geraram-se 30 matrizes M_{i0} com $i = 1, \dots, 30$, e para cada uma destas matrizes e fixado um valor de perturbação, δ , foram realizadas 20 perturbações de M_{i0} , dando origem a M_{ij}^δ , com $j = 1, \dots, 20$.

Para cada par (M_{i0}, M_{ij}^δ) foi calculado o valor do coeficiente de Goodman-Kruskal, aqui representado por $TGK_{ij}^2(\delta)$, e posteriormente determinado o valor médio desses coeficientes associados às várias perturbações para cada matriz inicial gerada.

Assim, para cada tipo de estrutura, calculou-se a sucessão de valores médios de TGK^2 ,

$$\overline{TGK_i^2(\delta)} = \frac{1}{20} \sum_{j=1}^{20} TGK_{ij}^2(\delta) \text{ com } i = 1, \dots, 30.$$

No que se segue descreve-se esta sucessão de valores através de algumas medidas de localização e dispersão.

	Geração Uniforme			
	$\delta = 0.05$	$\delta = 0.15$	$\delta = 0.25$	$\delta = 0.5$
Média	0.9371	0.7070	0.5985	0.3468
Desvio-padrão	0.0613	0.1486	0.1228	0.0952
Mínimo	0.8024	0.4096	0.2891	0.1559
P. 25%	0.9108	0.5783	0.5360	0.2988
P.50%	0.9616	0.7273	0.5841	0.3595
P.75%	0.9835	0.8360	0.6809	0.4165
Máximo	0.9988	0.9240	0.8231	0.5083

Quadro 6.3 – $\overline{TGK^2(\delta)}$, na Geração Uniforme

	Cadeias			
	$\delta = 0.05$	$\delta = 0.15$	$\delta = 0.25$	$\delta = 0.5$
Média	0.9143	0.7443	0.6183	0.3638
Desvio-padrão	0.0644	0.1097	0.1195	0.0875
Mínimo	0.7824	0.4451	0.3166	0.1466
P. 25%	0.8597	0.6973	0.5483	0.2998
P.50%	0.9332	0.7658	0.6327	0.3559
P.75%	0.9604	0.8348	0.7101	0.4410
Máximo	0.9942	0.8929	0.7670	0.5112

Quadro 6.4 – $\overline{TGK^2(\delta)}$, na Geração com Parâmetro de Forma, com $p = 0.025$

	Equilibradas			
	$\delta = 0.05$	$\delta = 0.15$	$\delta = 0.25$	$\delta = 0.5$
Média	0.9193	0.7227	0.6320	0.3917
Desvio-padrão	0.0880	0.1345	0.1013	0.1001
Mínimo	0.6897	0.4311	0.3745	0.1998
P. 25%	0.9262	0.6112	0.5684	0.3266
P.50%	0.9514	0.7474	0.6202	0.3992
P.75%	0.9670	0.8234	0.7253	0.4533
Máximo	0.9963	0.9389	0.7919	0.5670

Quadro 6.5 – $\overline{TGK^2(\delta)}$, na Geração com Parâmetro de Forma, com $p = 0.5$

Da análise dos quadros anteriores podemos concluir que o comportamento de TGK^2 é semelhante para os três tipos de estruturas consideradas. Nas simulações realizadas foram considerados valores de $\delta > 0.5$ contudo estes davam origem a valores de TGK^2 bastante pequenos¹, o que nos levou a reter apenas os resultados para valores de $\delta \leq 0.5$.

6.2.3 Comparação $M_i - H_i$

Neste ponto analisa-se o comportamento do coeficiente TGK^3 para os três tipos de estruturas e para os critérios de agregação em estudo.

Cada matriz M_{ij}^δ é submetida aos algoritmos de classificação dando origem a hierarquias H_{ij}^δ . Para cada tipo de estrutura, para cada valor de δ e para cada critério de agregação a comparação dos pares $(M_{ij}^\delta, H_{ij}^\delta)$, com $i = 1, \dots, 30$ e $j = 1, \dots, 20$, dá origem a um conjunto de valores de $TGK_{ij}^3(\delta)$, com $i = 1, \dots, 30$ e $j = 1, \dots, 20$.

Analogamente ao que foi feito em 6.2.2 calculou-se, para cada matriz inicial gerada, o valor médio desses coeficientes associados às várias perturbações. Assim, para cada tipo de estrutura, para cada valor de δ e para cada critério de agregação, calculou-se a sucessão de

valores médios $\overline{TGK_i^3}(\delta) = \frac{1}{20} \sum_{j=1}^{20} TGK_{ij}^3(\delta)$, com $i = 1, \dots, 30$.

Nos quadros e figuras que se seguem apresentam-se algumas medidas descritivas do conjunto de valores $\overline{TGK_i^3}(\delta)$.

¹ Para $\delta = 0.8$ o valor máximo dos valores $\overline{TGK_i^2}(\delta)$ era já inferior a 0.33, o que, atendendo à fórmula do coeficiente, nos indica que o número de concordâncias é já inferior a duas vezes o número de discordâncias.

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9351	0.9401	0.9441	0.9429	0.7495	0.8265	0.5936
Desvio Padrão	0.0637	0.0667	0.0581	0.0594	0.1300	0.0943	0.2015
Mínimo	0.8118	0.7454	0.8142	0.8135	0.4229	0.5695	0.2845
P. 25%	0.9009	0.9101	0.9183	0.9183	0.6351	0.7972	0.4288
P. 50%	0.9653	0.9653	0.9682	0.9681	0.7947	0.8406	0.6099
P. 75%	0.9900	0.9900	0.9901	0.9902	0.8282	0.8912	0.7872
Máximo	0.9995	0.9995	0.9995	0.9995	0.9657	0.9687	0.8904

Quadro 6.6 – Medidas descritivas de $TGK_i^3(0.05)$, na Geração Uniforme

EXTREMOS E QUARTIS

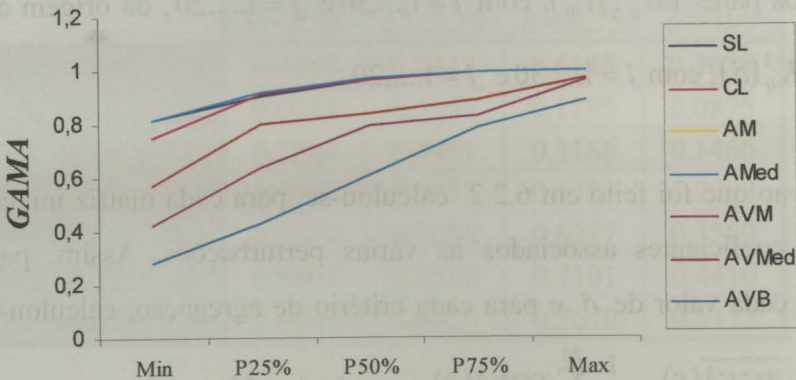


Fig 6.1 - $TGK_i^3(0.05)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.6981	0.7354	0.763135	0.7582	0.5844	0.6376	0.6166
Desvio Padrão	0.1384	0.1293	0.111248	0.1124	0.1373	0.1502	0.1366
Mínimo	0.4480	0.4989	0.552625	0.5474	0.3201	0.3113	0.3631
P. 25%	0.5963	0.6361	0.67052	0.6679	0.4666	0.5311	0.5256
P. 50%	0.6809	0.7269	0.748298	0.7427	0.5813	0.6522	0.5980
P. 75%	0.8398	0.8651	0.878754	0.8788	0.6782	0.7594	0.7043
Máximo	0.9146	0.9274	0.92967	0.9213	0.7975	0.8388	0.8656

Quadro 6.7 – Medidas descritivas de $TGK_i^3(0.15)$, na Geração Uniforme

EXTREMOS E QUARTIS

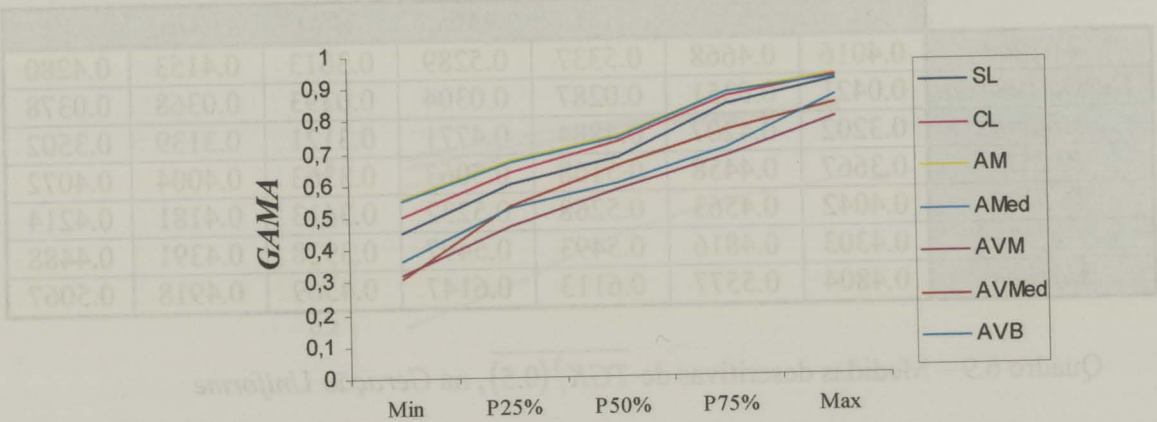


Fig. 6.2 - Extremos e quartis de $TGK_i^3(0.15)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.6142	0.6399	0.6840	0.6790	0.4958	0.5752	0.5823
Desvio Padrão	0.1117	0.1116	0.0920	0.0928	0.1019	0.0953	0.1203
Mínimo	0.3667	0.4526	0.5229	0.5098	0.3258	0.4031	0.3768
P. 25%	0.5575	0.5728	0.6353	0.6261	0.4318	0.5073	0.5041
P. 50%	0.5809	0.6184	0.6583	0.6572	0.4956	0.5686	0.5608
P. 75%	0.6915	0.7169	0.7482	0.7352	0.5481	0.6259	0.6752
Máximo	0.8255	0.8336	0.8425	0.8417	0.6723	0.7616	0.7955

Quadro 6.8 – Medidas descritivas de $TGK_i^3(0.25)$, na Geração Uniforme

EXTREMOS E QUARTIS

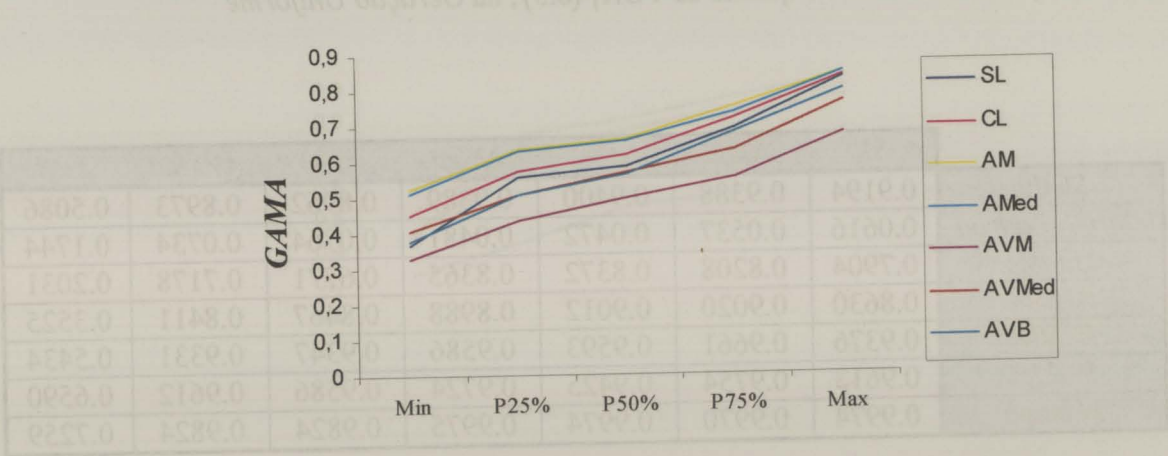


Fig. 6.3 - Extremos e quartis de $TGK_i^3(0.25)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.4016	0.4668	0.5337	0.5289	0.3613	0.4153	0.4280
Desvio Padrão	0.0421	0.0351	0.0287	0.0304	0.0393	0.0368	0.0378
Mínimo	0.3202	0.4207	0.4884	0.4771	0.3121	0.3139	0.3502
P. 25%	0.3667	0.4458	0.5168	0.5063	0.3363	0.4004	0.4072
P. 50%	0.4042	0.4563	0.5268	0.5232	0.3413	0.4181	0.4214
P. 75%	0.4303	0.4816	0.5493	0.5497	0.3928	0.4391	0.4488
Máximo	0.4804	0.5577	0.6113	0.6147	0.4569	0.4918	0.5067

Quadro 6.9 – Medidas descritivas de $TGK_i^3(0.5)$, na Geração Uniforme

EXTREMOS E QUARTIS

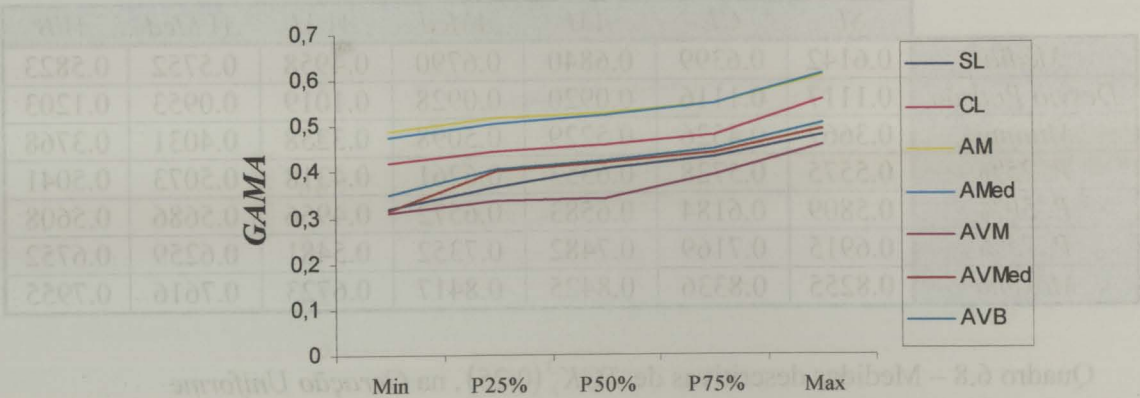


Fig. 6.4 - Extremos e quartis de $TGK_i^3(0.5)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9194	0.9388	0.9400	0.9389	0.8962	0.8973	0.5086
Desvio Padrão	0.0616	0.0537	0.0472	0.0481	0.0764	0.0734	0.1744
Mínimo	0.7904	0.8208	0.8372	0.8365	0.6771	0.7178	0.2031
P. 25%	0.8630	0.9020	0.9012	0.8988	0.8467	0.8411	0.3525
P. 50%	0.9376	0.9661	0.9593	0.9586	0.9347	0.9331	0.5434
P. 75%	0.9613	0.9754	0.9425	0.9724	0.9586	0.9612	0.6590
Máximo	0.9974	0.9970	0.9974	0.9975	0.9824	0.9824	0.7259

Quadro 6.10 – Medidas descritivas de $TGK_i^3(0.05)$, na G. Par. de Forma com $p = 0.025$

EXTREMOS E QUARTIS

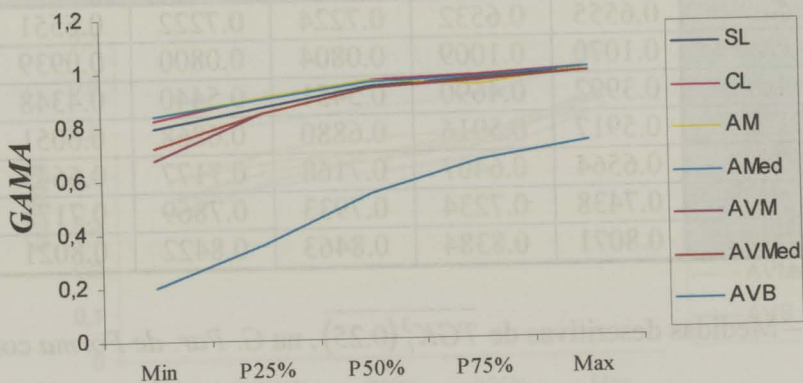


Fig. 6.5 - Extremos e quartis de $TGK_i^3(0.05)$, na G. Par. de Forma com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.7592	0.7707	0.8080	0.8064	0.7358	0.7457	0.5226
Desvio Padrão	0.1030	0.0960	0.0788	0.0804	0.1063	0.1009	0.1198
Mínimo	0.4523	0.5185	0.5874	0.5819	0.4100	0.4357	0.3222
P. 25%	0.7168	0.7064	0.7623	0.7552	0.6807	0.6921	0.4209
P. 50%	0.7660	0.7732	0.8083	0.8054	0.7484	0.7467	0.5295
P. 75%	0.8345	0.8307	0.8684	0.8690	0.8101	0.8151	0.5703
Máximo	0.9018	0.9234	0.9241	0.9244	0.8710	0.8724	0.7718

Quadro 6.11 – Medidas descritivas de $TGK_i^3(0.15)$, na G. Par. de Forma com $p = 0.025$.

EXTREMOS E QUARTIS

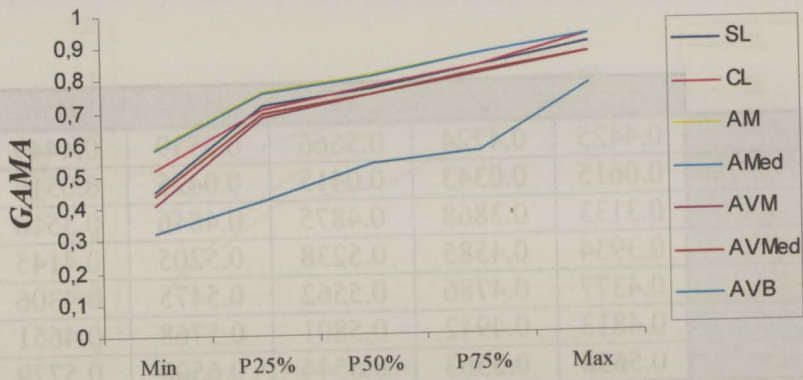


Fig. 6.6 - Extremos e quartis dos valores de $TGK_i^3(0.15)$, na G. Par. de Forma com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.6555	0.6532	0.7224	0.7222	0.6551	0.6586	0.5345
Desvio Padrão	0.1070	0.1009	0.0804	0.0800	0.0939	0.0909	0.1054
Mínimo	0.3992	0.4690	0.5451	0.5440	0.4348	0.4698	0.3541
P. 25%	0.5917	0.5916	0.6880	0.6868	0.6051	0.6043	0.4421
P. 50%	0.6564	0.6407	0.7168	0.7177	0.6645	0.6615	0.5160
P. 75%	0.7438	0.7234	0.7933	0.7869	0.7175	0.7147	0.6255
Máximo	0.8071	0.8384	0.8463	0.8422	0.8021	0.8032	0.7492

Quadro 6.12 – Medidas descritivas de $\overline{TGK_i^3(0.25)}$, na *G. Par. de Forma* com $p = 0.025$.

EXTREMOS E QUARTIS

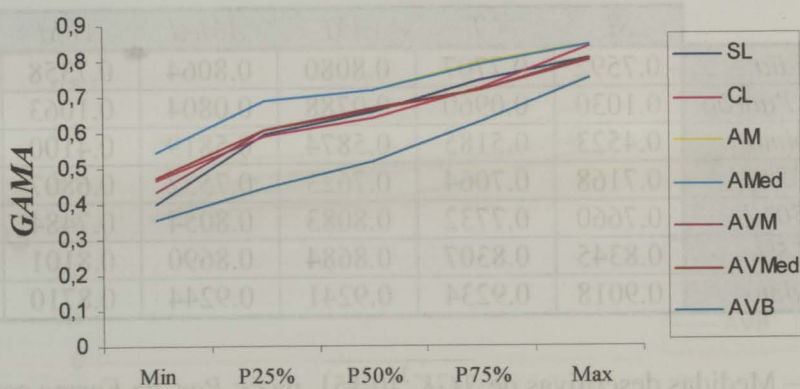


Fig. 6.7 - Extremos e quartis de $\overline{TGK_i^3(0.25)}$, na *G. Par. de Forma* com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.4425	0.4724	0.5566	0.5539	0.4449	0.4623	0.4187
Desvio Padrão	0.0615	0.0343	0.0415	0.0425	0.0517	0.0459	0.0504
Mínimo	0.3133	0.3868	0.4875	0.4846	0.3510	0.3943	0.3519
P. 25%	0.3934	0.4585	0.5238	0.5205	0.4145	0.4320	0.3799
P. 50%	0.4377	0.4786	0.5562	0.5475	0.4306	0.4487	0.4075
P. 75%	0.4813	0.4942	0.5801	0.5768	0.4651	0.4921	0.4431
Máximo	0.5638	0.5363	0.6544	0.6509	0.5779	0.5764	0.5382

Quadro 6.13 – Medidas descritivas de $\overline{TGK_i^3(0.5)}$, na *G. Par. de Forma* com $p = 0.025$.

EXTREMOS E QUARTIS

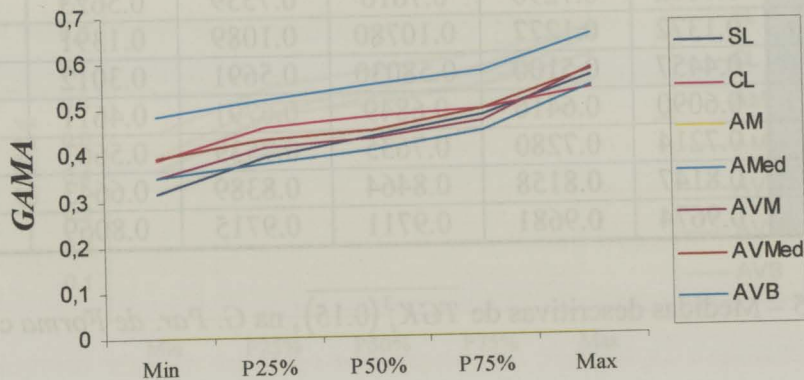


Fig. 6.8 - Extremos e quartis de $TGK_i^3(0.5)$, na *G. Par. de Forma* com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9189	0.9231	0.9300	0.9270	0.6166	0.7346	0.8239
Desvio Padrão	0.0888	0.0813	0.0728	0.0768	0.1948	0.1879	0.1142
Mínimo	0.6860	0.7300	0.7477	0.7451	0.2259	0.2395	0.4997
P. 25%	0.9341	0.9341	0.9366	0.9305	0.5151	0.5936	0.7637
P. 50%	0.9482	0.9523	0.9532	0.9534	0.5859	0.8052	0.8442
P. 75%	0.9703	0.9644	0.9730	0.9729	0.8018	0.8835	0.9093
Máximo	0.9978	0.9978	0.9978	0.9978	0.9430	0.9591	0.9918

Quadro 6.14 – Medidas descritivas de $TGK_i^3(0.05)$, na *G. Par. de Forma* com $p = 0.5$

EXTREMOS E QUARTIS

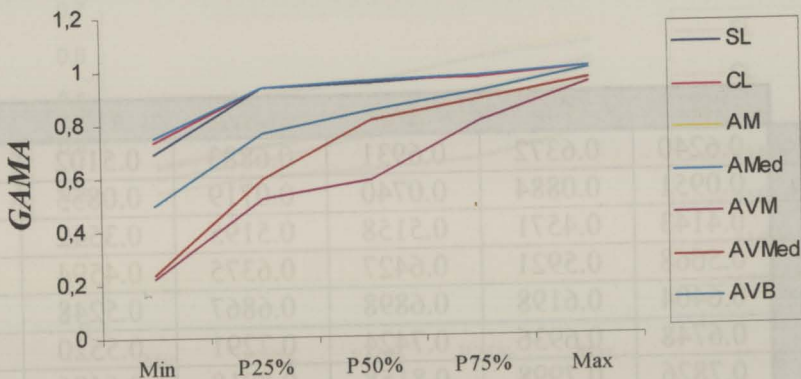


Fig. 6.9 - Extremos e quartis de $TGK_i^3(0.05)$, na *G. Par. de Forma* com $p = 0.5$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.7112	0.7290	0.7616	0.7559	0.5623	0.6478	0.6578
Desvio Padrão	0.1372	0.1277	0.10780	0.1089	0.1391	0.1390	0.1310
Mínimo	0.4457	0.5100	0.58030	0.5691	0.3012	0.3551	0.4181
P. 25%	0.6090	0.6416	0.6849	0.6791	0.4611	0.5597	0.5569
P. 50%	0.7214	0.7280	0.7635	0.7539	0.5663	0.6729	0.6881
P. 75%	0.8147	0.8158	0.8464	0.8389	0.6643	0.7493	0.7583
Máximo	0.9674	0.9681	0.9711	0.9715	0.8069	0.9045	0.9037

Quadro 6.15 – Medidas descritivas de $TGK_i^3(0.15)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

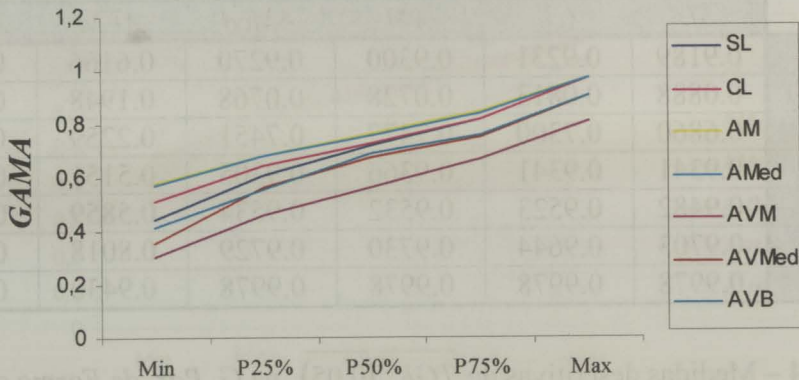


Fig. 6.10 - Extremos e quartis de $TGK_i^3(0.15)$, na G. Par. de Forma com $p = 0.5$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.6240	0.6372	0.6931	0.6883	0.5102	0.5898	0.6046
Desvio Padrão	0.0951	0.0884	0.0740	0.0719	0.0855	0.0779	0.0960
Mínimo	0.4143	0.4571	0.5158	0.5195	0.3542	0.4047	0.4151
P. 25%	0.5668	0.5921	0.6427	0.6375	0.4594	0.5600	0.5544
P. 50%	0.6404	0.6198	0.6898	0.6867	0.5248	0.5966	0.6109
P. 75%	0.6748	0.6936	0.7424	0.7291	0.5520	0.6368	0.6465
Máximo	0.7826	0.7998	0.8158	0.8018	0.6486	0.7220	0.7707

Quadro 6.16 – Medidas descritivas de $TGK_i^3(0.25)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

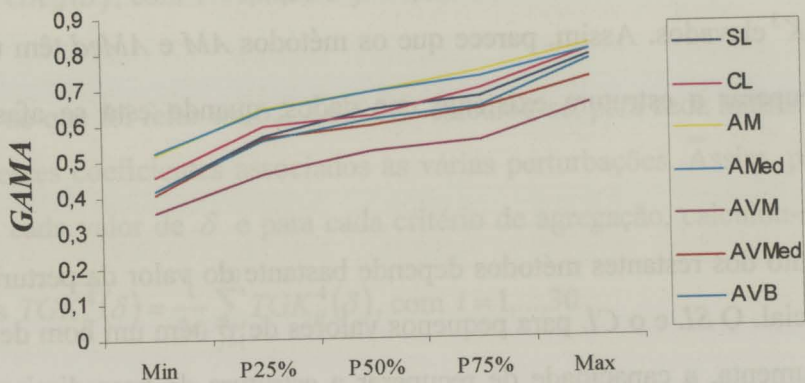


Fig 6.11 - Extremos e quartis de $TGK_i^3(0.25)$, na G. Par. de Forma com $p = 0.5$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.4200	0.4832	0.5486	0.5426	0.3543	0.4378	0.4476
Desvio Padrão	0.0628	0.0445	0.0445	0.0435	0.0364	0.0538	0.0608
Mínimo	0.3295	0.4119	0.4676	0.4720	0.2973	0.3509	0.3587
P. 25%	0.3659	0.4516	0.5169	0.5102	0.3312	0.3984	0.4010
P. 50%	0.4196	0.4698	0.5395	0.5361	0.3465	0.4308	0.4322
P. 75%	0.4621	0.5194	0.5925	0.5866	0.3677	0.4674	0.5051
Máximo	0.5415	0.5660	0.6360	0.6170	0.4261	0.5641	0.5820

Quadro 6.17 – Medidas descritivas de $TGK_i^3(0.5)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

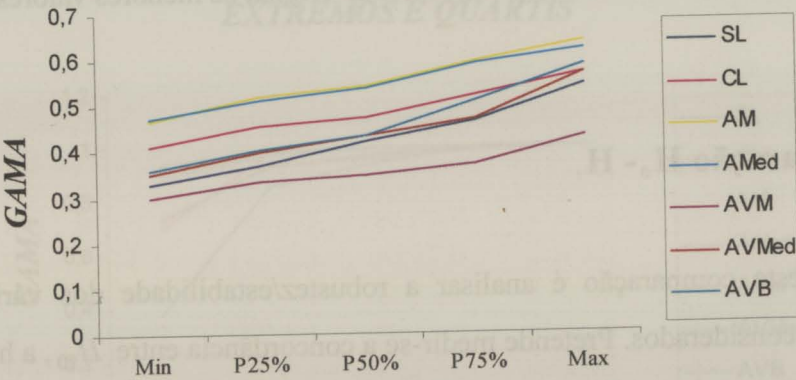


Fig. 6.12 - Extremos e quartis de $TGK_i^3(0.5)$, na G. Par. de Forma com $p = 0.5$

Dos quadros e figuras anteriores pode constatar-se que na generalidade os métodos que recuperam melhor a estrutura inicial são o *AM* e o *AMed*. De facto, quer para pequenos quer para grandes valores de δ , e para os três tipos de estruturas de dados, estes dois métodos têm valores de TGK^3 elevados. Assim, parece que os métodos *AM* e *AMed* têm uma capacidade grande em recuperar a estrutura existente nos dados, quando esta se afasta da estrutura ultramétrica.

O comportamento dos restantes métodos depende bastante do valor da perturbação e do tipo de estrutura inicial. O *SL* e o *CL* para pequenos valores de δ têm um bom desempenho, mas quando o δ aumenta, a capacidade de recuperar a estrutura de base diminui bastante. Em particular o *SL* quando a estrutura inicial é do tipo cadeia dá origem a valores de TGK^3 dos mais baixos.

O facto de os critérios de metodologia não *VL* constituírem o grupo que origina os valores mais elevados quando as perturbações são reduzidas está de acordo com as conclusões tiradas em 6.2.1. De facto para valores pequenos de δ a estrutura inicial é ainda muito próxima de uma ultramétrica e verificou-se que os métodos não *VL* recuperavam completamente uma estrutura inicial ultramétrica.

Dos critérios *VL* o que funciona melhor em geral é o *AVMed* e no caso da estrutura inicial ser do tipo equilibrado o *AVB* origina valores de TGK^3 semelhantes aos originados pelo *AVMed*. O *AVM* tem um funcionamento análogo ao *AVMed* para as estruturas do tipo cadeia, enquanto que para as estruturas do tipo equilibrado origina os menores valores de TGK^3 .

6.2.4 Comparação H_0 - H_i

O objectivo desta comparação é analisar a robustez/estabilidade dos vários critérios de agregação aqui considerados. Pretende medir-se a concordância entre H_{i0} , a hierarquia obtida da matriz de dados inicial, e H_{ij}^δ , a hierarquia obtida da matriz perturbada. Estuda-se o comportamento do coeficiente TGK^4 , para os três tipos de estruturas e para os critérios de agregação em estudo.

Para cada tipo de estrutura, para cada valor de δ e para cada critério de agregação temos a comparação dos pares (H_{i0}, H_{ij}^δ) , com $i = 1, \dots, 30$ e $j = 1, \dots, 20$, dando origem a um conjunto de valores de $TGK_{ij}^4(\delta)$, com $i = 1, \dots, 30$ e $j = 1, \dots, 20$.

Analogamente ao que foi feito em 6.2.2 e 6.2.3 calculou-se, para cada matriz inicial gerada, o valor médio desses coeficientes associados às várias perturbações. Assim, para cada tipo de estrutura, para cada valor de δ e para cada critério de agregação, calculou-se a sucessão de valores médios $\overline{TGK_i^4}(\delta) = \frac{1}{20} \sum_{j=1}^{20} TGK_{ij}^4(\delta)$, com $i = 1, \dots, 30$.

Nos quadros e figuras que se seguem apresentam-se algumas medidas descritivas do conjunto de valores $\overline{TGK_i^4}(\delta)$.

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9848	0.9682	0.9842	0.9768	0.9495	0.9484	0.8880
Desvio Padrão	0.0172	0.0618	0.0201	0.0293	0.0535	0.0629	0.1561
Mínimo	0.9351	0.6931	0.9053	0.8876	0.7265	0.7358	0.3752
P. 25%	0.9841	0.9766	0.9733	0.9671	0.9334	0.9222	0.8498
P. 50%	0.9900	0.9930	0.9922	0.9914	0.9583	0.9699	0.9436
P. 75%	0.9959	0.9951	0.9967	0.9947	0.9857	0.992	0.9974
Máximo	0.9993	0.9993	0.9993	0.9993	1.0000	1.0000	1.0000

Quadro 6.18 – Medidas descritivas de $\overline{TGK_i^4}(0.05)$, na Geração Uniforme

EXTREMOS E QUARTIS

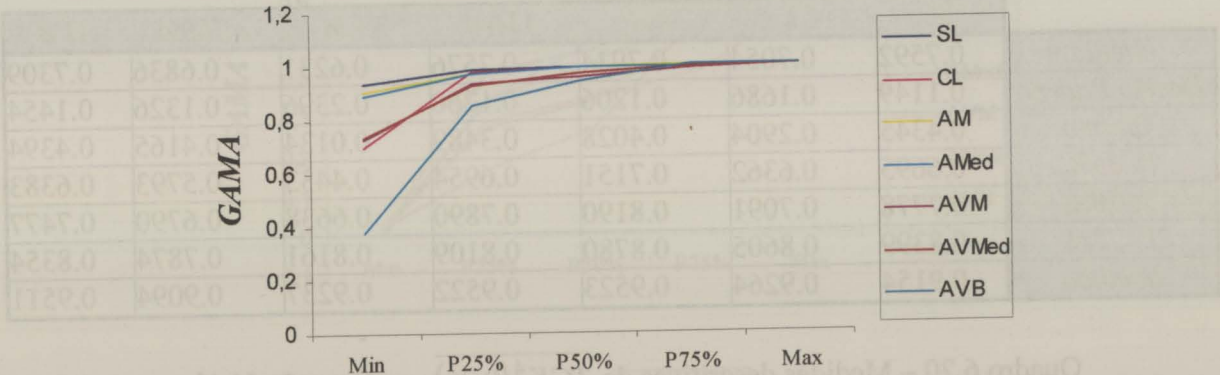


Fig. 6.13 - Extremos e quartis de $\overline{TGK_i^4}(0.05)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9085	0.8808	0.9249	0.8938	0.8522	0.8479	0.8332
Desvio Padrão	0.0970	0.1180	0.0874	0.0913	0.1502	0.1212	0.1635
Mínimo	0.5985	0.4883	0.5428	0.5485	0.4270	0.5009	0.3569
P. 25%	0.8655	0.8086	0.8996	0.8264	0.8218	0.8069	0.7425
P. 50%	0.9539	0.9265	0.9489	0.9192	0.9039	0.8685	0.8808
P. 75%	0.9723	0.9715	0.9778	0.9619	0.9602	0.9433	0.9480
Máximo	0.9952	0.9908	0.9908	0.9871	0.9893	0.9814	0.9980

Quadro 6.19 – Medidas descritivas de $TGK_i^4(0.15)$, na Geração Uniforme

EXTREMOS E QUARTIS

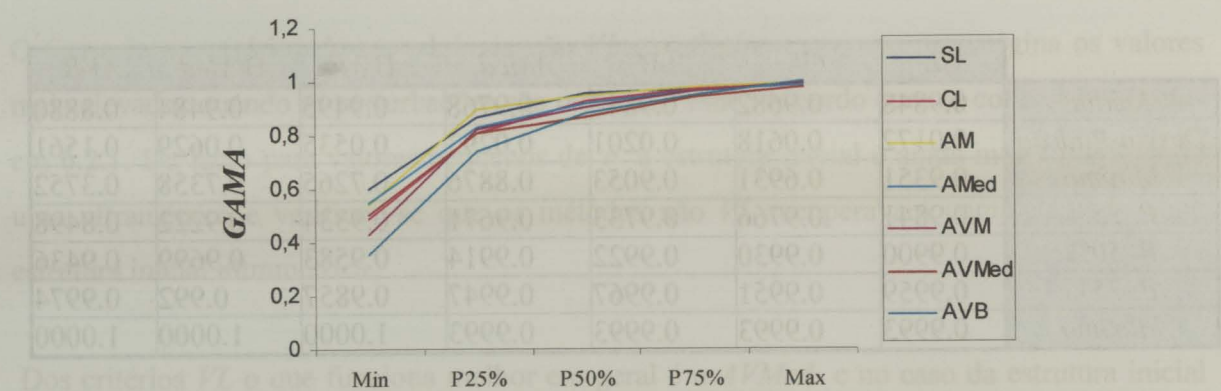


Fig. 6.14 - Extremos e quartis de $TGK_i^4(0.15)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.7592	0.7054	0.7934	0.7576	0.6231	0.6836	0.7309
Desvio Padrão	0.1149	0.1686	0.1206	0.1260	0.2396	0.1326	0.1454
Mínimo	0.4345	0.2904	0.4028	0.3489	0.0134	0.4165	0.4394
P. 25%	0.6695	0.6362	0.7151	0.6954	0.4433	0.5793	0.6383
P. 50%	0.7778	0.7091	0.8190	0.7890	0.6638	0.6790	0.7477
P. 75%	0.8399	0.8605	0.8780	0.8109	0.8161	0.7874	0.8354
Máximo	0.9154	0.9264	0.9523	0.9522	0.9237	0.9094	0.9511

Quadro 6.20 – Medidas descritivas de $TGK_i^4(0.25)$, na Geração Uniforme

EXTREMOS E QUARTIS

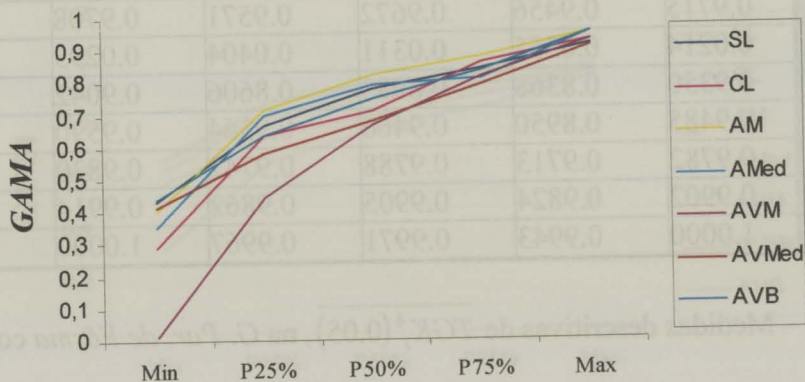


Fig. 6.15- Extremos e quartis de $TGK_i^4(0.25)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.4725	0.3543	0.4877	0.4433	0.3460	0.3761	0.4400
Desvio Padrão	0.1626	0.1326	0.1598	0.1500	0.2234	0.1866	0.1655
Mínimo	0.0000	0.0111	0.0145	0.0103	0.0230	0.0309	0.0002
P. 25%	0.4161	0.2934	0.4267	0.3489	0.1726	0.2795	0.3466
P. 50%	0.4718	0.3525	0.5221	0.4927	0.2705	0.3449	0.4421
P. 75%	0.5847	0.4193	0.5887	0.5440	0.5511	0.5294	0.5499
Máximo	0.7333	0.5904	0.6856	0.6765	0.7555	0.6486	0.6976

Quadro 6.21 – Medidas descritivas de $TGK_i^4(0.5)$, na Geração Uniforme

EXTREMOS E QUARTIS

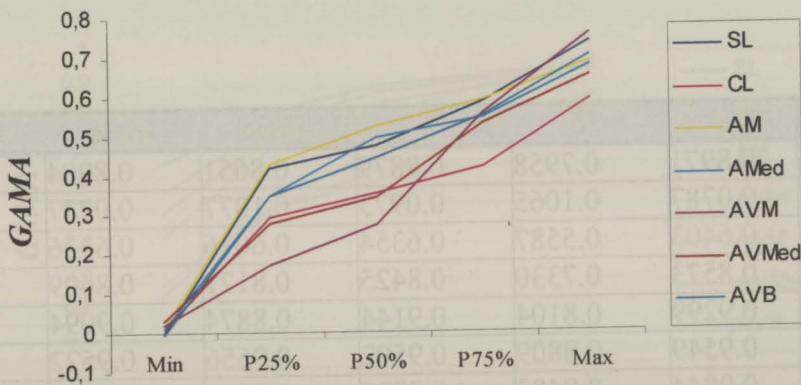


Fig. 6.16 - Extremos e quartis de $TGK_i^4(0.5)$, na Geração Uniforme

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9715	0.9456	0.9672	0.9571	0.9708	0.9564	0.8010
Desvio Padrão	0.0214	0.0555	0.0311	0.0404	0.0293	0.0398	0.1195
Mínimo	0.9350	0.8368	0.8983	0.8606	0.9042	0.8650	0.4901
P. 25%	0.9485	0.8950	0.9466	0.9364	0.9590	0.9410	0.7230
P. 50%	0.9782	0.9713	0.9788	0.9742	0.9830	0.9675	0.8279
P. 75%	0.9902	0.9824	0.9905	0.9862	0.9914	0.9868	0.8808
Máximo	1.0000	0.9943	0.9971	0.9967	1.0000	1.0000	0.9984

Quadro 6.22 – Medidas descritivas de $TGK_i^4(0.05)$, na G. Par. de Forma com $p = 0.025$

EXTREMOS E QUARTIS

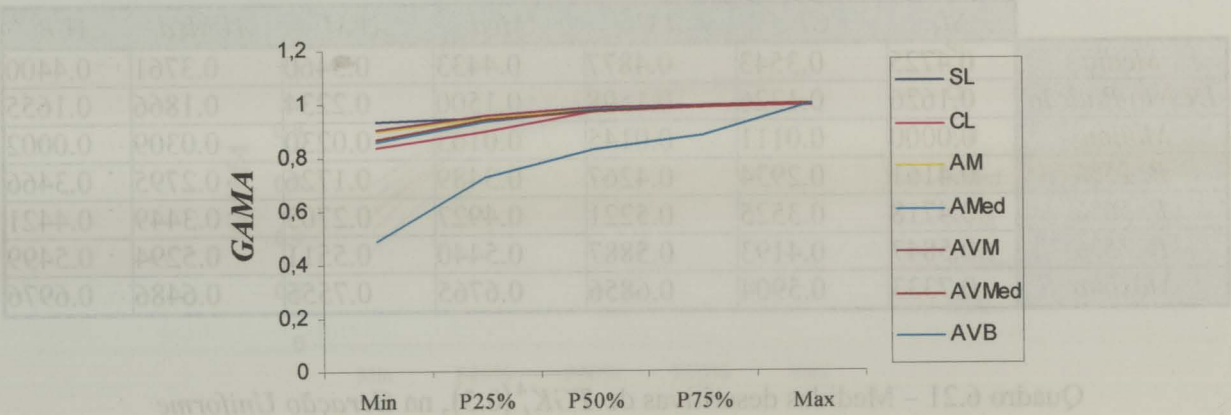
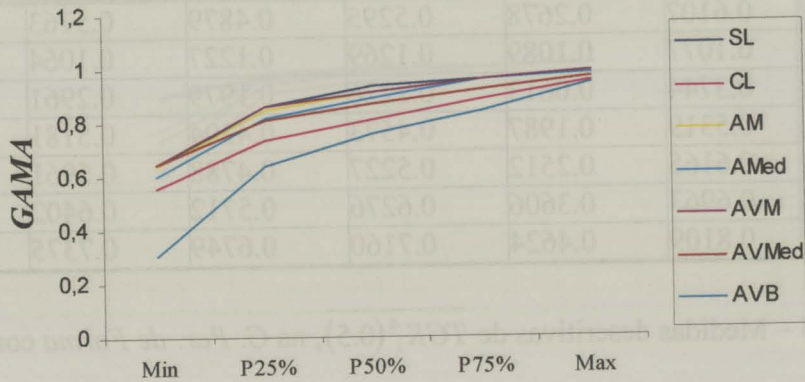


Fig. 6.17 - Extremos e quartis de $TGK_i^4(0.05)$, na G. Par. de Forma com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.8971	0.7958	0.8879	0.8651	0.8904	0.8547	0.7249
Desvio Padrão	0.0787	0.1065	0.0973	0.1077	0.0777	0.0892	0.1613
Mínimo	0.6403	0.5587	0.6354	0.6016	0.6436	0.6477	0.3062
P. 25%	0.8573	0.7330	0.8425	0.8171	0.8609	0.8078	0.6347
P. 50%	0.9299	0.8104	0.9144	0.8874	0.9094	0.8677	0.7591
P. 75%	0.9549	0.8809	0.9592	0.9556	0.9522	0.9176	0.8368
Máximo	0.9844	0.9483	0.9836	0.9766	0.9826	0.9639	0.9426

Quadro 6.23 – Medidas descritivas de $TGK_i^4(0.15)$, na G. Par. de Forma com $p = 0.025$

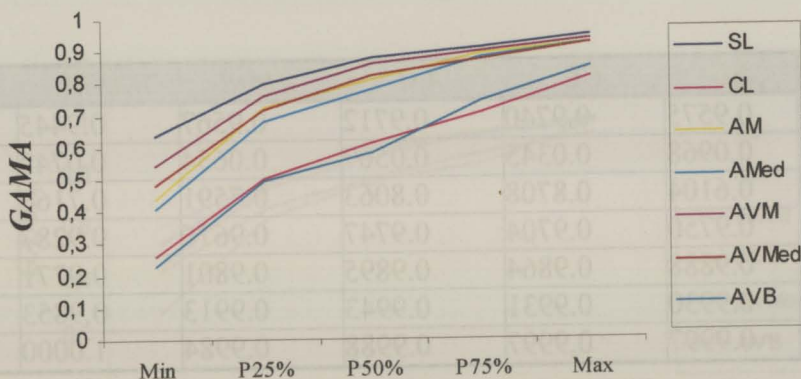
EXTREMOS E QUARTIS

Fig. 6.18- Extremos e quartis de $TGK_i^4(0.15)$, na *G. Par. de Forma* com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.8455	0.6009	0.7886	0.7540	0.8198	0.7765	0.5987
Desvio Padrão	0.0876	0.1508	0.1213	0.1308	0.1000	0.1103	0.1641
Mínimo	0.6367	0.2569	0.4386	0.4087	0.5447	0.4802	0.2254
P. 25%	0.7942	0.5014	0.7251	0.6773	0.7612	0.7148	0.4924
P. 50%	0.8781	0.6108	0.8103	0.7802	0.8590	0.8195	0.5808
P. 75%	0.9051	0.7034	0.8898	0.8781	0.8956	0.8674	0.7417
Máximo	0.9430	0.8127	0.9224	0.9222	0.9333	0.9204	0.8485

Quadro 6.24 – Medidas descritivas de $TGK_i^4(0.25)$, na *G. Par. de Forma* com $p = 0.025$

EXTREMOS E QUARTIS

Fig. 6.19 - Extremos e quartis de $TGK_i^4(0.25)$, na *G. Par. de Forma* com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.6107	0.2678	0.5295	0.4879	0.5763	0.5492	0.3455
Desvio Padrão	0.1077	0.1089	0.1269	0.1227	0.1064	0.1112	0.1792
Mínimo	0.3744	0.0613	0.2226	0.1979	0.2961	0.2094	0.0168
P. 25%	0.5319	0.1987	0.4573	0.4304	0.5181	0.4671	0.2223
P. 50%	0.6165	0.2512	0.5227	0.4788	0.5961	0.5599	0.3412
P. 75%	0.6963	0.3606	0.6276	0.5712	0.6402	0.6299	0.4083
Máximo	0.8109	0.4624	0.7160	0.6749	0.7375	0.7132	0.6751

Quadro 6.25 – Medidas descritivas de $TGK_i^4(0.5)$, na G. Par. de Forma com $p = 0.025$

EXTREMOS E QUARTIS

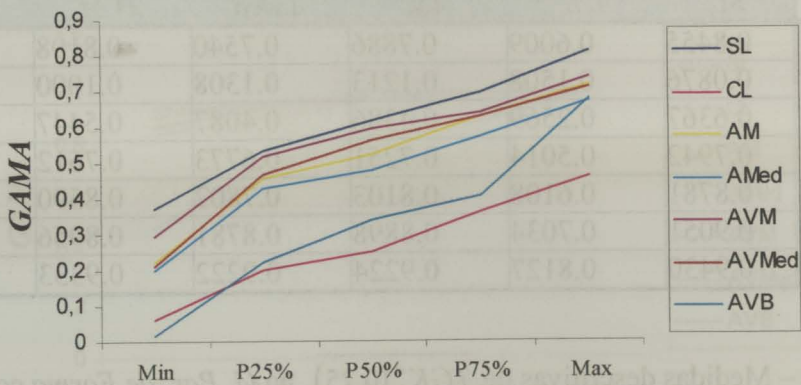


Fig. 6.20 - Extremos e quartis de $TGK_i^4(0.5)$, na G. Par. de Forma com $p = 0.025$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.9575	0.9740	0.9712	0.9567	0.9445	0.9164	0.9707
Desvio Padrão	0.0968	0.0345	0.0507	0.0654	0.0742	0.1012	0.0324
Mínimo	0.6104	0.8708	0.8063	0.7591	0.7166	0.5779	0.8873
P. 25%	0.9750	0.9704	0.9747	0.9676	0.9287	0.8928	0.9490
P. 50%	0.9888	0.9864	0.9895	0.9801	0.9771	0.9380	0.9835
P. 75%	0.9930	0.9931	0.9943	0.9913	0.9853	0.9885	0.9961
Máximo	0.9997	0.9997	0.9988	0.9984	1.0000	1.0000	1.0000

Quadro 6.26 – Medidas descritivas $TGK_i^4(0.05)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

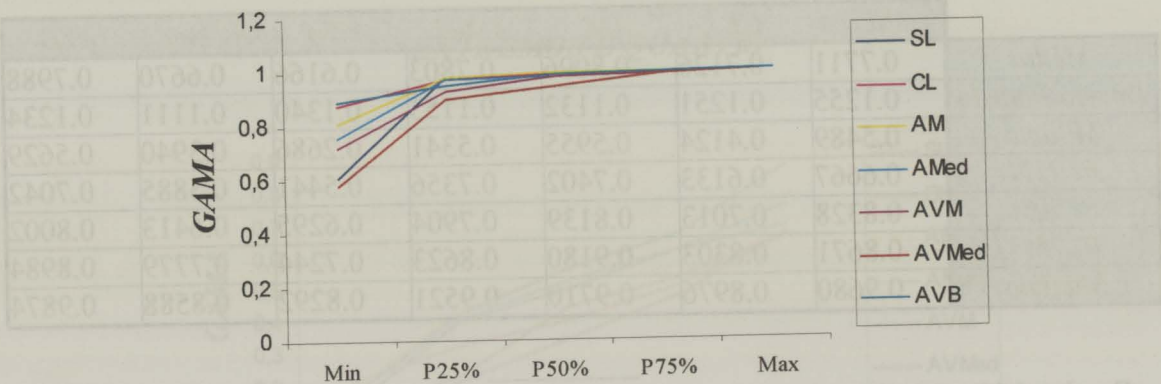


Fig. 6.21- Extremos e quartis de $TGK_i^4(0.05)$, na G. Par. de Forma com $p = 0.5$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.8350	0.8329	0.8708	0.8387	0.6857	0.7525	0.8367
Desvio Padrão	0.1301	0.1276	0.1053	0.1171	0.2848	0.1416	0.1581
Mínimo	0.5332	0.4540	0.6791	0.6011	-0.2706	0.4323	0.4529
P. 25%	0.7704	0.7373	0.7978	0.7627	0.6315	0.6678	0.7564
P. 50%	0.8902	0.8592	0.9218	0.8768	0.7624	0.7748	0.8976
P. 75%	0.9303	0.9384	0.9490	0.9348	0.8292	0.8654	0.9423
Máximo	0.9841	0.9762	0.9792	0.9688	0.9749	0.9303	0.9910

Quadro 6.27 – Medidas descritivas de $TGK_i^4(0.15)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

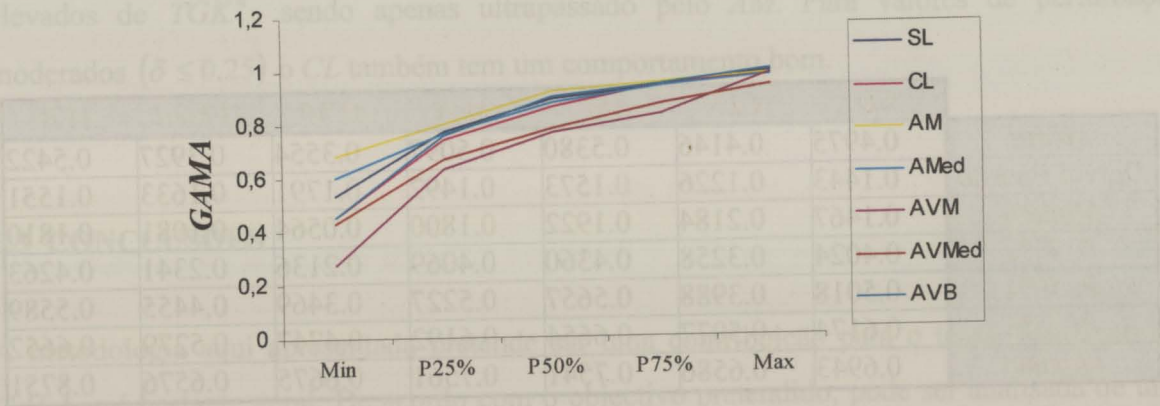


Fig. 6.22 - Extremos e quartis de $TGK_i^4(0.15)$, na G. Par. de Forma com $p = 0.5$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.7711	0.7126	0.8096	0.7803	0.6166	0.6670	0.7988
Desvio Padrão	0.1255	0.1251	0.1132	0.1153	0.1340	0.1111	0.1234
Mínimo	0.5489	0.4124	0.5955	0.5341	0.2686	0.3940	0.5629
P. 25%	0.6667	0.6133	0.7402	0.7356	0.5441	0.5885	0.7042
P. 50%	0.8328	0.7013	0.8139	0.7904	0.6295	0.6413	0.8002
P. 75%	0.8671	0.8303	0.9180	0.8623	0.7244	0.7779	0.8984
Máximo	0.9680	0.8976	0.9710	0.9521	0.8292	0.8588	0.9874

Quadro 6.28 – Medidas descritivas de $TGK_i^4(0.25)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

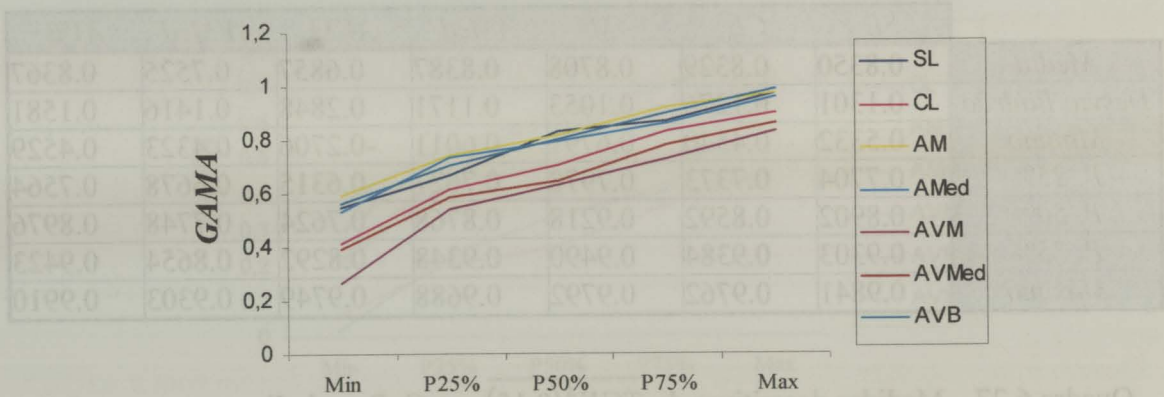


Fig. 6.23 - Extremos e quartis de $TGK_i^4(0.25)$, na G. Par. de Forma com $p = 0.5$

	SL	CL	AM	AMed	AVM	AVMed	AVB
Média	0.4975	0.4146	0.5380	0.5097	0.3554	0.3927	0.5422
Desvio Padrão	0.1443	0.1226	0.1573	0.1492	0.1791	0.1633	0.1551
Mínimo	0.1467	0.2184	0.1922	0.1800	0.0564	0.1081	0.1810
P. 25%	0.4024	0.3258	0.4360	0.4069	0.2136	0.2341	0.4263
P. 50%	0.5018	0.3988	0.5657	0.5227	0.3469	0.4455	0.5589
P. 75%	0.6174	0.5077	0.6654	0.6192	0.4747	0.5279	0.6652
Máximo	0.6943	0.6586	0.7541	0.7561	0.6675	0.6576	0.8751

Quadro 6.29 – Medidas descritivas de $TGK_i^4(0.5)$, na G. Par. de Forma com $p = 0.5$

EXTREMOS E QUARTIS

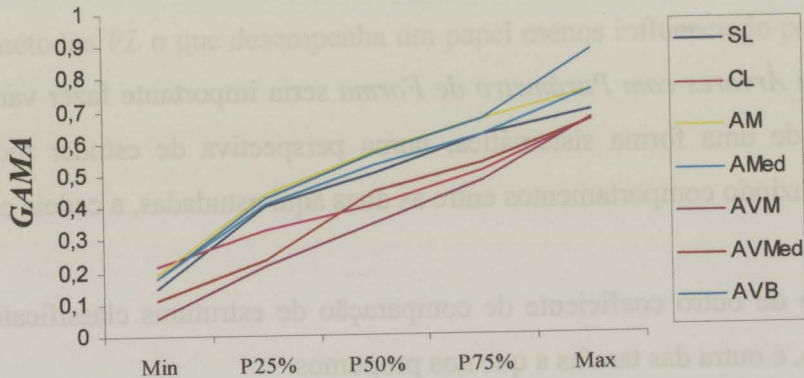


Fig. 6.24 - Extremos e quartis de $TGK_i^4(0.5)$, na *G. Par. de Forma* com $p = 0.5$

Da análise dos quadros e figuras anteriores, a primeira conclusão a tirar é que o desempenho dos métodos depende fortemente do tipo de estrutura existente nos dados. De uma forma global, pode afirmar-se que os métodos menos afectados pelas perturbações são o *SL* e o *AM* seguidos de perto do *AMed*. Quanto aos outros métodos o seu comportamento varia bastante com o tipo de estrutura.

No caso particular das estruturas predominantemente do tipo cadeia, os métodos *AVM* e *AVMed* têm um desempenho muito bom, enquanto que os métodos *CL* e *AVB* funcionam bastante mal. Quando em presença de árvores do tipo equilibrado, o *AVB* conduz a valores elevados de TGK^4 , sendo apenas ultrapassado pelo *AM*. Para valores de perturbação moderados ($\delta \leq 0.25$) o *CL* também tem um comportamento bom.

6.3 CONCLUSÕES

A metodologia aqui apresentada pretende ser uma contribuição para o tema Validação em *C.H.A.* e é bastante geral. De acordo com o objectivo pretendido, pode ser analisada de uma forma global ou usada apenas uma parte.

No conjunto das experiências apresentadas neste capítulo foi necessário proceder a algumas escolhas ou particularização de valores de parâmetros, de que é exemplo o número de elementos terminais das árvores geradas, M . Nos resultados da secção 6.2 considerou-se $M = 10$. É no entanto reconhecido que a variação da dimensão é um factor importante. Em trabalho futuro devem ser realizadas simulações com valores de $M \neq 10$.

Na *Geração de Árvores com Parâmetro de Forma* seria importante fazer variar o valor de p entre 0 e 0.5 de uma forma sistemática, numa perspectiva de estudar todos os tipos de estruturas, traduzindo comportamentos entre as duas aqui estudadas, a cadeia e a equilibrada.

A consideração de outro coeficiente de comparação de estruturas classificatórias, repetindo todo o processo, é outra das tarefas a que nos propomos.

O número de combinações possíveis, nas escolhas intervenientes, é enorme e os tempos de simulação e de análise dos resultados cresce de uma forma assustadora. Contudo, como já foi referido, algumas novas opções criteriosamente escolhidas devem ser concretizadas.

Esta metodologia permite, sob algumas condições, medir a eficiência dos critérios de agregação de classes bem como a estabilidade das estruturas por eles produzidas. Com um procedimento de simulação deste tipo é possível quantificar algumas características dos métodos de classificação, conhecidas da sua utilização prática.

Este estudo permite afirmar que os métodos de *C.H.A.*, aqui usados, de metodologia não *VL*, quando aplicados a uma matriz de dados ultramétrica, recuperam completamente essa mesma estrutura, enquanto que os métodos *VL* produzem hierarquias cujas estruturas podem ser um pouco diferentes. Este aspecto não é no entanto preocupante porque, como já anteriormente se referiu, os métodos *VL* e os métodos clássicos têm princípios subjacentes diferentes.

Quando a matriz de dados se afasta da estrutura ultramétrica os métodos que conduzem a valores de TGK^3 , concordância entre os dados e a estrutura obtida na classificação, mais elevados são os *AM* e *AMed*, seguidos do *SL*. O *CL* aproxima-se do comportamento do *SL* para valores de $\delta \leq 0.25$, não resistindo a valores de δ superiores. Dos critério *VL* o que funciona melhor em geral é o *AVMed*, mas quando a estrutura dos dados se aproxima do tipo cadeia o *AVM* e o *AVMed* têm comportamentos igualmente bons. Para estruturas do tipo equilibrado é a vez do *AVB* ter um bom desempenho.

Quando se comparam os pares $(H_{i0}, H_{ij}^{\delta})$, também aqui se constata que, a estabilidade das estruturas produzidas por alguns dos critérios, depende fortemente do tipo de estrutura existente nos dados. Em geral os métodos mais estáveis são o *AM*, *SL* e *AMed* e o menos estável é o *CL* seguido do *AVM*. Quando se consideram as árvores do tipo cadeia é a vez do *AVM* ser muito estável, e nas árvores equilibradas o *AVB* dá origem a valores de TGK^4 elevados. Dos métodos *VL* o que desempenha um papel menos influenciado pela estrutura é o *AVMed*.

CAPÍTULO 7
NOTAS FINAIS E
DESENVOLVIMENTOS FUTUROS

CAPÍTULO 7

NOTAS FINAIS E DESENVOLVIMENTOS FUTUROS

ÍNDICE

7.1 NOTAS FINAIS

7.1.1 Aspectos Computacionais

7.1.2 Conclusões

7.2 DESENVOLVIMENTOS FUTUROS

CAPÍTULO 7

NOTAS FINAIS E DESENVOLVIMENTOS FUTUROS

7.1 NOTAS FINAIS

Ao longo deste trabalho foi diversas vezes referida a necessidade e a importância da validação e da comparação das estruturas classificatórias. Vários estudos têm sido desenvolvidos nesta área, contudo o problema está ainda longe de ter uma solução aceite na globalidade. De facto não há nenhum algoritmo de Classificação Hierárquica, genericamente aceite como o melhor. Também não existe nenhuma abordagem ou procedimento global de validação dos resultados, sobre o qual os estudiosos estejam de acordo.

O tema aqui tratado é assim de grande actualidade e pertinência. A comprová-lo estão ainda as sessões dedicadas à Validação em Classificação, que têm feito parte dos programas de congressos e “workshops” internacionais, de que é exemplo recente a IFCS – International Federation of Classification Societies 2000.

Para a aplicação da metodologia aqui apresentada, foi necessário estudar e desenvolver três componentes distintas: algoritmos de C.H.A., a geração aleatória de dendrogramas e a comparação de estruturas de classificação.

Cada uma das referidas componentes tem subjacente, como se pôde antever pelos capítulos correspondentes, uma forte componente de programação. De seguida apresentam-se alguns aspectos associados a esta componente, que se pensa serem relevantes.

7.1.1 Aspectos Computacionais

Uma condição inicial ao conjunto de programas aqui desenvolvido, foi a de inteira autonomia de execução. De facto pretendia-se que os programas fossem integrados numa biblioteca já existente, *CLUSTER*, que obedece a essas características. Deste modo o recurso a bibliotecas numéricas, como por exemplo a *NAG*, foi desaconselhado.

Os referidos programas incluem as seguintes componentes:

- Programação e integração, na biblioteca *CLUSTER*, dos novos métodos de C.H.A. apresentados no capítulo 3.
- Algoritmos de geração aleatória de árvores de classificação, associados às metodologias propostas no capítulo 4.
- Comparação dos diferentes tipos de estruturas classificatórias envolvidas, recorrendo aos vários coeficientes estudados no capítulo 5. A comparação entre a (pre)ordenação inicial e a preordenação associada à hierarquia é efectuada a todos os níveis de agregação, ou nos níveis considerados mais significativos.
- Programação da metodologia de validação estocástica apresentada no capítulo 6, englobando os três pontos anteriores.

Neste conjunto de programas desenvolvidos a dimensão máxima possível, do número de elementos a classificar, não foi uma preocupação, dado que a execução da metodologia de validação apresentada não comportaria elevados valores de m . Contudo foram realizadas experiências com $m = 150$.

Uma vasta gama de critérios de agregação de classes, de metodologias directa e *VL*, encontra-se disponível. Também para a função de comparação entre elementos há um número elevado de opções à escolha.

Espera-se que brevemente tenha lugar a divulgação da biblioteca de programas actualizada, versão *CLUSTER7*, devidamente acompanhada de um Manual de Utilização.

Conjuntos de dados, com tipologias variadas e com diferentes dimensões, foram considerados para testar a sequência de programas, em particular as novas metodologias propostas neste trabalho. Consideraram-se dados tradicionalmente usados na literatura como exemplos e dados reais provenientes de diferentes áreas.

7.1.2 Conclusões

Como primeira conclusão da metodologia de validação aqui apresentada, pode afirmar-se que o comportamento de um método classificatório depende fortemente do tipo e da intensidade da estrutura dos dados.

Verificou-se que quando os métodos directos actuam sobre dados ultramétricos, reproduzem sempre essa mesma estrutura. Quanto aos métodos *VL* parece ser o *AVMed* que melhor recupera essa mesma estrutura.

Quando a estrutura dos dados se afasta de uma estrutura ultramétrica são, em termos gerais, os critérios *AM* e *AMed* que apresentam a maior capacidade em recuperar a estrutura subjacente. Para as estruturas do tipo cadeia, os critérios *AVM* e *AVMed* têm um bom comportamento, enquanto que para as estruturas equilibradas é a vez do *AVB* se colocar entre os melhores.

Quanto à comparação da hierarquia inicial com a hierarquia obtida após perturbação, pode concluir-se que em termos gerais os métodos mais resistentes são o *SL*, o *AM* e o *AMed*. Novamente quando as estruturas de base são do tipo cadeia, os métodos *AVM* e *AVMed* revelam-se bastante resistentes e no caso associado a árvores equilibradas o *AVB* tem um bom desempenho.

7.2 DESENVOLVIMENTOS FUTUROS

Algumas das perspectivas de desenvolvimento futuro foram já apontadas ou sugeridas ao longo do trabalho. Pensa-se que muito há a fazer neste domínio, quer a nível teórico ou analítico quer experimental ou de simulação.

Como primeiro ponto de desenvolvimento, refere-se a aplicação da metodologia experimental de validação a dados reais, e a dados simulados com estruturas diferentes das aqui consideradas. Em particular a não estrutura deve ser considerada, para informar da capacidade dos vários métodos em produzir estruturas hierárquicas artificiais.

Um estudo pormenorizado das propriedades dos vários métodos de *C.H.A.* propostos, deve ser efectuado, a nível teórico e por recurso à simulação. O desenvolvimento de técnicas, que simplifiquem e tornem mais objectiva, a tarefa de interpretação dos resultados de uma classificação é pertinente, atendendo a que as existentes se revelam insuficientes.

No âmbito da geração aleatória de estruturas é importante explorar a potencialidade das metodologias aqui propostas, nomeadamente a *Geração com Parâmetro de Forma*. Uma análise à complexidade algorítmica dos dois processos desenvolvidos e a sua comparação com a complexidade da metodologia de *Permutação Dupla* é certamente importante.

A comparação de estruturas merece ainda ser alvo de investigação, numa tentativa de encontrar bons coeficientes, para avaliar e comparar as estruturas hierárquicas, que reflectam de forma conveniente a sua natureza.

O recurso à simulação parece ser um caminho apropriado, tanto mais que o acentuado desenvolvimento técnico-informático torna mais simples e mais poderoso o acesso ao cálculo.

ANEXO 1
DEMONSTRAÇÃO DO RESULTADO 3.4

ANEXO 1

DEMONSTRAÇÃO DO RESULTADO 3.4

Como foi referido no texto apresenta-se neste anexo, a título de exemplo, a demonstração do resultado 3.4.

Resultado 3.4

Para todos os valores de m e n inteiros positivos e $x \in]0.5, 1[$, verifica-se que:

$$\begin{aligned} \int_{2x-1}^x t^m (1-2x+t)^n dt = \\ = \sum_{i=0}^n (-1)^i \frac{m!n!}{(m+1+i)!(n-i)!} x^{m+1+i} (1-x)^{n-i} + (-1)^{n+1} \frac{m!n!}{(m+n+1)!} (2x-1)^{m+n+1}. \end{aligned} \quad (1)$$

Demonstração:

A demonstração será feita recorrendo ao princípio de indução matemática, sobre n .

1. Considere-se $n = 1$, pretende-se verificar que, para todo o m inteiro positivo:

$$\begin{aligned} \int_{2x-1}^x t^m (1-2x+t) dt = \\ = \sum_{i=0}^1 (-1)^i \frac{m!}{(m+1+i)!(1-i)!} x^{m+1+i} (1-x)^{1-i} + (-1)^2 \frac{m!}{(m+2)!} (2x-1)^{m+2} \end{aligned} \quad (2)$$

Usando integração por partes, no primeiro membro de (2) tem-se:



$$\begin{aligned} \int_{2x-1}^x t^m (1-2x+t) dt &= \frac{1}{m+1} t^{m+1} (1-2x+t) \Big|_{2x-1}^x - \int_{2x-1}^x \frac{1}{m+1} t^{m+1} dt = \\ &= \frac{1}{m+1} x^{m+1} (1-x) - \frac{1}{(m+1)(m+2)} x^{m+2} + \frac{1}{(m+1)(m+2)} (2x-1)^{m+2}, \end{aligned} \quad (3)$$

que coincide com o segundo membro de (2).

2. Admita-se que (1) é uma proposição verdadeira quando n assume um qualquer valor natural não superior a k . Prove-se que, sob esta condição, (1) ainda é verdade para $k+1$, isto é, que é verdadeira a seguinte proposição:

$$\begin{aligned} \int_{2x-1}^x t^m (1-2x+t)^{k+1} dt &= \\ &= \sum_{i=0}^{k+1} (-1)^i \frac{m!(k+1)!}{(m+1+i)!(k+1-i)!} x^{m+1+i} (1-x)^{k+1-i} + (-1)^{k+2} \frac{m!(k+1)!}{(m+k+2)!} (2x-1)^{m+k+2}. \end{aligned} \quad (4)$$

Usando integrando por partes em (4) tem-se:

$$\begin{aligned} \int_{2x-1}^x t^m (1-2x+t)^{k+1} dt &= \\ &= \frac{1}{m+1} t^{m+1} (1-2x+t)^{k+1} \Big|_{2x-1}^x - \int_{2x-1}^x \frac{k+1}{m+1} t^{m+1} (1-2x+t)^k dt, \end{aligned} \quad (5)$$

usando a hipótese de indução para $n = k$, (5) é igual a

$$\begin{aligned} \frac{1}{m+1} x^{m+1} (1-x)^{k+1} - \frac{k+1}{m+1} \left[\sum_{i=0}^k (-1)^i \frac{(m+1)! k!}{(m+2+i)!(k-i)!} x^{m+2+i} (1-x)^{k-i} + \right. \\ \left. + (-1)^{k+1} \frac{(m+1)! k!}{(m+k+2)!} (2x-1)^{m+k+2} \right] = \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{m+1} x^{m+1} (1-x)^{k+1} + \sum_{i=1}^{k+1} (-1)^i \frac{m!(k+1)!}{(m+1+i)!(k-i+1)!} x^{m+1+i} (1-x)^{k-i+1} + \\
 &\quad + (-1)^k \frac{m!(k+1)!}{(m+k+2)!} (2x-1)^{m+k+2} = \\
 &= \sum_{i=0}^{k+1} (-1)^i \frac{m!(k+1)!}{(m+1+i)!(k-i+1)!} x^{m+1+i} (1-x)^{k-i+1} + (-1)^k \frac{m!(k+1)!}{(m+k+2)!} (2x-1)^{m+k+2},
 \end{aligned}$$

que é igual a (4), concluindo 2.

A veracidade conjunta das proposições (2) e (4), demonstra o resultado 3.4.

ANEXO 2

***DISTRIBUIÇÕES EXACTA E APROXIMADA
DA ESTATÍSTICA C PARA ORDENAÇÕES***

ANEXO 2

DISTRIBUIÇÕES EXACTA E APROXIMADA DA ESTATÍSTICA C PARA ORDENAÇÕES

Neste anexo apresentam-se algumas tabelas obtidas no estudo de comparação da distribuição exacta com a distribuição aproximada da estatística C para ordenações, isto é, o número de concordâncias entre duas ordenações de igual comprimento.

A cada valor de M , entre 4 e 10, associou-se uma tabela que contém a função de probabilidade da lei exacta de C , f_M , a distribuição aproximada de acordo com a expressão (5.41), $\tilde{f}_M$, e ainda a diferença entre as duas funções.

O objectivo deste anexo é tornar mais clara a leitura da parte final do ponto 5.1.4 deste trabalho.

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0.041667	0.044715	-0.00305
1	0.125	0.109375	0.015625
2	0.208333	0.212958	-0.00462
3	0.25	0.265905	-0.0159
4	0.208333	0.212958	-0.00462
5	0.125	0.109375	0.015625
6	0.041667	0.044715	-0.00305

Quadro 1 – Distribuições exacta e aproximada de C para $M = 4$

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0.008333	0.013743	-0.00541
1	0.033333	0.029462	0.003871
2	0.075	0.06713	0.00787
3	0.125	0.120881	0.004119
4	0.166667	0.172032	-0.00537
5	0.183333	0.193504	-0.01017
6	0.166667	0.172032	-0.00537
7	0.125	0.120881	0.004119
8	0.075	0.06713	0.00787
9	0.033333	0.029462	0.003871
10	0.008333	0.013743	-0.00541

Quadro 2 – Distribuições exacta e aproximada de C para $M = 5$

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0.001389	0.004267	-0.00288
1	0.006944	0.007818	-0.00087
2	0.019444	0.018059	0.001385
3	0.040278	0.036283	0.003995
4	0.068056	0.063401	0.004655
5	0.098611	0.096357	0.002254
6	0.125	0.127372	-0.00237
7	0.140278	0.146443	-0.00617
8	0.140278	0.146443	-0.00616
9	0.125	0.127372	-0.00237
10	0.098611	0.096357	0.002254
11	0.068056	0.063401	0.004655
12	0.040278	0.036283	0.003995
13	0.019444	0.018059	0.001385
14	0.006944	0.007818	-0.00087
15	0.001389	0.004267	-0.00288

Quadro 3 – Distribuições exacta e aproximada de C para $M = 6$

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0.000198	0.001333	-0.00114
1	0.00119	0.002099	-0.00091
2	0.003968	0.004699	-0.00073
3	0.009722	0.009619	0.000104
4	0.019444	0.018004	0.001441
5	0.033532	0.030812	0.00272
6	0.051389	0.048214	0.003175
7	0.07123	0.068982	0.002248
8	0.090278	0.090242	3.54E-05
9	0.105357	0.107943	-0.00259
10	0.11369	0.118055	-0.00436
11	0.11369	0.118055	-0.00436
12	0.105357	0.107943	-0.00259
13	0.090278	0.090242	3.54E-05
14	0.07123	0.068982	0.002248
15	0.051389	0.048214	0.003175
16	0.033532	0.030812	0.00272
17	0.019444	0.018004	0.001441
18	0.009722	0.009619	0.000104
19	0.003968	0.004699	-0.00073
20	0.00119	0.002099	-0.00091
21	0.000198	0.001333	-0.00114

Quadro 4 – Distribuições exacta e aproximada de C para $M = 7$

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0.000025	0.000418	-0.00039
1	0.000174	0.000573	-0.0004
2	0.00067	0.001226	-0.00056
3	0.001885	0.00247	-0.00059
4	0.004315	0.004683	-0.00037
5	0.008507	0.008353	0.000154
6	0.014931	0.014019	0.000911
7	0.023834	0.022138	0.001696
8	0.035094	0.032892	0.002202
9	0.048115	0.045982	0.002133
10	0.06183	0.060483	0.001348
11	0.074826	0.074855	-2.9E-05
12	0.085565	0.087168	-0.0016
13	0.092659	0.095508	-0.00285
14	0.095139	0.098462	-0.00332
15	0.092659	0.095508	-0.00285
16	0.085565	0.087168	-0.0016
17	0.074826	0.074855	-2.9E-05
18	0.06183	0.060483	0.001348
19	0.048115	0.045982	0.002133
20	0.035094	0.032892	0.002202
21	0.023834	0.022138	0.001696
22	0.014931	0.014019	0.000911
23	0.008507	0.008353	0.000154
24	0.004315	0.004683	-0.00037
25	0.001885	0.00247	-0.00059
26	0.00067	0.001226	-0.00056
27	0.000174	0.000573	-0.0004
28	0.000025	0.000418	-0.00039

Quadro 5 – Distribuições exacta e aproximada de C para $M = 8$

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0.000003	0.000132	-0.00013
1	0.000022	0.000159	-0.00014
2	0.000096	0.000324	-0.00023
3	0.000306	0.000635	-0.00033
4	0.000785	0.00119	-0.0004
5	0.001731	0.002135	-0.0004
6	0.00339	0.00367	-0.00028
7	0.006038	0.006039	-1.7E-06
8	0.009937	0.009518	0.000419
9	0.015281	0.014364	0.000916
10	0.022131	0.020759	0.001372
11	0.030371	0.028729	0.001642
12	0.039669	0.038072	0.001597
13	0.049485	0.048315	0.00117
14	0.05911	0.058715	0.000396
15	0.067747	0.068328	-0.00058
16	0.074606	0.076143	-0.00154
17	0.079021	0.081255	-0.00223
18	0.080545	0.083035	-0.00249
19	0.079021	0.081255	-0.00223
20	0.074606	0.076143	-0.00154
21	0.067747	0.068328	-0.00058
22	0.05911	0.058715	0.000396
23	0.049485	0.048315	0.00117
24	0.039669	0.038072	0.001597
25	0.030371	0.028729	0.001642
26	0.022131	0.020759	0.001372
27	0.015281	0.014364	0.000916
28	0.009937	0.009518	0.000419
29	0.006038	0.006039	-1.7E-06
30	0.00339	0.00367	-0.00028
31	0.001731	0.002135	-0.0004
32	0.000785	0.00119	-0.0004
33	0.000306	0.000635	-0.00033
34	0.000096	0.000324	-0.00023
35	0.000022	0.000159	-0.00014
36	0.000003	0.000132	-0.00013

Quadro 6 – Distribuições exacta e aproximada de C para $M = 9$

Valores de c	$f_M(c)$	$\tilde{f}_M(c)$	$f_M(c) - \tilde{f}_M(c)$
0	0	0.000042	-4.1E-05
1	0.000002	0.000045	-4.2E-05
2	0.000012	0.000087	-7.5E-05
3	0.000043	0.000165	-0.00012
4	0.000121	0.000303	-0.00018
5	0.000294	0.000538	-0.00024
6	0.000633	0.000925	-0.00029
7	0.001237	0.001541	-0.0003
8	0.002231	0.002488	-0.00026
9	0.003759	0.003889	-0.00013
10	0.005972	0.005889	8.25E-05
11	0.009007	0.008637	0.000369
12	0.012964	0.01227	0.000694
13	0.017882	0.016883	0.000998
14	0.023714	0.022501	0.001213
15	0.030316	0.029046	0.00127
16	0.037437	0.036317	0.001121
17	0.044736	0.043981	0.000755
18	0.051796	0.05159	0.000206
19	0.05817	0.058615	-0.00044
20	0.063418	0.064505	-0.00109
21	0.067156	0.068756	-0.0016
22	0.0691	0.070986	-0.00189
23	0.0691	0.070986	-0.00189
24	0.067156	0.068756	-0.0016
25	0.063418	0.064505	-0.00109
26	0.05817	0.058615	-0.00044
27	0.051796	0.05159	0.000206
28	0.044736	0.043981	0.000755
29	0.037437	0.036317	0.001121
30	0.030316	0.029046	0.00127
31	0.023714	0.022501	0.001213
32	0.017882	0.016883	0.000998
33	0.012964	0.01227	0.000694
34	0.009007	0.008637	0.000369
35	0.005972	0.005889	8.25E-05
36	0.003759	0.003889	-0.00013
37	0.002231	0.002488	-0.00026
38	0.001237	0.001541	-0.0003
39	0.000633	0.000925	-0.00029
40	0.000294	0.000538	-0.00024
41	0.000121	0.000303	-0.00018
42	0.000043	0.000165	-0.00012
43	0.000012	0.000087	-7.5E-05
44	0.000002	0.000045	-4.2E-05
45	0	0.000042	-4.1E-05

Quadro 7 – Distribuições exacta e aproximada de C para $M = 10$



FACULDADE DE ENGENHARIA

UNIVERSIDADE DO PORTO

BIBLIOTECA



0000067133