Número Ótimo de Aglomerados Estocásticos

Thales Santos Teixeira

Dissertação apresentada ao Instituto de Matemática
e Estatística da Universidade de São Paulo para
obtenção do grau de Mestre em Estatística

Área de concentração: **Processos multivariados estocásticos**

Orientador: **Prof. Adilson Simonis**

- São Paulo, outubro de 2003 -

Este trabalho teve apoio parcial do CNPq Processo: 132457/2001-6

Este exemplar corresponde à redação final devidamente corrigida e defendida por Thales Santos Teixeira e aprovada pela comissão julgadora.

São Paulo, outubro de 2003.

Banca examinadora:

- Prof. Adilson Simonis (Orientador) IME-USP
- Profa. Mancy Garcia IMECC-UNICAMP
- Prof. Sérgio Wesler IME-USP

Resumo

O número ótimo de aglomerados estocásticos proposto pelo método GAP de Tibshirani et. al (2000) é um procedimento para a determinação do número ideal de aglomerados em uma base de dados. Estudamos aqui a escolha do número de aglomerados quando a base tem uma evolução estocástica markoviana no tempo. Propomos avaliar a eficácia do método Gap e adequá-lo ao procedimento de evolução temporal. Além disto, pretende-se evidenciar por meio de simulações de Monte Carlo que a alteração proposta leva à escolha de um valor ótimo com propriedades estatísticas e computacionais desejáveis.

Abstract

The optimum number of stochastic clusters proposed by the GAP method of Tibshirani et. al (2000) is a procedure for the determination of the ideal number of clusters in a database. We study the choice of the number of clusters when the database has a markovian stochastic evolution in time. We also propose to evaluate the efficiency of the Gap method and to adequate it to the time evolving procedure. Moreover, it is intended to find evidence by means of Monte Carlo simulations that the change proposed leads to the choice of an optimum value with desirable statistical and computational properties.

Sumário

Capítulo 1

Introdução ao tema e considerações iniciais

O presente trabalho discute o problema de como segmentar um conjunto de elementos pelo grau de parecença e como obter o número ótimo de subconjuntos. Esta área do conhecimento estatístico esta inclusa na denominada análise de aglomerados (trad. de cluster analysis) e constitui uma poderosa ferramenta de aprendizado não supervisionado, onde há a ausência de uma amostra previamente classificada. O problema em questão é a definição a priori do número de aglomerados que o conjunto de elementos será particionado.

A dissertação discute a literatura em busca de metodologias existentes, apresenta alguns algoritmos e mostra resultados para uma variante do recente método Gap proposto por Tibshirani et. al. em 2002.

O trabalho está dividido em 4 capítulos, além desta introdução, divididos da seguinte forma: *capítulo 2* é composto por uma breve introdução à técnica de análise de agrupamentos e seus principais pontos críticos. O *capítulo 3* retrata o problema específico da definição do número mais adequado, ou ideal, de aglomerados para se dividir um conjunto de elementos. Nesse, será dado um conceito de "ideal" e, apartir daí, mostraremos alguns métodos e seus algoritmos. Na segunda metade deste capítulo, mais especificamente na *seção 2.42*, é introduzido ao leitor o método GAP. Com o conceito de uma distribuição de referência, que constitui num ponto central deste trabalho, provamos uma importante propriedade da estatística GAP. No *capítulo 4*, propomos um processo estocástico Markoviano de evolução de um conjunto de partículas como laboratório experimental para

mostrar algumas caraterísticas importantes do estimador do numero de aglomerados pelo método GAP. Avaliando por meio de simulações o modelo com dinâmica estocástica, fomos capazes de conjecturar pequenas alterações no cálculo e algoritmo do GAP de modo a obter um estimador do número ótimo de aglomerados no tempo. E, ao final do capítulo comparamos as propriedades e os resultados obtidos por simulações de Monte Carlo deste procedimento alternativo, propondo inclusive uma regra de parada para a ínferência do número de aglomerados no tempo. Ao final, no *capítulo 5*, ressaltamos a importância do artigo que deu origem ao método Gap, tecemos algumas considerações finais a respeito dos principais entraves encontrados durante a execução da pesquisa e aproveitamos para traçar um possível futuro norteador para seguir com novas abordagens ao tema tratado nesta dissertação.

Capítulo 2

Introdução ao modelo de análise de aglomerado

A análise de "clusters" (ou de aglomerados, ou de conglomerados) é uma técnica essencialmente exploratória não supervisionada de dados em virtude de não exigir uma classificação ou distribuição a priori dos elementos amostrais. No entanto, esta particularidade não diminui a importância desta técnica de análise multivariada de dados. Isto se deve em razão da sua ampla aplicabilidade em geração de grupos internamente coesos e isolados externamente entre si. Esta definição de aglomerado não é de consenso na literatura, o que acabou por gerar diferentes algoritmos de análise de aglomerados. A nossa noção intuitiva de grupos ideais, por vezes acaba indo contra a idéia de coesão interna e isolamento externo, almejado na maioria das técnicas de análise de aglomerados. Na figura abaixo, a intuição nos sugere que há 3 grupos coesos e isolados em (a) , 2 grupos apenas isolados em (b) e 2 grupos apenas coesos em (c) . Veja a figura 1.0 a seguir.

Apesar de divergências quanto à definição de aglomerado, uma unanimidade é a busca de classificação de elementos ou objetos em grupos distintos, cada qual contendo objetos similares. Em suma, traduzindo isto em linguagem estatística: análise de aglomerado busca identificar grupos de elementos amostrais (ou populacionais) com baixa variância dentro de cada grupo e alta variância de grupo para grupo.

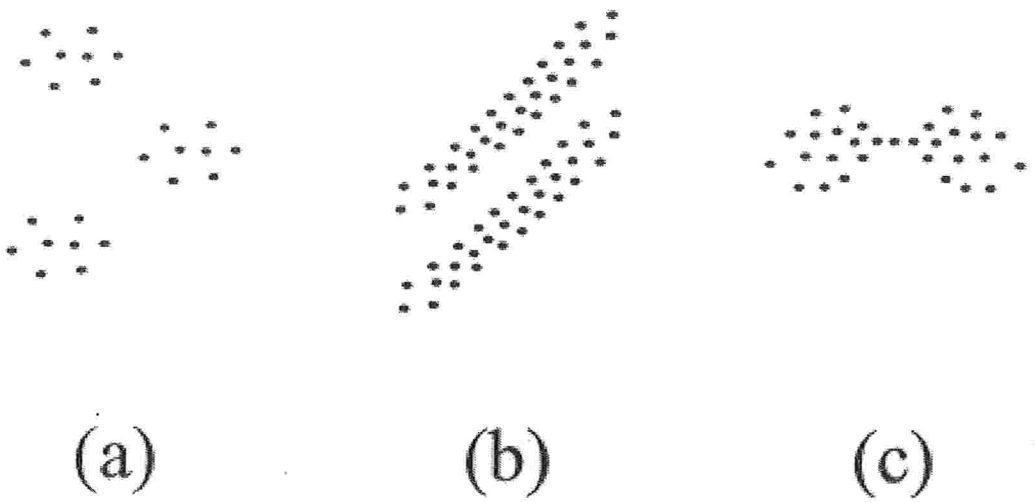

Figura 1.0 Grupos com e sem coesão interna e isolamento externo

Classificação, tal como é usado na literatura estatística, é distinto de técnica de discriminação, de identificação, de associação, de reconhecimento de padrões, de "clumping" e de análise de aglomerados, o tema dessa dissertação. Discriminação, ou análise discriminante, denomina o conjunto de técnicas de aprendizado supervisionado em que o objetivo é classificar um conjunto de elementos em grupos previamente conhecidos usando, para tal, um outro conjunto de elementos em que já se sabe o grupo em que cada elemento pertence. Daí o termo supervisionado, ou seja, tem-se a supervisão de uma amostra previamente classificada de elementos e objetiva-se encontrar a "lei" populacional por trás da classificação. Identificação ou associação é o que se faz após uma discriminação quando deparamos com um novo elemento e é necessário alocá-lo a um grupo pré-existente com base em uma regra. Já no reconhecimento de padrões, usa-se uma base de treinamento para se criar uma regra de associação e busca-se minimizar o erro de classificação. "Clumping", por sua vez, não define a metodologia de classificação, apenas denomina um conjunto de procedimentos para se descobrir grupos que em grande parte das vezes são sobrepostos uns aos outros. Por fim, análise de aglomerados é uma técnica de classificação não-supervisionada, em que não se conhece nem quais grupos e nem quantos grupos estão presentes em uma base de elementos populacionais ou amostrais.

Com a finalidade de descobrir quantos e quais grupos existem num determinado con-

junto de elementos, incluindo a questão de definir o que é um grupo, há que se avaliar quais características, as variáveis, que serão utilizadas no algoritmo que buscará otimizar um certo critério/função de classificação para a geração dos "melhores" aglomerados possíveis. Antes de avaliar os algoritmos existentes e os critérios de otimização, é necessário escolher as variáveis que irão representar um elemento e diferenciá-lo dos outros.

2.1 O problema da coleta de dados

Em um conjunto de n elementos, classificá-los em grupos homogêneos requer que se defina quais variáveis serão minimamente suficientes e estritamente necessárias para distinguir cada um dos n elementos de cada um dos $n-1$ elementos restantes. A medida de distinção, doravante denominada de dissimilaridade, ou contrariamente de parecença, denominada similaridade, deve representar fielmente o grau com que um certo elemento é parecido com outro ou é distinto desse. Um excesso de variáveis no cômputo desta dissimilaridade pode aproximar/afastar elementos outrora distintos/parecidos da mesma forma que o uso de um número muito reduzido de variáveis também podem fazer.

É notória a importância de buscar uma técnica de classificação no número mínimo de dimensões: primeiro, pelo aumento exponencial do tempo computacional exigido em uma análise de aglomerados e, segundo, pelo que se chama de "curse of dimensionality", algo como a maldição da dimensionalidade de criar novos relacionamentos numéricos em hiperespaços que superdimensionam relações inexistentes. Na análise discriminante, é possível calcular uma medida de quão separados estão os grupos e proceder-se a uma redução de dimensionalidade pela eliminação de variáveis cuja direção de variabilidade da mesma em relação aos dados não altere significativamente o critério de associação. Mas na análise de aglomerados isto não é possível pois não existe tal medida já que os grupos não são conhecidos a priori. Resta buscar uma redução de dimensionalidade para que tanto o tempo computacional quanto a estrutura matricial sejam tratáveis. Esta etapa inicial invariavelmente passa pela dificuldade de buscar um hiperespaço para projetar as variáveis sem sofrer grandes perdas na variabilidade inerente entre estes elementos. Entre algumas técnicas que podem ser utilizadas para mediar esta tarefa estão análises de componentes principais e escalonamento multidimensional. A primeira técnica é de uso amplo e trata a

redução de dimensionalidade pela transformação polinomial de componentes (nas direções de maior variabilidade) mas não é invariante a mudanças em escala das variáveis originais. No segundo caso, existem métodos não lineares para se avaliar o grau de perda de informação ao se reduzir um hiperespaço de D dimensões para D' dimensões. A referência (Hand 1981) trata de diversas situações neste contexto.

Dois critérios de avaliação desta redução são o critério de Kruscal, (Hand 1981):

$$\frac{\sum_{i<j}(d_{ij} - d'_{ij})^2}{\sum_{i<j} d_{ij}^2}$$

e o critério de Sammon, (Hand 1981):

$$\frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} (d_{ij} - d'_{ij})^2 / d_{ij}$$

onde d_{ij} é a distância de dissimilaridade entre elementos i e j no espaço D-dimensional, sendo d'_{ij} o mesmo cálculo no espaço D'-dimensional. No cálculo de dissimilaridade há diversas formas de definir a distância entre dois pontos no espaço euclidiano.

2.2 O problema do cálculo de dissimilaridades

Existe um problema crucial no início de uma análise de aglomerados: a escolha das variáveis que caracterizam tanto os elementos quanto os grupos que se busca criar. E atrelado a este problema está como definir uma função que calcula as $n(n-1)/2$ combinações de dissimilaridades d_{ij} entre os elementos i e $j \in 1, \ldots, N$, ou de similaridades s_{ij}. Na verdade a dissimilaridade é facilmente convertida em uma similaridade por uma função bijetora como a diferença ou o recíproco, da seguinte forma: $d_{ij} = 1 - s_{ij}$ ou por $d_{ij} = \frac{1}{s_{ij}}$.

Entre o conjunto de funções possíveis, destacam-se aquelas que são *métricas* e que obedecem às seguintes propriedades:

$$\forall\, i, j, k \in \mathbb{N};$$

$$d_{ij} \geq 0; d_{ij} = 0 \Longleftrightarrow i = j$$

$$d_{ij} = d_{ji}$$

$$d_{ij} \leq d_{ik} + d_{ki}$$

Já aqueles que são *coeficientes* de dissimilaridades obedecem às seguintes propriedades:

$$\forall\ i, j, k \in \mathbb{N};$$

$$d_{ij} \geq 0$$

$$d_{ij} = d_{ji}$$

$$d_{ij} \leq d_{ik} + d_{ki}$$

Uma medida de dissimilaridade enquadrada como métrica e, portanto também como coeficiente, é a generalização de algumas medidas conhecidas e amplamente utilizadas em técnicas de análise de aglomerado e em outras técnicas estatísticas desenvolvidas em espaços vetoriais. Conhecida como distância de *Minkowski*, ela é a generalização da distância euclidiana, com $\lambda = 2$ e da Manhattan ou city block, com $\lambda = 1$ em um espaço D-dimensional, sendo $x_i(l)$ o valor da coordenada do elemento i na variável l.

$$d_{ij} = \left[\sum_{l=1}^{D} |x_i(l) - x_j(l)|^{\lambda}\right]^{1/\lambda}$$

A escolha de uma função de cálculo de distância de dissimilaridade entre o elemento i e j envolve combinar todas as variáveis de diferentes tipos, intervalos e variabilidade em uma única medida escalar significativa para cada par de pontos no hiperspaço em estudo. Esta definição acaba por embutir na análise determinadas hipóteses que dizem respeito ao espaço em que serão feitos os agrupamentos. Por isso, há que se tomar em consideração a dificuldade de usar variáveis de diferentes tipos como as numéricas contínuas e discretas, as variáveis que são razões (proporções ou porcentagens), as nominais, as ordinais e as

binárias em uma análise de aglomerado. O resultado desta combinação de diferentes tipos de variáveis no cálculo de uma medida de dissimilaridade pode implicar em sua difícil interpretação numérica.

2.3 O problema do algoritmo: aglomeração ou partição

Computacionalmente, análise de aglomerados é uma sequência de cálculos matriciais seguida por uma sequência de decisões de alocação de modo a otimizar, em grande parte dos casos minimizar, uma estatística ou função dos dados originais.

Inicia-se com uma matriz $n \times p$ com n elementos nas linhas e p variáveis nas colunas. Em seguida, é calculado uma matriz $n \times n$ de dissimilaridades para todos os $n(n-1)/2$ pares de pontos distintos pela medida escolhida para todas as p variáveis. Após isto, há duas formas de se proceder.

Um caminho é o de alocar cada elemento para um aglomerado e ir juntando elemento a elemento, aglomerado a aglomerado, dos de menor distância para os de maior distância dada pela matriz. Algoritmos deste tipo que partem de n aglomerados e vão diminuindo o número de grupos a cada iteração até chegar a um grupo com todos os elementos são denominados de métodos hierárquicos. A única diferença entre eles diz respeito à forma como é calculada a distância de um aglomerado a outro. Este pode ser pela média das distâncias entre todos os pontos dos dois aglomerados (average linkage), pela distância dos elementos mais próximos (single linkage), pela distância dos elementos mais afastados (complete linkage), pela distância média dos grupos (centroid linkage), entre outros. A denominação 'hierárquico' advém do fato de que estes algoritmos aglomeram grupos a cada iteração criando grupos dentro de grupos, o que permite definir qualquer quantidade de aglomerados pois há uma hierarquia dentro de cada grupo. Esta propriedade permite que se desenhe um gráfico chamado dendrograma que mostra a proximidade de cada grupo no eixo das ordenadas e os elementos pertencentes a cada grupo em ordem de proximidade no eixo das abscissas.

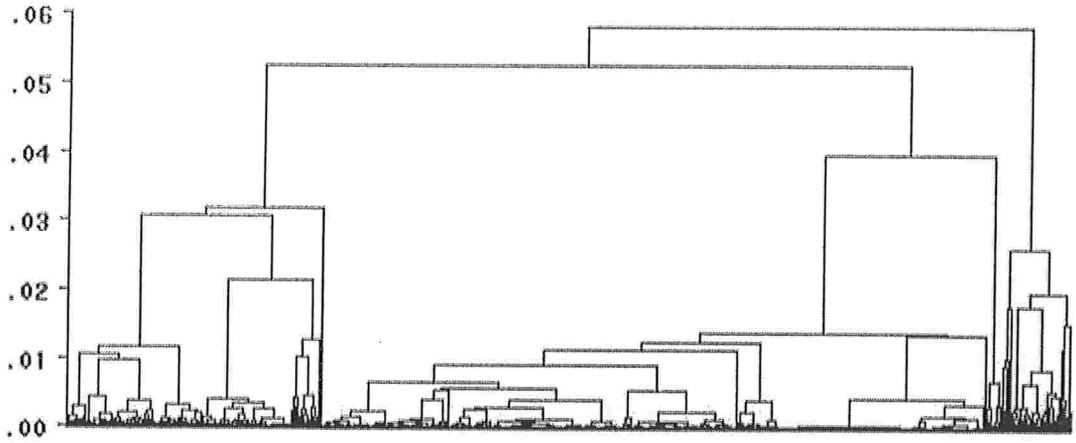

Figura 2.0 Exemplo de dendrograma de clusterização hierárquica com elementos no eixo da abscissa e medida de dissimilaridade na ordenada

Alguns métodos "top-to-bottom" iniciam-se com i aglomerados e aumentam o seu número em uma unidade a cada iteração, são os métodos aglomerativos hierárquicos, ao passo que outros "bottom-to-top" seguem o caminho inverso e são denominados particionamentos hierárquicos. O número k de aglomerados é uma decisão a posteriori que varia entre 1 e n.

O outro caminho é pela decisão de um número pré-definido de aglomerados k e pela escolha de um particionamento inicial com k elementos alocados cada um deles a um dos grupos, estes sendo os centróides iniciais. A cada iteração, calcula-se as $k(n-k)$ distâncias entre todos os outros pontos e os centróides, alocando-se cada ponto ao aglomerado cuja distância ao centróide seja a mínima dentre as k distâncias existentes. Após a alocação, recalcula-se os centróides como a média das coordenadas de todos os elementos de cada um dos novos aglomerados para cada variável. Com novos centróides, repete-se o processo iterativo até que uma certa medida de qualidade dos aglomerados não se altere mais ou que sua alteração se dê por menos de um valor mínimo pré-definido. Esta medida de qualidade, no caso de um dos algoritmos mais conhecidos e utilizados, o método das k-médias, é a soma das distâncias ao quadrado entre todos os pares distintos de pontos dentro de cada grupo, ponderada pelo número de elementos em cada grupo, somado em todos os grupos. Isto é,

$$W_k = \sum_{r=i}^{k} \frac{1}{n_r} \sum_{(i,j \in r)} d_{ij}^2$$

onde n_r é o número de elementos no cluster $r = 1, \ldots, k$. Este método, ao buscar o particionamento dos n elementos em k aglomerados de modo a minimizar a quantia W_k, não é um método otimizador global, apenas otimiza passo-a-passo (stepwise optimum) ao buscar o mínimo W_k. A única garantia de otimização global seria a avaliação do W_k para todas as partições, isto é,

$$\frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{n}{k} (k-i)^n$$

avaliações, algo computacionalmente demorado para grandes valores de n. O W_k é uma função monótona decrescente com o k, mas como os pacotes estatísticos possuem diferentes algoritmos para se tentar chegar ao W_k advindo da partição ótima, no caso da técnica das k médias, essa função pode ter leves incrementos de k para $k+1$. Em algoritmos do tipo "nested" onde 2 grupos se formam pela desagregação de um outro existente, isto nunca irá ocorrer. Um conjunto específico deste tipo é formado pelos algoritmos hierárquicos. Mas no caso de métodos de alocação iterativa como o do k médias, o objetivo é minimizar W_k para cada k, restrito a um número de iterações. Assim, situações anômalas em que $W_{k+1} > W_k$ podem ocorrer, pois a cada nova partição em k aglomerados, a partição para $k+1$ não é levada em consideração. Pela definição de W_k, com a distância euclidiana:

$$W_k := \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,j \in C_r} \sum_{p=1}^{P} \sqrt{(x_{ip} - x_{jp})^2},$$

podemos perceber que a somatória em $i, j \in C_r$, que é o r-ésimo aglomerado, é reduzida em alguns termos cruzados (de um aglomerado para outro) quando se aumenta o k mantendo-se o resto inalterado, devido à característica hierárquica. Portanto, em teoria W_k só pode diminuir com o aumento de k. Casos em que W_k pode crescer ligeiramente com o aumento de k, são infreqüentes e podem ser evitado com o aumento do número de iterações, pela escolha inteligente dos centróides iniciais ou pela eliminação de elementos que podem estar causando este problema. Se W_k fosse monótona decrescente, para um k fixado teríamos

um limite superior W_{k-1} e um limite inferior W_{k+1} para o qual se poderia avaliar o quão próximo uma partição qualquer em k aglomerados estaria do melhor ou pior caso possível. Em razão desta problemática o método das k-médias não fornece garantias de achar uma partição adequada, especialmente em grandes matrizes de dados pois o número de iterações necessárias cresce exponencialmente. A ampla aplicabilidade deste algoritmo se deve à simplicidade de interpretação e ao fornecimento de uma partição localmente ótima, pela otimização passo a passo, com extrema rapidez. Daí para uma partição ótima leva um longo tempo computacional. Evidências empíricas indicam que os aglomerados finais são altamente dependentes da partição inicial adotada, ou melhor, dos centróides iniciais escolhidos. Outros métodos de alocação iterativa de elementos entre aglomerados diferem na medida que se quer otimizar, podendo ser ela a soma dos traços ou o determinante da matriz de variância-covariância dentro de cada aglomerado. Uma crítica deste método é a possibilidade dele rodar várias iterações "sem sair" do lugar, alocando e desalocando os mesmos pontos para um aglomerado, caindo rapidamente num ótimo local em que não se sai mais. Isto se deve ao recálculo dos centróides a cada alocação de elementos. Numa iteração um ponto está mais perto de um centróide e no recálculo do centróide ele fica mais longe, mas ao se escolher outro ponto ele fica mais perto, e assim sucessivamente.

Observa-se que, seja a priori no caso dos algoritmos de alocação iterativa, seja a posteriori nos métodos aglomerativos e de particionamento, todos eles exigem a escolha do número k de aglomerados em algum momento para se ter o resultado completo da análise. Resultado este que indica a qual grupo cada um dos elementos pertence e quais são os limites, as médias e as variâncias dos grupos formados. Assim, depara-se com o problema de determinar o número de grupos populacionais existentes baseados numa amostra ou mesmo baseado em toda a população de elementos, tema central desse trabalho.

2.4 O problema do número de aglomerados

Em todos os algoritmos de análise de aglomerados é necessário fornecer o número de grupos em que serão divididos os elementos. No caso dos algoritmos de alocação iterativa, como o método das k-médias, o k é um parâmetro de entrada. E no caso de algoritmos hierárquicos, a escolha do número de grupos deve ser feito durante o algoritmo, mais precisamente depois

do cálculo da matriz de distâncias e da criação do dendrograma, este podendo ser utilizado como forma de avaliação visual dos "saltos" das medidas de dissimilaridade, na tomada de decisão do número k de aglomerados. Outros métodos computacionais advindos da teoria de grafos se baseiam em árvores de mínimo(a) alastramento ou envergadura (trad. de minimum spanning tree) e necessitam de um parâmetro de entrada como distância máxima para corte, o que leva diretamente à escolha do número de grupos, ou vice-versa.

A metodologia para a escolha do número de grupos, em especial em técnicas de aprendizado não-supervisionado, não é consensual e existem artigos como o de (Milligan 1985), que são verdadeiras compilações de mais de 30 técnicas e algoritmos para a determinação do número de aglomerados em uma base de dados. Dentre os mais utilizados, em função da simplicidade e da eficiência, estão o método de Calinski e Harabasz (1974), o de Duda e Hart (1973), o "Cubic Clustering Criterion" utilizado pelo software estatístico $SAS^{®}$, entre outros. A grande maioria destes procedimentos busca avaliar o ganho na diminuição da variabilidade interna dos grupos e o ganho no aumento da variabilidade entre elementos de grupos diferentes quando se aumenta o número de aglomerados de k para $k+1$. A medida que quantifica estas variabilidades é o que diferencia um procedimento do outro. Os mais tradicionais utilizam as matrizes W_k, matrizes de distâncias quadráticas ponderadas entre elementos dentro dos k aglomerados e B_k, matrizes de distâncias quadráticas ponderadas entre elementos entre os k aglomerados, para a mensuração da variabilidade interna dos grupos e da variabilidade entre elementos de grupos diferentes, respectivamente. Outras variantes de procedimentos de determinação do número de aglomerados avaliam funções, em k, de medidas que incluem o traço ou o determinante, ponderado ou não, das matrizes W_k, doravante denominada matriz de dispersão interna, e B_k, denominada matriz de dispersão externa.

Seja $C_p(i)$, $i \in k$, a coordenada do elemento i na variável p e seja $\text{centroide}_p(k)$ a média da variável p no agrupamento k. O elemento $w^k_{pp'}$ da matriz W_k é tal que:

$$centróide_p(k) \equiv \frac{1}{n_k} \sum_{i=1}^{n_k} C_p(i)$$

$$w^k_{pp'} = \sum_{i=1}^{n_k} (C_p(i) - centróide_p(k)) \times (C'_p(i) - centróide_{p'}(k))$$

onde $p, p' = 1, \ldots, P$; n_k é o número de elementos do cluster k

Assim, o que vários destes procedimentos buscam sistematizar é o conceito de avaliar quando o incremento do número de aglomerados escolhidos provoca uma forte queda na dispersão interna e um forte aumento da dispersão externa, seguido por uma estabilização das dispersões quando se aumenta o número k de aglomerados daí em diante. Esta sistemática está baseada na diferença da inclinação (a derivada se fosse derivável) da função que leva em conta W_k em função de k, $k \in \{1, n\}$. Graficamente, busca-se visualizar ao plotar esta função um "cotovelo" onde uma forte queda acompanha uma estabilização desta, ou de outra função similar. A seguir serão expostos alguns procedimentos tradicionais para a determinação do número k ideal de aglomerados, seus pontos fortes e fracos e, principalmente, as idéias intuitivas por detrás da metodologia, algumas das quais serão incorporadas na metodologia usada neste trabalho.

Capítulo 3

Técnicas tradicionais para determinação do número de aglomerados

De modo geral, as metodologias disponíveis para a aferição do número ideal de aglomerados em um conjunto de dados são caracterizadas por um algoritmo e por uma regra de parada. O cálculo de uma função de k, por exemplo, é o algoritmo. Já o critério de escolha é uma regra de parada quando se busca evitar calcular a função do algoritmo para todo k; o ponto ótimo, o máximo (ou mínimo) da função ou combinação da função com outros parâmetros também atuam como funções de decisão para a escolha do número de aglomerados.

Dentre as diversas técnicas existentes, (Milligan 1985), analisaram e aplicaram cerca de 30 das técnicas mais conhecidas e abrangentes que independem do algoritmo de aglomeração e expuseram seus resultados para diferentes bases de dados com "aglomerados naturais", ou seja, visivelmente espaçados uns dos outros. Muitas destas técnicas advém da estatística, da computação e da taxonomia usando a análise normal multivariada, a teoria dos grafos ou os métodos não-paramétricos, entre outros. A seguir, serão abordados alguns dos principais procedimentos escolhidos não tanto pela eficiência com que estes conseguiram determinar o número ótimo de aglomerados mas pelas idéias inovadoras que surgiram com a sua concepção.

3.1 Definições

A seguir, serão apresentadas as definições utilizadas neste capítulo.
n = número de observações, $n \in \mathbb{N}$
k = número de aglomerados, $k \in \{1, 2, 3, \ldots, N\}$
p = número de variáveis, $p \in \mathbb{N}$
X = matriz de dados, $X \in \mathbb{R}^N \times \mathbb{R}^P$
$\overline{X}$ = matriz de médias, $\overline{X} \in \mathbb{R}^k \times \mathbb{R}^P$
Z = matriz de indicadores de aglomerados, $Z \in \mathbb{R}^N \times \mathbb{R}^k$
onde $z_{ik} = \mathbf{1}_i\{k\}$ e $Z'Z$ é matriz diagonal com $\{z'z\}_{ii}$ igual ao número de observações no aglomerado i, $i = 1, \ldots, N$. Algumas vezes utilizaremos n e p para representar o valor em somatória de N e P quando estiver claro no contexto a notação.

Assim, $\overline{X} = (Z'Z)^{-1}Z'X$, é a matriz $k \times p$ com as médias de cada variável p para todos os pontos em cada aglomerado k. E, temos as seguintes matrizes $p \times p$ de dispersões entre os pontos:

$$T_k = X'X$$

é a matriz de dispersão total para os k aglomerados;

$$B_k = \overline{X}'Z'Z\overline{X}$$

é a matriz de dispersão entre os k aglomerados;

$$W_k = (X - Z\overline{X})'(X - Z\overline{X}) = X'X - \overline{X}'Z'Z\overline{X} = T_k - B_k$$

é a matriz de dispersão dentro dos k aglomerados;

Observe que, $traço(W_k) \equiv$ soma de distâncias euclidianas ao quadrado de cada observação ao centróide do aglomerado pertencente, tal como foi definido anteriormente em linguagem não matricial.

3.2 Técnicas descritivas

(Calinski and Harabasz 1974) - Esta técnica busca comparar a perda no acréscimo dos valores da matriz de dispersão externa, B_k, ao se aumentar k, com o ganho na redução dos

valores da matriz de dispersão interna, W_k ao se aumentar k. Para algoritmos hierárquicos o primeiro é monótono crescente em k e o segundo monótono decrescente, como já foi dito anteriormente. Assim, o que se busca é determinar o ponto em que estes valores se compensam para um k ideal. A avaliação do máximo da função abaixo define o k ideal.

$$CH(k) = \frac{\frac{Traço(B_k)}{(k-1)}}{\frac{Traço(W_k)}{(n-k)}}$$

Um ponto negativo desta técnica é que $CH(1)$ e $CH(n)$ não estão definidos, então a técnica não retornará o valor de $k = 1$ (ou n), os dados sempre serão particionados. Mesmo no caso de total dispersão ou homogeneidade dos dados, a aglomeração será sugerida pela técnica, produzindo $K > 1$ e $\neq n$, algo nem sempre defensável.

(Duda and Hart 1973) - Nesta técnica, o procedimento é calcular a razão entre a dispersão interna com 2 e com 1 aglomerado, para o caso do particionamento de todos os aglomerados existentes. Compara-se o valor calculado com um valor crítico específico, função da dimensionalidade e do número de elementos, entre outras variáveis, e se a razão for menor, rejeita-se a hipótese de um aglomerado. Ou seja, para cada decisão de particionar um conjunto de elementos, calcula-se o DH inicial. Se este for menor do que um valor de referência, particiona-se a base em dois grupos e cada um dos grupos em dois novos grupos, e assim sucessivamente até a não rejeição do teste.

$$DH = \frac{\sum_{i=1}^{k=2} d_{ij}^2}{d_{ij}^2}$$

Contrariamente à técnica anterior, este procedimento permite avaliar se o particionamento de todo o conjunto de dados é vantajoso ou não.

(SAS Institute user's guide: statistics 2003) - Este é o padrão adotado no software estatístico $SAS^{®}$ e consiste de um procedimento elaborado para avaliar a hipótese nula de que os dados vieram de uma distribuição uniforme numa hipercaixa, contra a hipótese alternativa dos dados terem sidos amostrados de uma mistura de distribuições normais multivariadas de mesma variância. A escolha do k ideal é dado pela fórmula que segue.

$$CCC = \max_k \ln\left(\frac{(1-\mathbb{E}(R^2))}{(1-R^2)}\right) \cdot \frac{\sqrt{\frac{np}{2}}}{\sqrt[5]{(\frac{1}{1000}+\mathbb{E}(R^2))^6}}$$

onde R^2 é a proporção de variância dos aglomerados e $\mathbb{E}(R^2)$ é calculada sob a hipótese de que a distribuição é a uniforme num hipercubo,; p é a estimativa da dimensionalidade da variação entre aglomerados e as constantes são resultados de simulações. O interessante deste procedimento é a introdução de uma distribuição de referência, na hipótese nula, como forma de comparar valores calculados na amostra com valores teóricos esperados nesta distribuição de referência uniforme multivariada. O valor em k que provoca o maior afastamento da função (neste caso no R^2) em relação ao seu valor teórico esperado será o momento em que a partição usada detectar os aglomerados ideais.

(Hubert and Levin 1976) - Aqui, padroniza-se o valor da dispersão dos pontos para valores distintos de k e acha-se o mínimo da função abaixo, onde se obtém o número ótimo de aglomerados.

$$\text{C-index(k)} = \min_k \frac{d_w(k) - \min(d_w)}{\max(d_w) - \min(d_w)}$$

onde d_w é a soma das distâncias dentro do aglomerado. Não é exatamente o W_k mas busca mensurar de forma semelhante a dispersão dos pontos dentro de cada aglomerado, somado para todos os aglomerados.

As técnicas acima geralmente funcionam bem, segundo (Milligan 1985), para aglomerados bem espaçados mas que podem não ser homogêneos quanto ao número de elementos. Para aglomerados sobrepostos, as técnicas de modelos de mistura (mixture models) são mais bem sucedidas em inferir o número e o tipo da família de distribuições do qual os dados foram extraídos. Muitas técnicas buscam penalizar o erro ao se fundir dois aglomerados diferentes em detrimento do erro associado ao se particionar um aglomerado em dois outros. Existem diversas outras metodologias como o cálculo do $Traço(W_k)$, $k^2|W_k|$, $n \cdot \log \frac{T_k}{W_k}$ e $Traço(W_k^{-1}B_k)$; muitas baseando-se em medir a relação das matrizes de dispersão interna, externa ou total, como função de k, de modo a visualizar um ponto ótimo ou uma queda acentuada seguida pela estabilização.

As técnicas acima, como já foi dito, possuem idéias bastante simples por detrás de

suas fórmulas. Calinski e Harabasz (1974) buscam avaliar a relação da matriz de dispersão interna com a matriz de dispersão externa em todo o conjunto de dados para definir o número ideal de aglomerados. Já Duda e Hart (1973) buscam um teste local, para cada aglomerado ou subconjunto dos dados, avaliando-os sob uma mesma medida, a dispersão interna. O Cubic Clustering Criterion (2003) busca comparar uma medida de variabilidade explicada dos dados com o seu valor teórico se o mesmo fosse extraído de uma distribuição de referência. Herbert e Levin (1976) avaliam a dispersão padronizando ao utilizar os valores máximos e mínimos encontrados para números de aglomerados diferentes. Algumas destas idéias serão incorporadas nos procedimentos que seguirão, com a mesma finalidade de detecção do número de aglomerados.

3.3 Técnicas estatísticas

O procedimento que segue é ligeiramente diferente dos anteriores na medida em que contempla um conceito diferente de agrupamento do usual que até então vem sendo adotado. Em "Estimating the Number of Clusters" (Cuevas, Febrero and Fraiman 2000), a idéia do número ótimo de aglomerados como sendo o número de "modas locais" é alcançada pela avaliação do número de componentes conectados de uma estimativa do conjunto $\{f(\cdot) > c\}$, o subespaço do suporte da distribuição de probabilidades cujo valor da função densidade seja maior que uma constante pré-fixada. Um componente conectado é, de certa forma, um conjunto de pontos próximos uns aos outros e que indicam haver na região um aglomerado.

Seja o número k de aglomerados em uma população d-variada definida como sendo o número de componentes conectados do conjunto $\{f > c\}$, f a função densidade de probabilidade em $\mathbb{R}^d$ e c uma constante arbitrária maior que zero. Então k é o número de componentes conectados de $\{f > c\}$ e o estimador deste conjunto é a união de bolas com centros em pontos apropriados da subamostra que é selecionada via um estimador não-paramétrico da densidade f. Intuitivamente esta definição de aglomerado em c busca detectar a quantia de conjuntos próximos de pontos onde existam modas locais na função de densidade que originou esses pontos.

Inicialmente, para inferir o k, será necessário um estimador $\hat{f}_n$ não-paramétrico da densidade f. E dado $c > 0$, deve-se estimar $T_n(S)$, o número de componentes conectados

do conjunto de nível $S(f;c) = \{f > c\}$ de uma amostra $x_1, \ldots, x_n$ de $f(\underline{x})$. Mas, para tal, teremos que estimar o conjunto $\{f > c\}$ pelo conjunto $\{\hat{f}_n > c\}$ e, portanto, teremos $\hat{S}_n(\hat{f}_n; c)$ que será a união de bolas de raio $\varepsilon > 0$ centradas nos pontos x_i tal que $x_i \in \{\hat{f}_n > c\}$. E por fim, o número k de aglomerados em c será dado por $T_n(\hat{S}_n)$ que é o número de c-aglomerados existentes, ou seja, a quantia de grupos de bolas (ou hiper-esferas) diferentes que se interceptam entre si. Formalmente,

$$T_n = T(\hat{S}_n),$$
$$\hat{S}_n = \bigcup_{i=1}^{q_n} B(X_i, \varepsilon_n) \qquad X_i \in \{\hat{f}_n > c\},$$

onde $B(X_i, \varepsilon_n)$ é uma bola fechada de raio $\varepsilon_n > 0$ com centro em X_i e q_n é o número aleatório de pontos no conjunto, acima do nível $\hat{f}_n > c$. No artigo orininal prova-se que o estimador $\hat{S}_n$ é fortemente consistente e a variabilidade de T_n é atestada por bootstrap, mas, agora o problema passa de definir o número ótimo de aglomerados para o de definir valores adequados de c e ε_n. No artigo original isto é feito por meio de simulação.

Diversas técnicas de avaliação do número ótimo de aglomerados por modas ou máximos locais das densidades de probabilidade conjunta também, em algum sentido criam outros problemas de avaliação subjetiva de novos parâmetros ou constantes. Mas vale ressaltar a contribuição da metodologia acima para a identificação de aglomerados em modelos de mistura de famílias de distribuições de probabilidades, contrariamente aos métodos apresentados até então que buscam detectar aglomerados isolados.

3.4 Estatística GAP

Uma recente contribuição ao tema de estimação do número de aglomerados veio, em 2000, com o artigo "Estimating the number of clusters in a data set via the GAP statistics", (Tibshirani, Walter and Hastie 2001a). Eles propuseram uma metodologia que funciona para qualquer algoritmo de análise de aglomerados e para qualquer cálculo de distância entre pontos. Este método visa formalizar a escolha do número ótimo de aglomerados no ponto onde ocorre o conhecido "fenômeno do cotovelo", que é o ponto na função W_k, definido na página 8, em função de k, em que o mesmo sofre uma queda abrupta ao passar

de $k-1$ para k, ficando $W_{k-1} \gg W_k$, e em seguida sofre quedas muito leves de k para $k+1$ em diante. Formalmente, escolhe-se k^* tal que:

$$\{(W_k - W_{k+1})/k < k^*\} \gg \{(W_k - W_{k+1}/k) \geq k^*\}.$$

onde $\gg$ denota "muito maior que". A função $W(k)$ é monótona decrescente em algoritmos hierárquicos, em que um aglomerado é quebrado/fundido a cada passo gerando mais/menos distâncias para o cálculo de $W(k)$, enquanto que as outras distâncias se mantêm. Já no caso de algoritmos de alocação iterativa,o caso de k-médias, a cada escolha de k, da partição inicial e do número máximo de iteração para alcance do mínimo $W(k)$, o algoritmo gera uma partição totalmente diferente da anterior e, mesmo buscando o $W(k)$ mínimo, ele é apenas ótimo local. A razão está em que o algoritmo é interrompido quando não há mais decréscimo no W_k pela simples mudança de um único elemento de um grupopara outro, mas poderia haver decréscimo se fossem mudados 2 ou mais elementos simultaneamente de grupos. Por isso, neste caso, esta função que é monótona decrescente, pode aumentar levemente para alguns incrementos de k em alguns algoritmos implementados em softwares estatísticos, especialmente nos casos em que o número de pontos é muito grande e denso em certas regiões do espaço. Isto não costuma atrapalhar o método proposto pois esta anomalia pode ser contornado na maioria dos casos pela escolha de uma partição inicial distinta.

Procedendo à metodologia, defina uma distância entre os elementos i e j, por exemplo a euclidiana ao quadrado:

$$d_{ij} = \sum_{p=1}^{P} (x_{ip} - x_{jp})^2$$

e sejam $C_1, \ldots, C_k$ os conjuntos de elementos em cada um dos k aglomerados, e $n_1 = |C_1|, \ldots, n_k = |C_k|$ a quantidade de elementos em cada um dos k aglomerados. Assim, define-se a soma das distâncias dentro de cada aglomerado r e a dispersão interna para k aglomerados, respectivamente, como:

$$D_r = \sum_{i,j \in C_r} d_{ij}$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r.$$

Com estas medidas, a idéia central é padronizar o $\log W_k$ comparando-o com a $E^*\{\log W_k\}$ calculada sob uma distribuição de referência, análogo ao conceito de calcular o quanto a curva da dispersão dos dados se desvia do máximo da curva esperada. A estatística GAP proposta no artigo, é definida como:

$$GAP(k) = E^*(\log W_k) - \log W_k$$

onde, após levar-se em conta a distribuição amostral da estatística GAP e da distribuição de referência, será escolhido o k, tal que o GAP seja máximo. Segue o algoritmo completo para a determinação do número ótimo de aglomerados em uma base de dados, fixado um algoritmo de aglomeração e uma métrica.

1. Aglomerar a matriz $n \times p$ dos dados para $k = 1, \ldots, n$, calculando W_k para cada k.

2. Gerar, por Monte Carlo, B vetores de amostras de referência $x_1^*, \ldots, x_n^*$, calculando W_{kb}^* para cada b e k. Calcular:

$$E_n^*\{\log W_k\} = \frac{1}{B} \sum_{i=1}^{B} \log(W_k^*)_i$$

$$\sigma(k) = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} \left(\log W_k^{*i} - E_n^*\{\log W_k\}\right)^2}$$

$$s_k := \sqrt{\frac{B+1}{B}} \sigma(k)$$

3. Calcular a estatística GAP(k) e escolher o k^* ótimo tal que :

$$k^* = \arg\min_k \{k | GAP_n(k) \geq GAP_n(k+1) - s_{k+1}\}$$

Algumas abordagens e escolhas usadas na metodologia do GAP merecem comentários. A mais relevante é a distribuição de referência uniforme com suporte no intervalo de variação de cada uma das p variáveis dos dados. Ocorre que, na ausência de informações acerca de uma distribuição originária dos dados, a uniforme é a mais conservadora pois propicia aglomerados mais esparsos possíveis. Em (Tibshirani et al. 2001a), é provado que o ínfimo da razão entre o erro quadrático médio (EQM) com k aglomerados sobre o EQM com 1 grupo, tomado na classe de distribuições com um componente é alcançado pela distribuição uniforme do intervalo[0,1]. Isto significa que dentre todas as distribuições unimodais no espaço unidimensional, a uniforme é a que gera aglomerados mais esparsos possível pelo método GAP. Outro teorema provado neste mesmo artigo indica que não existe uma distribuição que satisfaça a propriedade de ter mínima razão do EQM de k sobre 1 aglomerado, em dimensão maior que 1. Então, conclui-se que a geometria da distribuição de referência afeta o resultado do cálculo do número de aglomerados pelo método proposto no artigo. A abordagem mais simples usado por (Tibshirani et al. 2001a), foi a de gerar variáveis aleatórias uniformes independentes em cada uma das p dimensões dos dados com suporte dado pela janela de amplitude equivalente aos valores observados em cada dimensão. Observe que a escolha do suporte utilizado para a distribuição de referência, no caso de uniformes independentes, não altera o cálculo do número de aglomerados pelo método GAP. Este resultado será mostrado em detalhes na Seção 2.6.

Outro ponto importante na metodologia usada é que os autores atestam que não usaram um fator multiplicador para o desvio padrão s_k (ou fator igual a 1), mas que um maior conservadorismo pode ser alcançado se o valor mais adequado for avaliado por meio de simulações de amostras da base original de dados. Além do mais, os autores afirmam que descobriram empiricamente que o fator 1 funciona. No entanto, verificamos, também empiricamente, que quanto menor for este fator, maior (ou igual) é a estimativa do número ótimo de aglomerados. Isto será levado em consideração nos exemplos a seguir.

Entretanto, a regra do argumento mínimo em k para a escolha ótima é a mais discutível delas pois favorece o menor número de aglomerados. Isto é, a desigualdade acima não leva em conta o grau (ou o módulo do valor) com que esta desigualdade ocorre entre $k's$ pequenos e próximos entre si. Isto pode ser questionado pois o custo ao se optar por k ao invés de $k+1$ não é levado em conta, o que é crítico em várias aplicações de análise de aglomerados, è.g., naquelas em que o número de aglomerados pode ser relativamente alto.

Em suma, o procedimento pela estatística GAP alcança muito bem o seu propósito de detectar quantos aglomerados isolados existem numa base de dados e merece atenção na escolha da distribuição de referência. Os seus propositores evidenciaram esta afirmação ao comparar este com diversos outros métodos definidos e compilados em (Milligan 1985). Mas o GAP se torna discutível ao escolher o menor número de aglomerados que passa na condição estipulada, sem levar em conta o grau com que alguns $k's$ passam por esta condição. Ou seja, o algoritmo não leva em conta o comportamento da curva GAP de k^* (o valor escolhido) até $k = n$ pois ele estuda apenas o primeiro argumento que minimiza a diferença entre GAP(k) e GAP(k+1).

3.5 Inferência de K pelo GAP em pequenas amostras

A seguir serão apresentados 3 exemplos criados com configurações construídas com 49 pontos no quadrado de lado unitário, a saber, um exemplo com mistura de 5 normais bivariadas, outro exemplo com 8 normais bivariadas desigualmente espaçadas (pelo vetor de médias) e um terceiro caso com 8 pentágonos deterministicamente distribuídos. O primeiro conjunto de pontos foi gerado aleatoriamente e é formada com o objetivo de se criar 5 grupos não sobrepostos com mesma quantidade de elementos pela mistura de 5 normais bivariadas com médias centradas e igualmente espaçadas na diagonal do quadrado, desvios padrões de 0,05 e variáveis não correlacionadas.

O cálculo das funções $GAP(k)$, o intervalo com barras de mais e menos um desvio padrão e a função $Lacuna(k)$, definido na sequência, são traçados no gráfico abaixo para a configuração de mistura de 5 normais. Define-se $Lacuna(k)$ como sendo a diferença entre o GAP em k e $k+1$ somado o desvio padrão do GAP em $k+1$, uma forma simples de reescrever o algoritmo GAP. Cabe a observação de que o desvio padrão do GAP tende a aumentar com k e que, pela influência desse, não é verdade que o número ótimo de aglomerados será o valor de k que maximiza a função. A proposta original de (Tibshirani et al. 2001a), conforme explanado anteriormente é de que o valor procurado seja o menor

valor de k tal que o GAP neste ponto seja maior do que no próximo ponto, descontado o desvio padrão do GAP no próximo ponto. Deste modo, o número de aglomerados será o primeiro ponto acima do valor zero no eixo das ordenadas da função Lacuna(k), no caso do exemplo abaixo, $k = 5$.

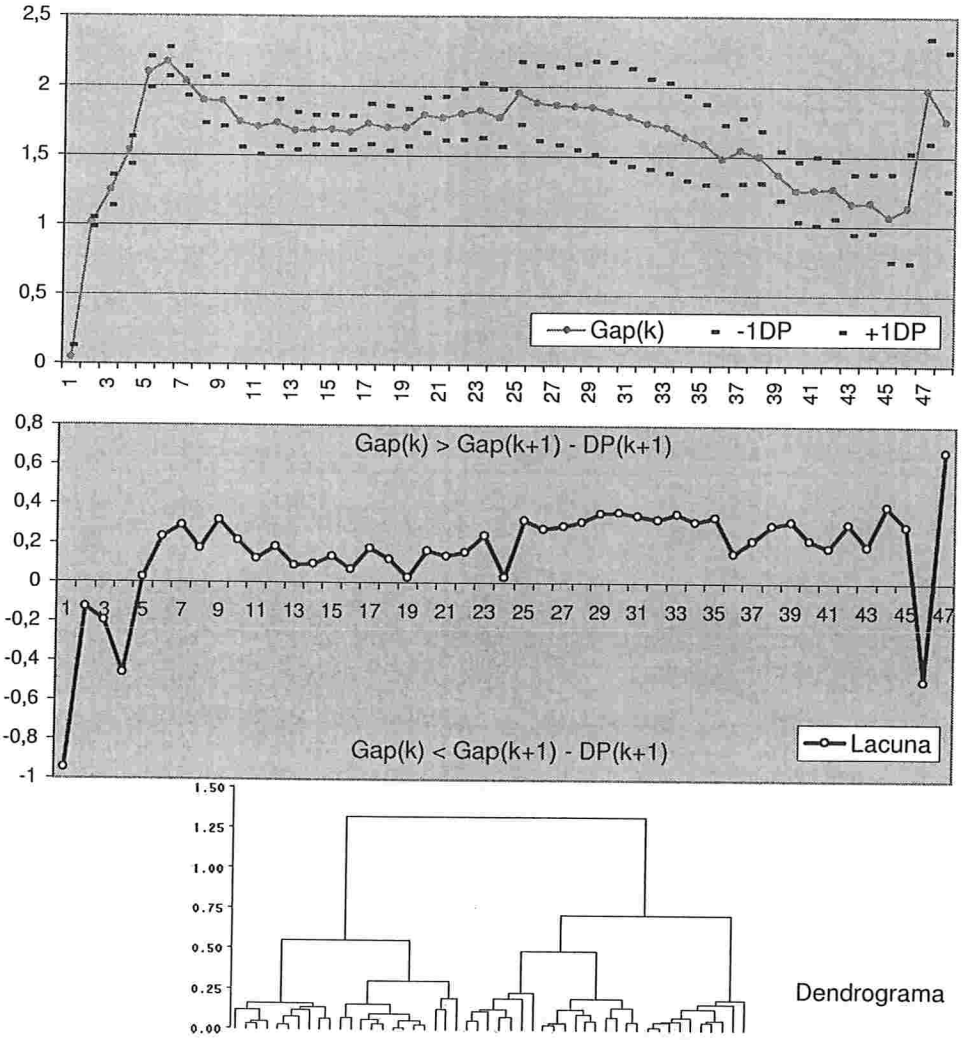

Figura 3.0 Função GAP(k) e Lacuna(k), onde o eixo da abscissa é a iteração; em baixo o dendrograma do ex. 1

A análise de aglomerados em todos os exemplos foi feita em SAS, escolhendo-se a distância euclidiana e o método hierárquico de ligação média. A título de ilustração também serão apresentados os dendrogramas das junções feitas a cada passo como forma

de avaliar a distância de corte sugerida pelo algoritmo. Como regra empírica, é sugerido que o número de aglomerados seja aquele que resulte do corte da maior diferença de junções subsequêntes no dendrograma. Este segue ao gráfico da Lacuna.

O próximo exemplo contém 49 pontos igualmente distribuídos em 8 Normais com vetores de médias $(0,1;0,1)$, $(0,3;0,1)$, $(0,1;0,3)$, $(0,3;0,3)$, $(0,7;0,7)$, $(0,7;0,9)$, $(0,9;0,7)$, $(0,9;0,9)$, com desvio padrão de $0,025$ e cujas variáveis não são correlacionadas. Visualmente, tem-se ou 2 grupos distantes ou 8 grupos, sendo 4 próximos entre si.

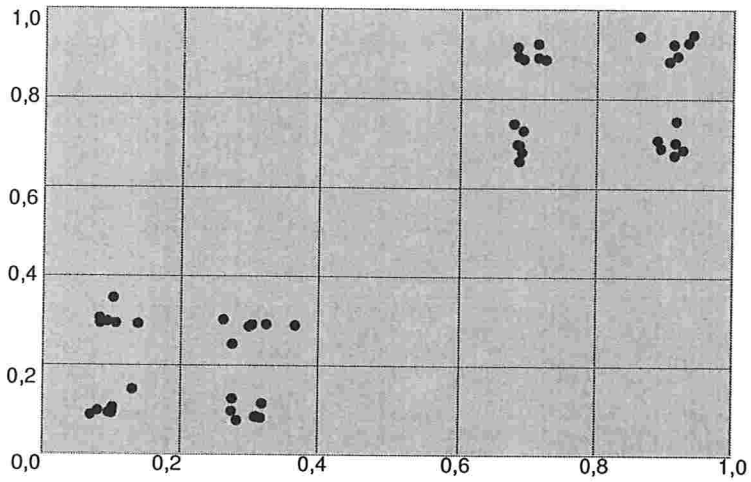

Figura 4.0 Gráfico de dispersão dos pontos no ex. 2

Esta construção visa avaliar empiricamente como a curva GAP se comportará em situação ambígua. E pode-se ver pelas funções no gráfico abaixo que este fenômeno é captado tanto pela função GAP na forma de um forte crescimento seguido por uma queda, quanto pela Lacuna, que passa de um valor negativo para um valor positivo nos $k's = 2$ e 8.

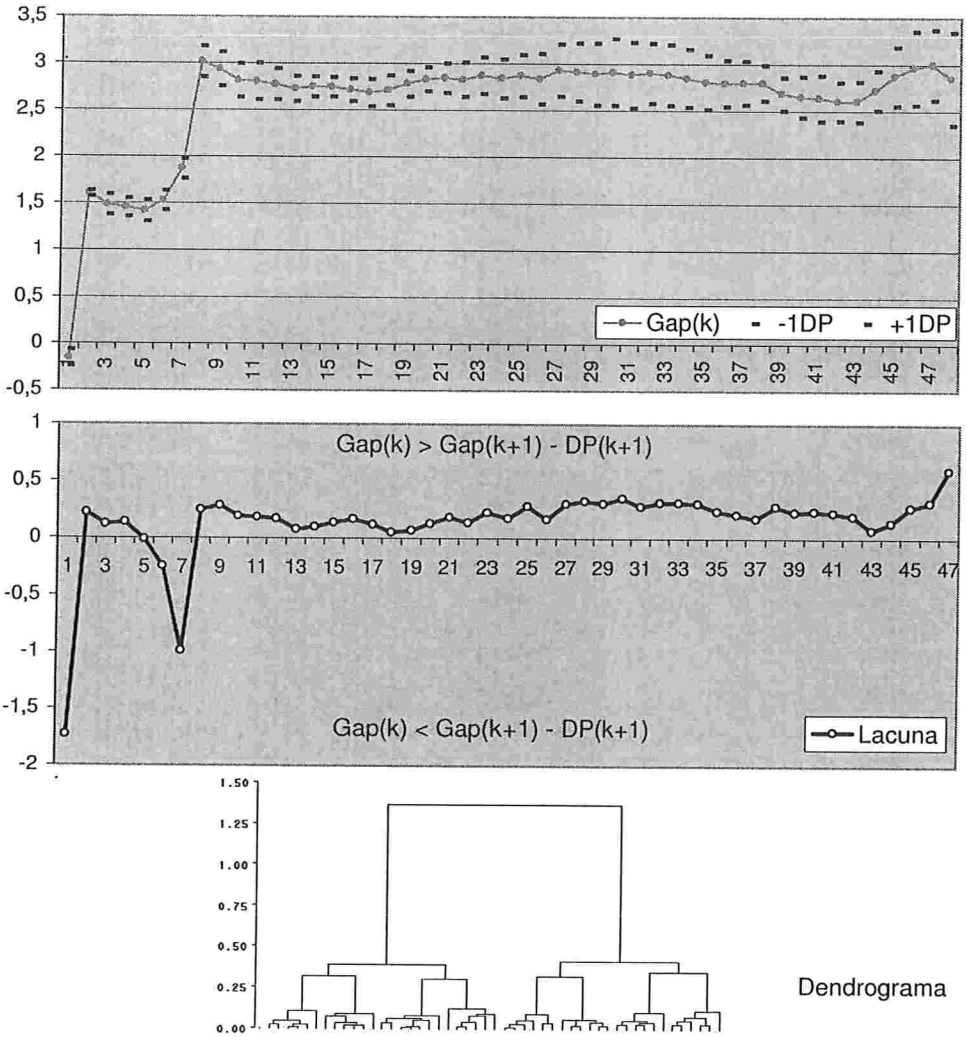

Figura 4.1 Função GAP(k) e Lacuna(k), onde o eixo da abscissa é a iteração; em baixo o dendrograma do ex. 2

O dendrograma, por sua vez, não distingue tão bem a ambiguidade potencial já que o salto de dissimilaridade é muito mais intenso de 2 para 1 aglomerado, podendo levar à escolha de $k = 2$. Mas mesmo nesta situação, o método do GAP sugere dois grupos como sendo a partição mais adequada, em detrimento de 8 que é o valor máximo na função GAP, descartado por força do efeito do argumento mínimo utilizado na sua formulação.

No terceiro exemplo, foi concebida uma outra configuração com o objetivo de potencializar o efeito de ambiguidade anterior pela divisão dos 49 pontos em 8 grupos de 6 elementos, cada um com a estrutura de pontos nos vértices do pentágono com um pon-

to no centro (5 + 1). Os pentágonos são dispostos no reticulado bidimensional de modo a ter um no extremo superior esquerdo, um no extremo inferior direito e os outros seis igualmente empaçados em pares no centro. Assim, pode-se visualizar os oito grupos de mesmo tamanho, pode-se visualizar 3 grupos (um grande no centro e dois pequenos nas extremidades) mas também pode-se visualizar 5 grupos (os dois das extremidades e os 3 pares de pentágonos no centro).

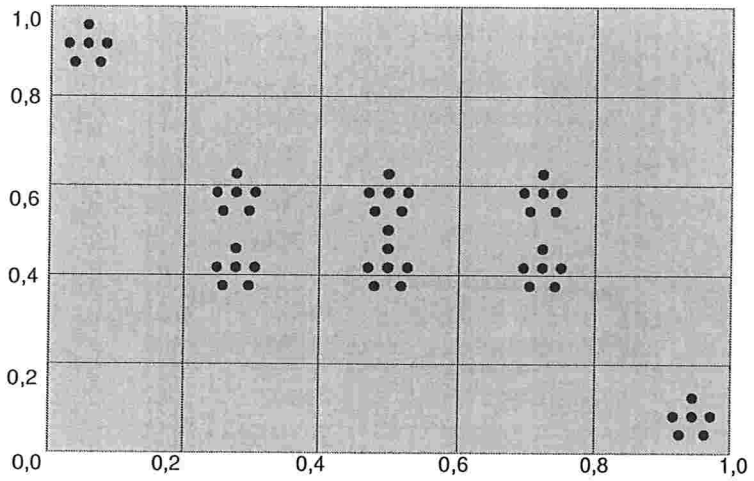

Figura 5.0 Gráfico de dispersão dos pontos no ex. 3

A análise das funções GAP e Lacuna evidencia o ponto 8 onde ocorre o máximo da função GAP mas o algoritmo indica $k = 5$ aglomerados porque, ao levar em conta um desvio padrão e o crescimento da função até este ponto, ele acaba satisfazendo à desigualdade requerida e, como podemos ver na função Lacuna, é o menor argumento a fazê-lo. Com relação à escolha de 3 grupos, não há evidências de que esta poderia ser uma opção razoável, pelo menos do ponto de vista da função Lacuna ela poderia ser tão razoável quanto 2 ou 4 grupos.

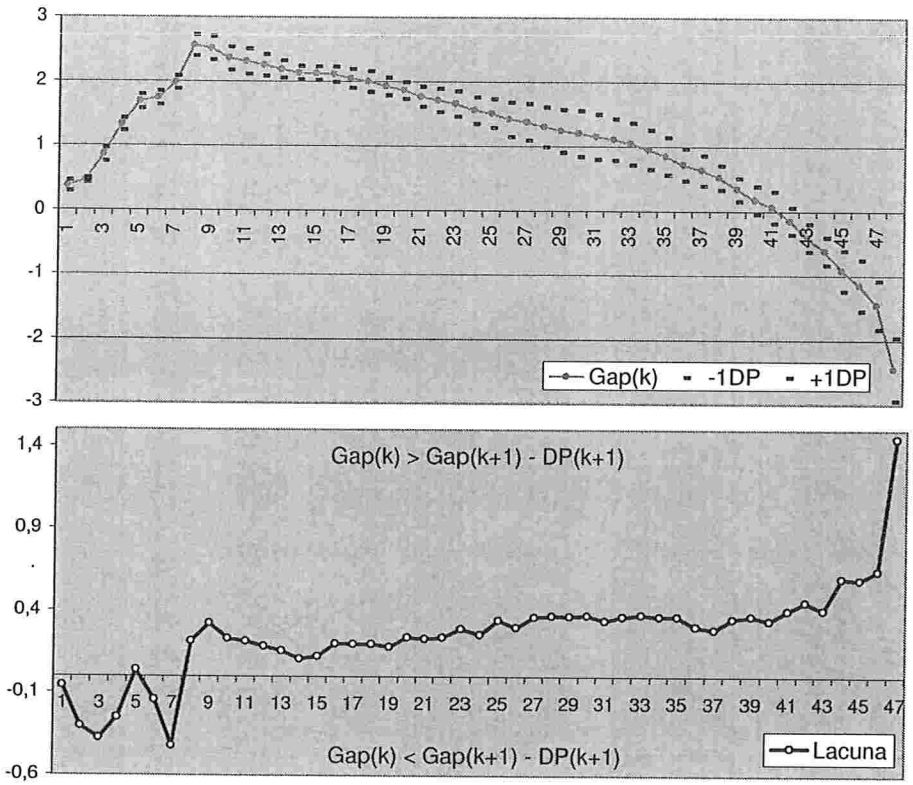

Figura 5.1 Função GAP(k) e Lacuna(k), onde o eixo da abscissa é a iteração, do ex. 3

Em suma, os exemplos prospectivos acima nos evidenciam o comportamento das curvas GAP e Lacuna, principalmente a última que particiona o conjunto dos $k's$ possíveis em valores acima do zero e valores abaixo do zero. E como a metodologia do GAP escolhe o primeiro valor acima do eixo, pode-se especular que todos os valores acima do eixo que seguem um valor abaixo do eixo são potenciais candidatos a serem $k's$ adequados. Isto se deve porque a passagem de um valor negativo na função Lacuna para um valor positivo é o efeito de crescimento seguido de queda ou estabilização na função GAP, um máximo local. E, por sua vez, estes máximos locais na função GAP são evidências de crescimento da distância entre o $\log W_k$ dos dados e $E^*\{\log W_k\}$ da distribuição de referência, justamente o "cotovelo" heurístico que procuramos. Tendo isto em consideração, pode-se especular que uma pequena alteração na formulação da escolha do número de aglomerados, usando a função Lacuna, pode gerar resultados diferentes e eventualmente mais adequados, como por exemplo 8 no exemplo 2 e 3 ou 8 no exemplo 3.

3.6 Constância dos Parâmetros de Escala para a Distribuição de Referência Uniforme

Iremos mostrar que o critério de escolha do número ótimo de aglomerados pela estatística GAP, k_{GAP} é invariante ao suporte da distribuição de referência, se esta for uniforme, e que, neste caso, a função GAP é multiplicada por uma constante.

Seja a distribuição de referência gerada por variáveis aleatórias independentes com densidades:

$$x_p \sim Uniforme(0, a), \qquad a > 0$$

para cada dimensão $p = 1, \ldots, P$ que é o número de variáveis presente na base de dados e utilizadas na análise de aglomerados. Assim, teremos, calculando a esperança da dispersão interna pela distância euclidiana ao quadrado:

$$W_k := \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,j \in C_r} \sum_{p=1}^{P} (x_{ip} - x_{jp})^2$$

$$E^*\{W_k\} = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,j \in C_r} \sum_{p=1}^{P} E\{(x_{ip} - x_{jp})^2\}$$

onde,

$$E\{(x_{ip} - x_{jp})^2\} = 2(Var\{x_i\} + E\{x_i\}^2) - 2E(x_i, x_j) = 2Var\{x_i\}$$

pois x_i e x_j são independentes e identicamente distribuídos. Deste modo,

$$\begin{aligned} E^*\{W_k\} &= \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,j \in C_r} \sum_{p=1}^{P} 2Var\{x_i\} \\ &= \sum_{r=1}^{k} \frac{1}{2n_r} \binom{n_r}{2} 2PVar\{x_i\} \\ &= \sum_{r=1}^{k} \frac{P(n_r - 1)Var\{x_i\}}{2} \end{aligned}$$

$$= \frac{P(n-k)Var\{x_i\}}{2}$$

Caso multipliquemos o suporte da distribuição uniforme em uma constante $C > 0$, teremos

$$U_p \sim Uniforme(0, Ca) \qquad Var(U) = \frac{(Ca-0)^2}{12}$$

$$E^*\{W_k^C\} = \frac{P(n-k)Var\{u_i\}}{2}$$

$$= \frac{P(n-k)C^2Var\{x_i\}}{2}$$

$$\Rightarrow E^*\{W_k^C\} = C^2E^*\{W_k\}$$

Pela desigualdade de Jensen,

$$E^*\{\log W_k\} \leq \log E^*\{W_k\}$$

podemos concluir que, dado:

$$\log E^*\{W_k^C\} = \log(C^2E^*\{W_k\}),$$

$$E^*\{\log W_k^C\} + \delta_k = 2\log C + E^*\{\log W_k\} + \delta_k$$

$$\therefore E^*\{\log W_k^C\} = 2\log C + E^*\{\log W_k\}$$

A passagem anterior é justificada pelo fato de que a diferença do log da esperança de W_k e da esperança do log de W_k pode ser transformada em uma igualdade pela inserção de uma constante δ que depende de k mas que não depende de C. Portanto, a aproximação feita do lado esquerdo da igualdade acima é cancelada com a aproximação feita do lado direito. Calculando o GAP,

$$GAP(k, C) = 2\log C + E^*(\log W_k) - \log W_k$$

$$= C' + GAP(k)$$

Calculando a função estimadora de k, denominada Lacuna,

$$k - \text{ótimo} := \arg\min_k \{GAP(k, C) - GAP(k+1, C) + s_{k+1}\}$$

$$Lacuna(k, C) := GAP(k, C) - GAP(k+1, C) + s_{k+1} =$$

$$C' + GAP(k) - (C' + GAP(k+1)) + \sqrt{\frac{B+1}{B}\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\Big(\log(W_{k+1})_i - E^*(\log W_{k+1})\Big)^2}} =$$

$$GAP(k) - GAP(k+1) + \sqrt{\frac{B+1}{B}\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\Big(\log C + \log(W_{k+1})_i - \log C - E^*(\log W_{k+1})\Big)^2}} =$$

$$Lacuna(k)$$

Portanto, com o auxílio da função Lacuna, provamos que o GAP(k) é aumentado por uma constante e o k-ótimo não varia com o aumento do tamanho do suporte da distribuição de referência uniforme se esta for independente para cada dimensão p. Ademais, é evidente que um deslocamento em qualquer dimensão do suporte da distribuição de referência terá o mesmo efeito. Para provar isto, basta acrescentarmos constantes C_1 e C_2 aos parâmetros a e b da Uniforme e proceder aos cálculos de modo análogo ao feito acima.

Capítulo 4

Cálculo do Número de Aglomerados pela Estatística GAP em Sistemas Estocásticos

A seção anterior versou sobre algumas das técnicas mais proemientes e utilizadas correntemente para o cálculo do número ideal de aglomerados em uma base de n elementos por p variáveis. A exposição introduziu a técnica do GAP como um modelo da escolha do número de aglomerados pela comparação de uma função de dispersão calculada sobre a amostra e comparada com a esperança da mesma, avaliada em uma distribuição de referência. Além disso, mostrou-se que o suporte da distribuição de referência, no caso dela ser uniforme em cada uma das p dimensões, independente e identicamente distribuída, não afeta o k dado pela estatística GAP.

Na seqüência, mostraremos como se comporta a estatística GAP(k) e a escolha do número de aglomerados por esta técnica avaliados sob uma base de dados que possui uma evolução estocástica no tempo. Ou seja, a cada instante de tempo a base de dados muda pela alteração aleatória de um dos seus elementos de modo que a configuração futura dependa da configuração presente, definindo um sistema markoviano de ordem 1. E a cada iteração será calculado o GAP(k) e o número k de clusters até que ocorra uma estabilização destes valores pela entrada no equilíbrio dinâmico do processo markoviano.

4.1 Modelo estocástico adotado

O modelo adotado consiste em um conjunto de n elementos em p dimensões igualmente espaçados em um hipercubo de tamanho equivalente ao suporte da distribuição de referência adotada. E a cada nova iteração, escolhe-se um destes n pontos com igual probabilidade e aloca-o para a fronteira mais longe (o ponto central do lado oposto) de um dos $2p$ hipercubos adjacentes ao que contém os pontos iniciais. A cada iteração, existe uma região de destino de modo que estão ordenados os $2p$ possíveis destinos. Caso seja escolhido um dos pontos que já esteja alocado a um destino, este é novamente realocado ao destino da vez. Para que os pontos não se sobreponham nos destinos, gerando distâncias nulas em cálculos posteriores, será adicionado um ruído branco: a cada ponto alocado é adicionada uma variável aleatória de distribuição normal centrada no destino com variância σ^2.

Como ilustração, desenvolvemos uma evolução simplificada do modelo proposto com $n = 16$ pontos, $p = 2$ dimensões, $\sigma^2 = 0.001$. Neste caso, temos quatro pontos de destino de coordenadas (0,5;-1), (2;0,5), (0,5;2) e (-1;0,5). A ordem de alocação não afeta o estudo e pode ser definida com o sentido horário partindo de qualquer ponto de destino, o Sul, por exemplo, o destino (0,5;-1) como foi feito na figura 6.0 .

O tempo médio para se chegar ao equilíbrio dinâmico, onde todos os n pontos foram escolhidos pelo menos uma vez e, portanto alocados, é $n \ln n$, dado pela soma das esperanças de variáveis aleatórias com distribuição geométrica de parâmetros diferentes. Podemos calcular o tempo médio de escolha de todos os pontos da seguinte forma:

Seja $\mathbb{P}\{X\}$ a probabilidade de se escolher todos os ponto ao menos uma vez e X_i, o instante da escolha do ponto i:

$$\begin{aligned}
\mathbb{E}\{X\} &= \sum_{i=1}^{n} \mathbb{E}\{X_i\}, \qquad X_i \sim \textit{Geométrica}(p_i) \\
&= \sum_{i=1}^{n} \frac{1}{p_i} = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \dots \\
&= n\Big(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \dots + 1\Big) \\
&\simeq n \int_1^n \frac{1}{x} dx = n \ln n
\end{aligned}$$

A título de ilustração do comportamento evolutivo das partículas neste processo, será necessário definir uma configuração inicial, por exemplo, a que todos os pontos estão igualmente espaçados em um reticulado de duas dimensões e, associando a estes, dois vetores, um de variáveis aleatórias com distribuição discreta entre 1 e n, e outro vetor com variáveis aleatórias com distribuição normal para provocar o ruído branco. Desta forma, conseguimos determinar todos os estados, ou configurações, dos pontos a cada tempo e, portanto, fazer a análise de aglomerados calculando as estatísticas necessárias a escolha no número ótimo de grupos em cada tempo. Segue a evolução do conjunto de pontos em uma simulação realizada, para iterações desigualmente espaçadas no tempo (número da iteração no canto superior esquerdo):

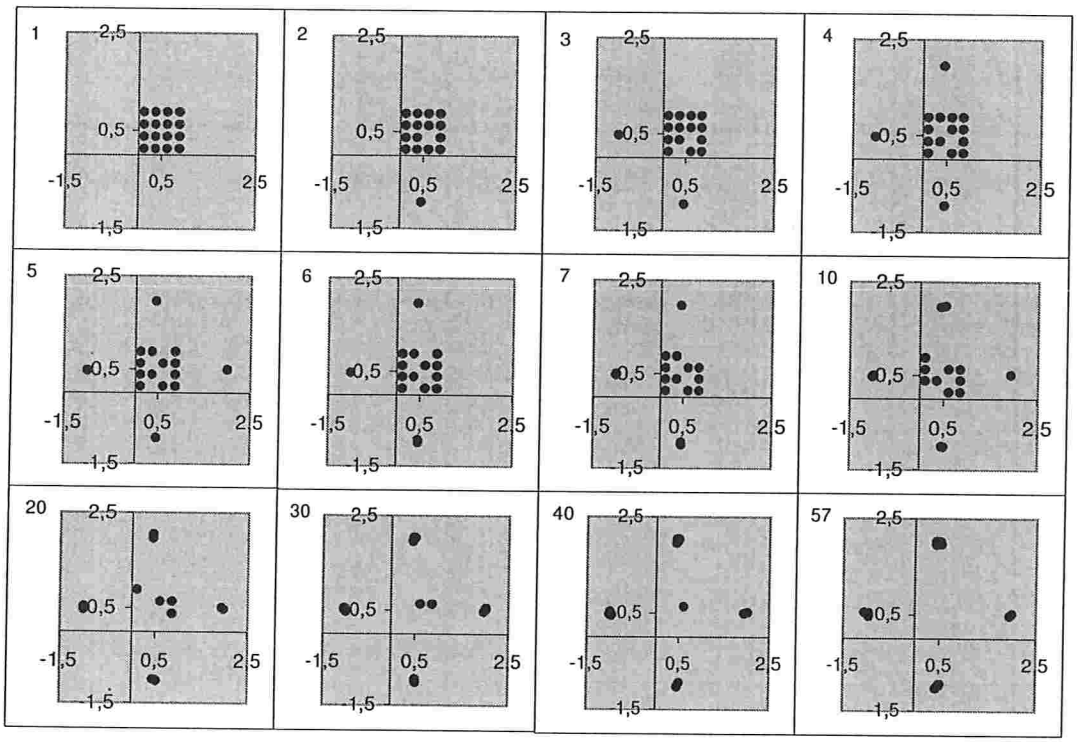

Figura 6.0 Configuração do modelo proposto até o equilíbrio dado após 57 iterações (nº da iteração no canto superior esquerdo)

Esta evolução foi idealizada com o propósito de estimar, não apenas um número de aglomerados em uma base estática, mas para estimar o número de aglomerados para o

qual o algoritmo estimador do GAP tenderá a indicar, ao ser calculado a cada alteração aleatória de uma base dinâmica. Os pontos vão sendo depositados em lugares previamente conhecidos para que se tenha controle do número final de aglomerados que será o número de destinos. Em suma, busca-se o k pela avaliação dos k_t dados pela estatística GAP.

4.2 O Cálculo do Número de Aglomerados no tempo

Procederemos ao cálculo do número ótimo de clusters $k, k = 1, ..., 16$, em cada iteração do modelo. Este será avaliado mediante a maior diferença entre a esperança do logaritmo de W_k usando a distribuição de referência e o log de W_k calculado sobre a base de dados. Fazendo-se isto e adicionando-se um multiplicador ao desvio padrão da esperança, estimado por replicações de Monte Carlo, tem-se um valor de k para cada passo até que se chegue ao valor final de 4 clusters, quando o processo se estabiliza e todos os pontos estão em um dos 4 destinos finais, a menos do ruído branco.

Conforme dito anteriormente, a escolha do k dada pelo menor argumento em que GAP(k) decresce, pode ser expresso como o argumento k que minimiza a função Lacuna, definidos por:

$$k - \text{ótimo} := \arg\min_k \{GAP(k) - GAP(k+1) + \sigma_{k+1} > 0\}$$

onde σ_{k+1} = desvio padrão de $\log W_k$. O artigo de (Tibshirani et al. 2001a), propõe um multiplicador conservador para este desvio padrão equivalente a 1, mas outros valores, maiores ou menores, devem ser propostos mediante avaliação do número de replicações de Monte Carlo feitas para o cálculo da esperança do logaritmo de W_k e mediante simulações do cálculo de k em uma base de dados específica. Por hora não será alterado este multiplicador mas mais adiante evidenciaremos como o número ótimo de aglomerados é sensível a estes valores para algumas bases de dados. Observe que um multiplicador maior, ou não alterará, ou poderá diminuir o número de aglomerados escolhidos, nunca aumentá-lo. Esta constatação é trivial uma vez que a função Lacuna desloca-se para cima em quantidades que variam com o k (σ_{k+1} tende a crescer em k) e este deslocamento só pode resultar em um argumento mínimo menor ou igual ao anterior.

Calculando a função Lacuna(k) para todo k e para toda iteração t do modelo mar-

koviano, podemos ter uma boa idéia do comportamento desta função, que está definida apenas em $\mathbb{N}$, mas que, por questão de conveniência visual, será graficada como uma função contínua, interpolada entre os pontos de descontinuidade. Segue um subconjunto de funções Lacuna para todos os $k's$ de alguns intervalos de tempo de uma simulação do modelo. Para cada t, existem duas curvas, a curva superior usa um multiplicador de 1 para o s_{K+1}, o desvio padrão do GAP no ponto $K+1$, enquanto que a outra curva não leva em conta o desvio padrão na escolha do número de aglomerados, ficando sempre por baixo da anterior.

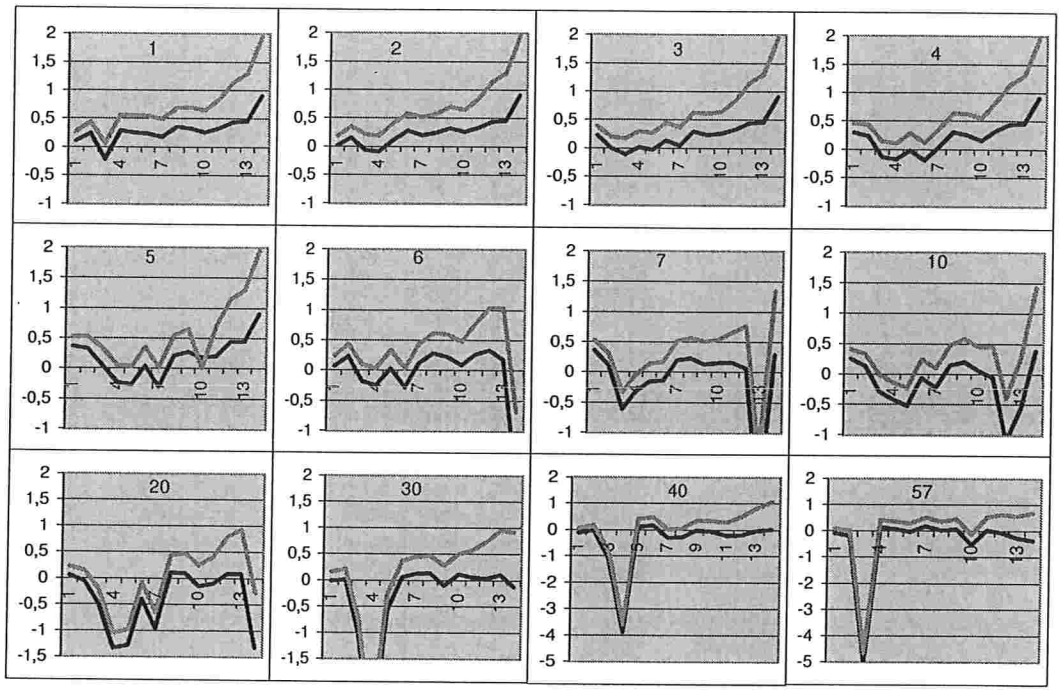

Figura 7.0 Funções Lacuna(k)para a realização anterior, onde o eixo da abscissa é o nº de aglomerados (nº da iteração no centro superior)

Ao avaliarmos a evolução da estimativa do número de aglomerados no decorrer do tempo, até o valor final de 4, podemos constatar que o método do GAP tende a subestimar o número de grupos na base de dados e que o método estima como existindo apenas um grupo nas primeiras iterações, mais especificamente, nas 27 iterações do primeiro exemplo simulado. Isto ocorre ao usarmos o conjunto de curvas mais baixas do gráfico acima, sem

o desvio padrão. Obviamente, se fosse usado o multiplicador 1 para o DP, este valor de casos em que $k = 1$ só poderia crescer, o que é inaceitável do ponto de vista prático em que claramente nota-se que o método não está detectando a evolução e o crescimento do número de grupos rápido o suficiente.

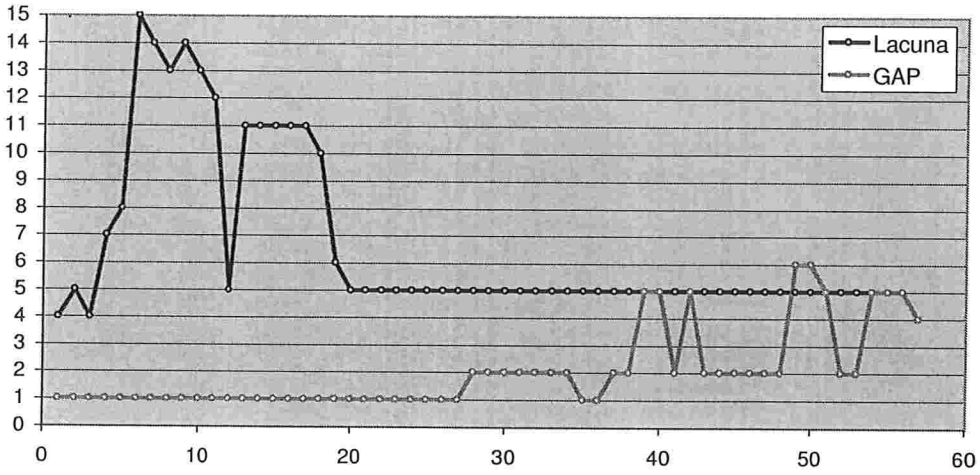

Figura 8.0 Evolução de K_{GAP} no tempo para a realização anterior, onde o eixo da abscissa é a iteração(tempo)

A conjectura deste fenômeno perturbador é de que ele advém do uso do argumento mínimo como regra de decisão do k baseado na seqüência GAP(1) a GAP(n). Desta forma, em diversos casos estudados, a determinação do número de grupos não é decidida tomando-se como base toda a curva, e sim apenas os valores de GAP(k) até o menor k tal que $GAP(k) - GAP(k+1) + \sigma_{k+1} > 0$. Acredita-se que o resto da curva tenha informações importantes que não devem ser eliminadas para a escolha do k, principalmente no caso de uma base com evolução estocástica, com forte dependência do estado anterior, como é o caso do exemplo usado como laboratório.

A fim de sanar a deficiência considerada acima, é necessário propor uma variante na metodologia da escolha do k pelo GAP de modo a avaliar o número de agrupamentos pelo passado recente, por todas as possibilidades do número de grupos e pela busca da convergência de um k para o processo. Enfim, se faz necessário o estudo da função no tempo de maneira global.

4.3 Proposta alternativa no contexto de Aglomerados Estocásticos

Na seção anterior prospectamos que a metodologia da escolha do número de aglomerados pela estatística GAP apresentou falhas em alguns aspectos para a determinação de grupos não sobrepostos em um modelo estocástico com evolução markoviana (simples) no tempo. Um ponto a ser considerado é que o k dado pelo GAP tende a subestimar a quantidade de agrupamentos a cada iteração, isto foi constatado empiricamente e depois traduzido como sendo resultado do uso da função "argumento mínimo".

Propomos aqui algumas alterações no algoritmo GAP visando conseguir um estimador com propriedades melhores para o k_t em cada instante t do processo estocástico em estudo. Deste modo, procedemos a uma variante do algoritmo usando a função Lacuna, advinda do GAP.

O procedimento alternativo consiste essencialmente em calcular o GAP(k) para todo k e todo instante t no modelo com evolução estocástica, sendo a escolha do k_t feita da seguinte forma:

$$\begin{aligned} k_t &= \arg\max_k\{-Lacuna_t(k)\} + 1 \\ &= \arg\max_k\{GAP_t(k+1) - GAP_t(k) - \sigma_{k+1,t}\} + 1 \\ \text{onde } \; Lacuna_t(k) &= GAP_t(k) - GAP_t(k+1) + \sigma_{k+1,t} \end{aligned}$$

Com esta alteração, procura-se (1): evitar que se subavalie a quantidade de agrupamentos pela substituição da regra de escolha do argumento que primeiro satisfaça uma condição para a escolha do argumento que maximize uma função de interesse. E esta função deve trazer consigo informações do grau com que W_k se estabiliza, resultando no estudo do comportamento do "cotovelo" almejado desde o início. Outra propriedade desejada (2): o estimador deve convergir o mais rápido possível para o seu valor verdadeiro no equilíbrio final. Não só buscamos um bom estimador para k_t a cada iteração mas também um estimador k geral para o processo sem que sejam feitas todas as análise de aglomerados até que o equilíbrio ocorra, algo próximo de $n^2 \ln n$, para n grande. Isto é moroso computacionalmente e por isso, uma convergência rápida é importante. Além disso, é sempre

interessante termos (3): um procedimento que gere $k_t's$ com baixa variabilidade ao redor do verdadeiro valor de convergência.

Visando encontrar evidências para nossa variante, já que a demonstração analítica destas propriedades é de difícil desenvolvimento por não haver forma analítica das distribuições de W_k, do GAP e da Lacuna, foram utilizadas replicações simuladas de variáveis aleatórias independentes na evolução do modelo estocástico e foi estimado k_t a cada replicação de Monte Carlo para todo o conjunto de k possíveis. Foram realizadas 100 replicações do modelo exposto anteriormente com 16 pontos. K_t^b é o valor estimado pela regra acima do número de aglomerados k para a iteração t da replicação b de Monte Carlo. Temos:

$$K_t = \frac{\sum_{i=1}^{B} K_t^i}{B} \mathbf{1}_{\{K_t^b\}}(1, \ldots, N).$$

O resultado dos cálculos na simulação está apresentado na figura a seguir, indicando a média e a mediana para estimativas de k no tempo calculadas pelo método GAP e pelo método alternativo, aqui proposto, denominado Lacuna:

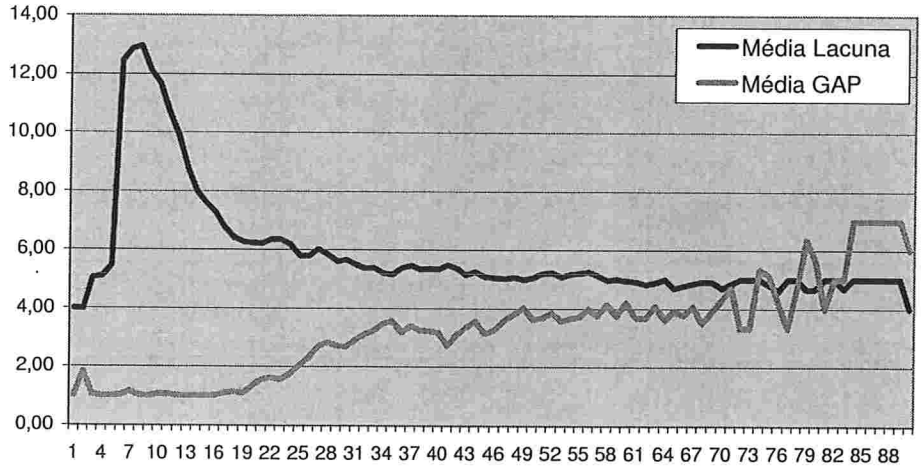

Figura 9.0 Médias de K_T nas 100 replicações pelo método GAP e alternativo, onde o eixo da abscissa é a iteração(tempo)

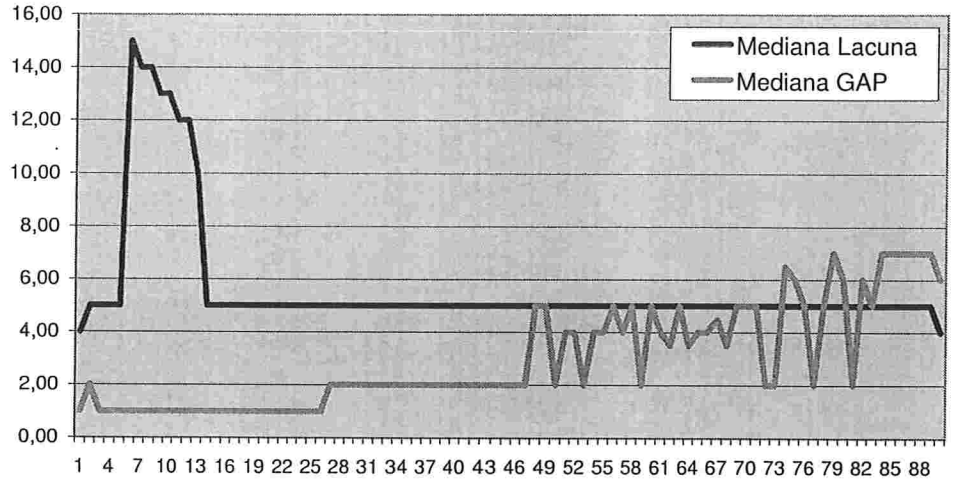

Figura 10.0 Medianas de K_T nas 100 replicações pelo método GAP e alternativo, onde o eixo da abscissa é a iteração(tempo)

4.4 Desenvolvendo uma regra de parada para K

Pelo resultado das simulações feitas, constata-se que o método GAP de estimação do número de aglomerados parece subestimar, em média e em mediana, o número ideal de grupos nos conjuntos de dados dinâmicos utilizados. Este viés é provocado em parte pelo uso da função "arg min" na metodologia original. Além disto, notamos que a variância do estimador está crescendo com o tempo (vide figura 11) e que a esperada convergência do k_t estimado para um valor final único não ocorre. Em suma, podemos dizer pelos resultados apresentados nas simulações que o método GAP aparentemente não é eficiente na determinação do número de grupos quando os elementos possuem esta evolução estocástica. Atenta-se para o fato de que o estado de equilíbrio final do processo, apesar de ser 4 grupos, não é o valor buscado $k = 5$ que é o número de grupos imediatamente anterior ao fim do processo e que permanece no seu decorrer pois a formulação geométrica foi definida para ter 4 aglomerados de destino e mais 1 grande aglomerado que deveria ser consumido com o tempo, em média levando $n \ln n$ iterações.

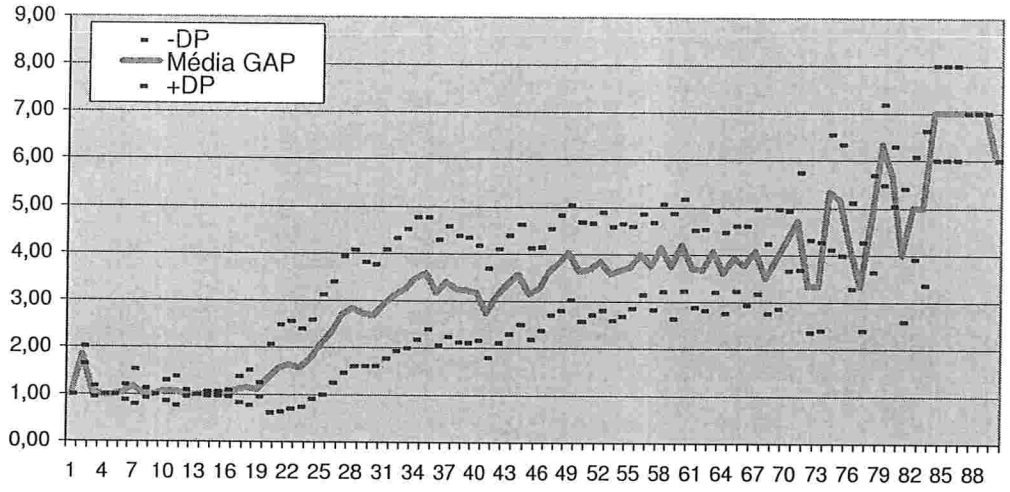

Figura 11.0 Médias e bandas de um desvio padrão de K_T pelo GAP a cada iteração

Contrariamente ao GAP, o método alternativo, denominado método *Lacuna*, superestima o número de aglomerados em média e em mediana no início mas a partir de um instante (aleatório) este fenômeno decai rapidamente. Além disso, o k_t estimado por este método converge rapidamente para o valor final ideal, tendo a propriedade de ter uma variância que decresce com o tempo (vide figura 12).

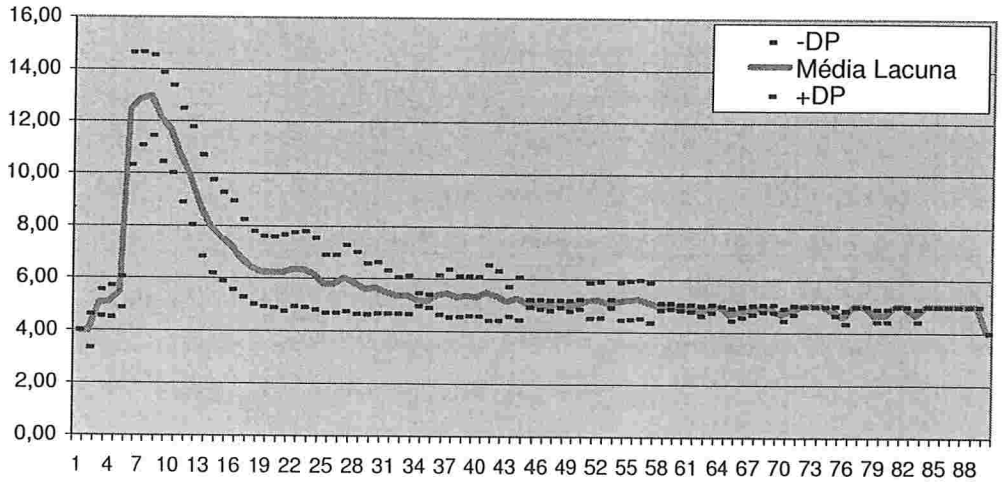

Figura 12.0 Médias e bandas de um desvio padrão de K_T pela Lacuna a cada iteração

Comparando a média estimada pela Lacuna com a mediana, optamos pela última por acertar o número de agrupamentos ideais após cerca de 16 iterações de um total de 90. Outros trabalhos semelhantes de estimação do número de grupos atentam para o fato da mediana do k dado por alguns algoritmos ser mais robusto do que a média, (Fridlyand 2001). E como já foi explicitado que em média o processo é próximo da duração $16 \ln 16$, necessitou-se de pouco mais de $1/\ln 16$, aproximadamente $1/3$ (ou 0,3607) do tempo médio de duração do processo para se avaliar com grande margem de acerto a quantidade ideal de aglomerados estocásticos (com elementos evoluindo estocasticamente no tempo) pelo método da Lacuna.

O objetivo de definir uma regra de parada, isto é, uma forma de interromper o cálculo do número de aglomerados no tempo baseado nos valores anteriormente calculados, de k_1 a k_{t-1} é alcançado não necessariamente após ou antes de $1/3$ da vida útil do processo, onde vida útil é o tempo necessário para que todos no n elementes seja escolhidos e alocados a um dos 4 destinos. As simulações indicaram que logo no início da evolução estocástica, quando os primeiros elementos vão sendo escolhidos e alocados, o método da Lacuna tende a estimar o número de aglomerados por valores cada vez mais altos. Isto é em parte explicado pela configuração geométrica dos pontos iniciais. Como eles estão igualmente afastados uns dos outros e os destinos vão sendo preenchidos, sendo estes elementos quase que ficando sobrepostos (a menos do ruído branco), a avaliação é de que existem tantos grupos quanto elementos não sobrepostos. O método calcula $\log W_k$ e o compara com a esperança desta função calculada sob uma distribuição uniforme, que possui seus pontos também igualmente espaçados. Daí a curva do GAP e , portanto, a curva da Lacuna, evidenciar valores que beiram a quantidade de pontos inicial. Veja no gráfico que este efeito rapidamente decai com a escolha de mais elementos para serem alocados aos 4 destinos até que ele desaparece completamente e resta apenas os 4 aglomerados de destino, já devidamente preenchidos por pelo menos um ponto, e no aglomerado central restam mais alguns pontos para serem escolhidos. Ou melhor, no início, cada ponto é traduzido pelo método como sendo um aglomerado mas, em poucas iterações, os elementos, antes igualmente espaçados,vão sendo afastados uns dos outros e a distância relativa entre eles cresce, fazendo com que o algoritmo comece a "perceber" os grupos de destino e o grupo originário apenas.

Em razão das evidências apresentadas advindas da simulação de Monte Carlo e do estudo do comportamento das funções utilizadas tanto no cálculo do GAP quanto da

Lacuna, podemos lançar a proposta de escolher a regra de parada como sendo o instante, após uma sequência de fortes quedas no k_T, em que a mediana se estabiliza em um valor e se mantém constante. A partir do instante em que mais iterações são observadas, maiores são as chances de já ter sido alcançado o k ideal mas acredita-se que basta olhar até o instante que representa $1/\ln n$ do tempo médio de duração do processo para se observar este fenômeno. Caso o fenômeno descrito antes, ocorra, sugere-se que se calcule o k_T até o instante $1/\ln n$, para evitar uma possível recorrência do mesmo. Se até esse instante o fenômeno da queda abrupta não tenha ocorrido, deve-se continuar até a sua primeira ocorrência.

Formalmente, para $\varepsilon < 1$, seja:

$$T_1^{\varepsilon} = \inf\left[t \geq 0 : \frac{k_t}{k_{t-1}} < \varepsilon\right]$$

O tempo de parada T proposto é:

$$T^{\varepsilon} = \max\left[T_1^{\varepsilon}, n\right]$$

Isto equivale a dizer, que no processo aleatório em estudo, o número ideal de aglomerados deve ser: o valor de k no instante que representa $1/\ln n$ da proporção de duração média do referido processo ou o primeiro instante em que a taxa da razão de queda dos k_t´s sofra um decaimento abrupto, o que for maior. Nas simulações feitas, foi usado o valor de $\varepsilon = 1/2$. Vale ressaltar que esta regra de parada só vale para condições equivalentes às simulações de Monte Carlo, não se tem a pretensão de generalizá-lo neste trabalho.

Capítulo 5

Resultados e Considerações Finais

A definição de aglomerado não é consensual, também não o é a definição do número ótimo de aglomerados. Daí, podemos intuir do por quê da existência de diversos métodos estatísticos e não-estatísticos para abordar o problema. Ademais, um outro fator complicador é a definição de "ideal", tanto que neste trabalho não nos atrevemos a deixar o ambiente mais conturbado pela criação de outra definição nova. Buscamos contornar o problema diferentemente dos outros trabalhos nesta área. Enquanto a grande maioria dos outros trabalhos busca avaliar os métodos de inferência do número de agrupamentos gerando quantidades expressivas de conjuntos de dados advindos de uma ou mais distribuições multivariadas, a nossa proposta foi a de gerar amostras de um mesmo processo estocástico evoluindo no tempo. Ao invés de gerarmos configurações diferentes baseadas em uma mesma lei de densidade de probabilidade, deixamos que as diversas configurações (iterações do processo) fossem geradas aleatoriamente baseadas numa lei pré-definida de transição markoviana, em que o estado futuro da configuração dos elementos se baseiam apenas na configuração presente e numa variável aleatória. Com esta decisão, acreditamos que poderíamos contornar o problema de definir qual o número ideal de aglomerados esperados pela introdução de um modelo com dinâmica estocástica cujo estado final dos elementos (pontos num hiperespaço) seja equivalente ao número de agrupamentos buscado. Na literatura, existe a divisão entre procedimentos que buscam avaliar o número de agrupamentos gerados por distribuições com médias centradas longe umas das outras, grupos ditos não sobrepostos, e quando há sobreposição entre grupos, avaliados em modelos de mistura de distribuições. No nosso caso, não há necessidade de se optar por um ou por outro porque

pode ou não haver sobreposição no decorrer do processo, o importante é que sabemos com certeza o número final de grupos pois estes são os destinos dos pontos da configuração inicial, isto é, sua configuração invariante limite.

Por não haver a prerrogativa de que o método proposto neste trabalho visa descobrir grupos sobrepostos ou não, acredita-se que a técnica de análise de aglomerados subjacente a este método também tenha que ter a mesma propriedade de generalidade de uso. Ou seja, há que se ter um método de estimação do número de agrupamentos que independa do algoritmo de aglomeração utilizado. Isto não é a regra geral na literatura recente de análise de aglomerados mas felizmente é o caso do método proposto neste trabalho. Isto se deve porque o método *Lacuna* busca descobrir e mensurar o já mencionado "fenômeno do cotovelo" que é o ponto em que uma medida de variabilidade passa de uma forte queda para uma forte estabilização. E este fenômeno, por sua vez, independe do algoritmo de análise de aglomerados, apenas depende da forma como vai ser medida a variabilidade e de como são definidos o que é uma forte queda e uma estabilização.

A maneira mais adequada de mensuração destas características foi encontrada nos trabalhos de Tibshirani et al. A contribuição deles foi a definição de uma função denominada GAP que compara uma função de dispersão dos elementos com a esperança da mesma função caso esta fosse calculada sobre uma distribuição de referência. Assim, esse trabalho foi o núcleo de partida da dissertação em questão, que teve seus primeiros passos constituídos pelo entendimento, por meio de diversas simulações, da metodologia de cálculo do número de aglomerados pelo método GAP. Para termos a certeza de que este trabalho realmente era inovador e não havia algo mais chamativo neste campo de estudo, uma pesquisa foi feita e a cada método diferente que buscava atacar o problema, cresceu a intuição de que o método GAP realmente tinha boas propriedades aliadas a idéias interessantes. À medida em que fortalecia o comprometimento com a abordagem usada por Tibshirani et al. e que simulações foram sendo executadas, a decisão de tentar alterar este método para contemplar outros casos em que o método GAP não estava tendo um bom desempenho foi fortalecida. Mais especificamente, surgiram indícios de que o método subestimava o número de aglomerados em bases de dados com uma estrutura geométrica evidente e espaçada em vários polígonos de elementos iguais. Estruturas de dados com pouca dispersão em poucos grupos estavam sendo bem diagnosticadas, em contrapartida, estruturas definidas em vários grupos não eram captadas. No entanto, percebeu-se que não

era o núcleo do algoritmo, a função GAP, que não captava a real quantidade de grupos e sim, a metodologia empregada que falhava nestes casos. O uso da função "arg min" das diferenças sucessivas entre o GAP em k e o GAP em $k+1$ (formalmente o menor k tal que esta diferença seja positiva, descontado o desvio padrão) era onde residia o problema. Neste momento, a partir da constatação empírica desta situação, surgiu a oportunidade de alterar o método proposto para contemplar um conjunto particular de estruturas geométricas de conjuntos de pontos cujos grupos não estavam sendo bem estimados. Desta forma, originou o que convencionamos chamar de aglomerados estocásticos e cujo problema inerente era o de calcular o número ótimo de grupos no tempo, conjuntos internamente coesos e externamente isolados que evoluem segundo uma dinâmica markoviana no tempo.

Os subsídios deste estudo forma relativamente escassos no que concerne a existência de resultados provados, teoremas, lemas, etc., principalmente porque a distribuição de probabilidades das variáveis aleatórias utilizadas nos diversos cálculos são difíceis de serem avaliadas. Estas variáveis costumam ser funções de medidas de distâncias entre elementos e que, por si só, já oferecem bastante complexidade à tentativa de descoberta da densidade de probabilidades. Assim, nos restou trabalhar com a esperança e a variância e, atrelados aos teoremas para famílias de distribuições unimodais de referências, derivar um resultado importante: a constatação e referida prova de que o número estimado de aglomerados tanto no método GAP quanto no Lacuna não são alterados por variações no suporte da distribuição uniforme de referência. Com esta prova, foi possível definir a janela do modelo estocástico adotado, livremente do suporte da distribuição de referência.

Ao avaliar o comportamento do método GAP sob o processo estocástico que foi definido e possuindo os subsídios teóricos dos outros modelos mais citados na literatura, concebemos um procedimento alternativo baseado na função GAP para avaliar o número de grupos no tempo com maior precisão (em termos de acerto do número k) e com maior eficiência (em termos da quantidade de iterações necessárias). Chegamos no método Lacuna por meio de uma pequena alteração na formulação do GAP e avaliamos comparativamente a sua precisão mediana. Por intermédio de repetições de simulações independentes, atestamos que a mediana era mais preciso do que pelo GAP. No entanto, a eficiência ficaria discutível pois enquanto esse não necessita do cálculo da função GAP para todos os possíveis valores de k (pois pega o argmin das diferenças de GAP's subsequentes), o nosso método necessitaria de todos os valores, de $k = 1, ..., k = n$. Consequentemente, o nosso

método precisaria ser tão preciso quanto havíamos atestado mas em muito menos tempo de decorrência do processo. E com isto, surgiu a necessidade de criar uma regra de parada para tal propósito. Mesmo não tendo subsídios teóricos para provar que o estimador no tempo de parada possui todas as características desejáveis de um bom estimador, as simulações por Monte Carlo nos permite lançar mão de poderosas evidências, uma das quais sendo que necessitamos de cerca de 1/3 da duração média do processo até o equilíbrio para termos uma boa estimativa da quantidade terminal de grupos, isto sendo apenas uma conjectura.

A generalidade tão arduamente perseguida neste trabalho é sacrificada no fim às custas de se ter um resultado forte indicado pelas simulações executadas. Isto não tira de forma alguma a qualidade das conclusões apenas as restringe por um lado, por outro, nos fornece grandes perspectivas de continuação do trabalho em busca de atestar (1) a precisão do novo método em outros sistemas de partículas markovianos, (2) a eficiência da regra de parada para mais elementos/partículas e mais (ou menos) grupos de destino.

Concluímos dizendo que este trabalho consta de uma certa audácia, fruto da tentativa de compilar poderosas idéias que foram concebidas nos últimos anos; da tentativa de contribuir com a ciência pela criação de uma outra metodologia para a resolução de um problema clássico; da tentativa de quebrar os paradigmas atuais e propor uma outra forma de mensurar o grau com que o método consegue atingir o seu propósito; e, por fim, expor com honestidade e simplicidade o longo e tortuoso caminho das pedras, passo a passo.

FIM

Referências Bibliográficas

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in statistics* **3**: 1–27.

Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters, *The Canadian Journal of Statistics* **28**: 367–382.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis*, John Wiley and Sons, New York.

Fraley, C. and Raftery, A. (1998). How many clusters? which clustering method? answers via model-based cluster analysis, *The Computer Journal* **41**: 578–588.

Fridlyand, Y. (2001). *Resampling methods for variable selection and classification: application to genomics*, Graduate Devision of Statistics, University of California, Berkeley.

Gordon, A. (1999). *Classification*, Chapman and Hall, London.

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

Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall, *Psychological Bulletin* **83**: 1072–1080.

Marriott, F. (1971). Practical problems in a method of cluster analysis, *Biometrics* **27**: 501–514.

Milligan, G. W. e Cooper, M. C. (1985). An examination of procedures for determining the number of groups in a data set, *Psicometrika* **50**: 159–179.

Pollard, D. (1982). A central limit theorem for k-means clustering, *The Annals of Probability* **10**: 919–926.

SAS Institute user´s guide: statistics, R. (2003). Cubic clustering criterion, *http://onlinedoc.sas.com.*

Tibshirani, R., Walter, G. and Hastie, T. (2001a). Estimating the number of clusters in a data set via the gap statistics, *Journal of the Royal Statistical Society, Ser.* **B 63**: 411–423.

Tibshirani, R., Walter, G., Botstein, D. and Brown, P. (2001b). Cluster validation by prediction strenth, *http://www-stat.stanford.edu/ tibs/research.html.*