ANÁLISE DE CORRESPONDÊNCIA

ANA MUSETTI RAMOS DE SOUZA

DISSERTAÇÃO APRESENTADA

AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR:

PROF: DR. ADOLPHO WALTER PIMAZONI CANTON

- SÃO PAULO, ABRIL DE 1982 -

AGRADECIMENTOS

Gostaria de agradecer a todos aqueles que, de um modo ou de outro, colaboraram na execução deste trabalho. Agradeço em particular:

Ao Professor Doutor Adolpho Walter Pimazoni Canton, meu orientador.

Aos colegas e amigos da Assessoria Técnica do Centro de Computação Eletrônica da Universidade de São Paulo, pela grande ajuda prestada no uso do computador.

Ao Professor Carlos Roberto Azzoni, que muito colaborou nas aplicações numéricas deste trabalho.

E ao Sr. João Baptista Esteves de Oliveira, pela dedicação ao trabalho de datilografia.

- 0 -

ÍNDICE

INTRODUÇÃO

Os métodos de análise de dados têm como objetivo fornecer representações suscintas das grandes tabelas de dados, ou seja, as tabelas que contêm os dados de conjuntos de várias variáveis medidas, possivelmente, em muitas unidades experimentais. Estas tabelas, em geral, apresentam dimensões muito grandes, tornando difícil interpretar os resultados e informações que contêm. Os princípios básicos de todos os métodos de análise de dados são semelhantes.

Um desses métodos, a *Análise de Correspondência*, se aplica às grandes tabelas cruzadas ou tabelas de contingência, ou seja, se aplica basicamente para variáveis discretas. Entretanto pode também ser aplicada para variáveis contínuas desde que estas estejam codificadas de maneira apropriada, como por exemplo, divididas em classes. A análise de correspondência é um algoritmo de redução de dados, que fornece uma imagem simplificada e deformada da realidade, mas com regras de interpretação que permitem retornar às estruturas reais à partir das imagens obtidas.

Comparando o domínio de aplicação da análise de correspondência com o da análise de componentes principais, temos que enquanto este se resume a um conjunto de variáveis (que não têm caracter repetitivo) e um conjunto de observações (que têm caracter repetitivo) dessas variáveis, o domínio de aplicação da

análise de correspondência se estende para dois ou mais conjuntos de variáveis (sem caracter repetitivo) cruzadas. A técnica é aplicada a cada conjunto de variáveis de modo a obter, para cada um deles, eixos fatorais ou fatores que sejam em menor número que o número de variáveis e que contenham o máximo possível da informação das variáveis. O objetivo final e fundamental deste método é obter a melhor representação simultânea de dois ou mais conjuntos de variáveis através de gráficos representando cada variável nos planos de projeção formados pelos primeiros eixos fatoriais cruzados dois a dois.

No caso em que temos dois conjuntos de variáveis cruzados, a análise é dita análise de correspondência binária e no caso em que temos mais de dois conjuntos de variáveis cruzadas, a análise é dita de correspondência múltipla.

As tabelas de contingências são também chamadas tabelas de dependência. A análise de correspondência se aplica, em geral, quando a hipótese de independência é rejeitada. Há então um interesse em se estudar as relações que existem entre os conjuntos de variáveis que estão sendo cruzados. Se existe relação entre os elementos de diferentes conjuntos de variáveis, diz-se que estes conjuntos estão em correspondência. Daí o nome análise de correspondência.

O número máximo de fatores calculado é igual ao número do menor conjunto de variáveis, mas em geral só os dois, três ou quatro primeiros fatores contêm uma informação realmente significante.

A interpretação dos resultados da análise é, entretanto,

muito ampla e subjetiva e só o treino e a prática do estatístico e/ou a experiência e o conhecimento do pequisador podem auxiliar, facilitar e tornar plausível esta interpretação.

É difícil fazer um histórico preciso da técnica de análise de correspondência. A primeira publicação foi feita por Hirschfeld (1935) mas a partir daí a técnica foi amplamente negligenciada e só mais tarde foi redescoberta independentemente por vários autores: Willians (1952), Kendall & Stuart (1961), Benzécri (1969) e Lancaster (1969). O desenvolvimento dado por Hirschfeld não foi citado por Fisher em 1940 quando este desenvolveu uma técnica semelhante sob o nome de "análise canônica para tabelas de contingência" e assim Fisher é freqüentemente apontado como o introdutor do método. Os trabalhos mais recentes são devidos a Benzécri e seus colaboradores entre os quais destacam-se: Lebart, Morineau, Tabard e Fenelon.

O objetivo principal deste trabalho é apresentar um estudo sobre a validade e a qualidade dos resultados da análise de correspondência. Para tanto, o mesmo foi desenvolvido da seguinte forma:

No Capítulo 1 apresentamos a Análise de Correspondência Binária, depois de introduzirmos a base teórica comum aos números métodos de análise de dados sob o nome de "análise geral".

No Capítulo 2 apresentamos a técnica da Análise de Correspondência Múltipla, que é uma generalização da técnica binária apresentada no Capítulo 1.

No Capítulo 3 fazemos uma discussão sobre a qualidade e

a validade dos resultados fornecidos pela Análise de Correspondência.

No Capítulo 4 apresentamos exemplos de aplicação do método da Análise de Correspondência, dando as devidas interpretações aos resultados e discutindo a qualidade e a validade dos mesmos. Foram usados programas de computador desenvolvidos pela CESIA (Centre de Statistique et d'Informatique Appliquées) e implantados no CCE-USP.

CAPÍTULO 1

ANÁLISE DE CORRESPONDÊNCIA BINÁRIA

Para facilitar a apresentação das técnicas de análise de correspondência binária e múltipla (Capítulo 2), iniciamos este capítulo com a base teórica comum aos numerosos métodos de análise de dados, que chamamos de "análise geral". Apresentamos, a seguir, a Análise de Correspondência Binária, quando somente dois conjuntos de variáveis são postos em correspondências.

O método de Análise de Correspondência se aplica principalmente às tabelas de contingência, também chamadas tabelas de dependência ou tabelas cruzadas. Mas sua aplicação é permitida no caso de tabelas contendo variáveis contínuas desde que estas estejam separadas em classes.

Uma vez verificada a dependência entre as linhas e as colunas da tabela, o método de Análise de Correspondência procura a "melhor" representação simultânea dos dois conjuntos de variáveis constituídos pelas linhas e colunas da tabela, fornecendo gráficos onde aparecem nuvens de pontos-linhas e pontos-colunas nos planos de projeção formados pelos primeiros eixos fatoriais tomados dois a dois. A interpretação dos resultados é feita com a ajuda de alguns coeficientes como a "contribuição asboluta" e "contribuição relativa", que relacionam os fatores encontrados na análise com as variáveis.

1.1 - ANÁLISE GERAL

Considere uma tabela de valores numéricos (representada por uma matriz) com n linhas e p colunas. Procuramos uma técnica de redução da dimensão do espaço que se aplique a diversos tipos de tabelas, extraindo o essencial da informação contida na mesma.

A tabela $\underset{\sim}{X}$ de valores numéricos permite duas representações geométricas:

em $\mathbb{R}^n$: temos uma nuvem de p pontos cujas coordenadas são os n elementos das linhas da tabela (cada coluna pode ser considerada como um vetor com n componentes);

em $\mathbb{R}^p$: temos uma nuvem de n pontos cujas coordenadas são os p elementos das colunas da tabela (cada linha pode ser considerada como um vetor com p componentes).

Apresentamos a técnica de redução, sucessivamente, nos espaços vetoriais $\mathbb{R}^p$ e $\mathbb{R}^n$.

1.1.1 - Ajustamento por um sub-espaço vetorial de $\mathbb{R}^p$

Em $\mathbb{R}^p$ temos uma nuvem de n pontos (pontos-linhas). O ajustamento por meio de um sub-espaço vetorial consiste em encontrar um sub-espaço $\mathbb{R}^q$ com $q << p$, e que contenha, ou melhor, represente a nuvem dos pontos-linhas.

Iniciamos procurando um sub-espaço vetorial com uma di-

mensão, ou seja, queremos uma reta que passe pela origem e que seja o "melhor", por algum critério, ajustamento possível da nuvem de pontos.

Seja $\underset{\sim}{u}$ um vetor (coluna) unitário, isto é, $\underset{\sim}{u}'\underset{\sim}{u} = 1$

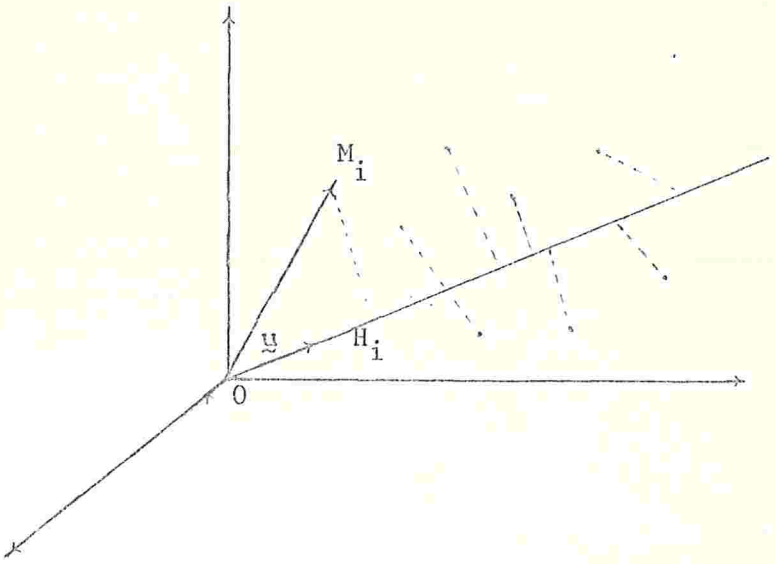


Figura 1.1 - Projeções dos pontos sobre o sub-espaço de uma dimensão

A Figura 1.1 mostra a projeção OH_i de um vetor OM_i sobre o sub-espaço de dimensão um definido por $\underset{\sim}{u}$. Esta projeção é o produto escalar de OH_i por $\underset{\sim}{u}$, ou seja, a soma dos produtos, termo a termo, das componentes de OM_i e de $\underset{\sim}{u}$.

Cada linha de $\underset{\sim}{X}$ é um vetor de $\mathbb{R}^p$ e o produto $\underset{\sim}{X}\cdot\underset{\sim}{u}$ é uma matriz coluna (vetor) com n componentes, onde cada componente é o produto escalar de uma linha de $\underset{\sim}{X}$ por $\underset{\sim}{u}$. Logo, as n componentes de $\underset{\sim}{X}\ \underset{\sim}{u}$ são as n projeções dos n pontos da nuvem sobre $\underset{\sim}{u}$.

Usaremos o critério clássico dos "Mínimos Quadrados" para ajustar a nuvem de n pontos ao sub-espaço, ou seja, deve-

mos minimizar a soma de quadrados dos desvios:

$$\sum_{i=1}^{n} (M_iH_i)^2 .$$

Aplicando o Teorema de Pitágoras a cada um dos n triângulos retângulos H_iOM_i, i=1,...,n, segue a relação:

$$\sum_{i=1}^{n} (M_iH_i)^2 = \sum_{i=1}^{n} (OM_i)^2 - \sum_{i=1}^{n} (OH_i)^2 .$$

A soma $\sum_{i=1}^{n} (OM_i)^2$ é fixa pois é característica da nuvem. Logo, minimizar

$$\sum_{i=1}^{n} (M_iH_i)^2$$

é equivalente a maximizar $\sum_{i=1}^{n} (OH_i)^2$.

Exprimindo $\sum_{i=1}^{n} (OH_i)^2$ em função de $\underset{\sim}{X}$ e $\underset{\sim}{u}$ vem:

$$\sum_{i=1}^{n} (OH_i)^2 = (\underset{\sim}{X}\underset{\sim}{u})'\underset{\sim}{X}\underset{\sim}{u} = \underset{\sim}{u}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{u}$$

Para encontrar $\underset{\sim}{u}$, devemos maximizar a forma quadrática $\underset{\sim}{u}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{u}$ em relação a $\underset{\sim}{u}$ sujeito à condição $\underset{\sim}{u}'\underset{\sim}{u} = 1$.

Seja $\underset{\sim}{u}_1$ o vetor encontrado. O sub-espaço de duas dimensões que se ajusta "melhor" a nuvem contém o sub-espaço formado por $\underset{\sim}{u}_1$. Devemos procurar, então, um vetor $\underset{\sim}{u}_2$, segundo vetor de base do sub-espaço, que seja ortogonal a $\underset{\sim}{u}_1$ e que maximize $\underset{\sim}{u}_2'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{u}_2$.

De modo análogo, chegamos ao "melhor" sub-espaço a q dimensões ($q << p$).

1.1.2 - O máximo de uma forma quadrática

Colocando o problema de uma forma mais geral, procuramos um vetor $\underset{\sim}{u}$ que maximiza a quantidade $\underset{\sim}{u}'\underset{\sim}{A}\underset{\sim}{u}$ com a condição

$$\underset{\sim}{u}'\underset{\sim}{M}\underset{\sim}{u} = 1$$

onde $\underset{\sim}{A}$ e $\underset{\sim}{M}$ são matrizes simétricas e $\underset{\sim}{M}$ é definida não-negativa. (No nosso caso temos $\underset{\sim}{A} = \underset{\sim}{X}'\underset{\sim}{X}$ e $\underset{\sim}{M} = \underset{\sim}{I}$, I = identidade).

Daremos duas demonstrações para a solução deste problema. A primeira usando os "multiplicadores de Lagrange" e a segunda usando "propriedades espectrais" de matrizes simétricas.

a) Primeira demonstração:

Para achar o máximo de $\underset{\sim}{u}'\underset{\sim}{A}\underset{\sim}{u}$ com a condição de $\underset{\sim}{u}'\underset{\sim}{M}\underset{\sim}{u} = 1$ devemos ter:

$$\frac{d}{d\underset{\sim}{u}}[\underset{\sim}{u}'\underset{\sim}{A}\underset{\sim}{u} - \lambda(\underset{\sim}{u}'\underset{\sim}{M}\underset{\sim}{u} - 1)] = 0$$

onde λ é o multiplicador de Lagrange.

Derivando em relação a $\underset{\sim}{u}$, temos:

$$2\underset{\sim}{A}\underset{\sim}{u} - 2\lambda\underset{\sim}{M}\underset{\sim}{u} = 0 \Rightarrow \underset{\sim}{A}\underset{\sim}{u} = \lambda\underset{\sim}{M}\underset{\sim}{u}. \qquad (1.1)$$

Pré-multiplicando por $\underset{\sim}{u}'$ e como $\underset{\sim}{u}'\underset{\sim}{M}\underset{\sim}{u} = 1$ temos:

$$\lambda = \underset{\sim}{u}'\underset{\sim}{A}\underset{\sim}{u},$$

onde λ é o máximo procurado.

Como $\underset{\sim}{M}$ é inversível de (1.1) segue: $\underset{\sim}{M}^{-1}\underset{\sim}{A}\underset{\sim}{u} = \lambda\underset{\sim}{u}$, ou seja $\underset{\sim}{u}$ é o vetor característico (v.c.) da matriz $\underset{\sim}{M}^{-1}\underset{\sim}{A}$ correspondente a maior raiz característica (r.c.) λ.

Seja $\underset{\sim}{u}_1$ o v.c. correspondente a maior r.c. λ_1 tal que (1.1)

é verificado. Procuramos agora o vetor $\underset{\sim}{u}_2$, M-ortogonal a $\underset{\sim}{u}_1$ ($\underset{\sim}{u}_2'\underset{\sim}{M}\underset{\sim}{u}_1 = 0$), unitário ($\underset{\sim}{u}_2'\underset{\sim}{M}\underset{\sim}{u}_2 = 1$) e que minimize $\underset{\sim}{u}_2'\underset{\sim}{M}\underset{\sim}{u}_2$.

Usando Lagrange novamente:

$$\frac{d}{d\underset{\sim}{u}_2}[\underset{\sim}{u}_2'\underset{\sim}{A}\underset{\sim}{u}_2 - \lambda_2(\underset{\sim}{u}_2'\underset{\sim}{M}\underset{\sim}{u}_2 - 1) - \mu_2'\underset{\sim}{M}\underset{\sim}{u}_1] = 0$$

onde λ_2 e μ_2 são os multiplicadores de Lagrange. Temos:

$$2\underset{\sim}{A}\underset{\sim}{u}_2 - 2\lambda_2\underset{\sim}{M}\underset{\sim}{u}_2 - \mu_2\underset{\sim}{M}\underset{\sim}{u}_1 = 0.$$

Pré-multiplicando por $\underset{\sim}{u}_1'$ e sabendo que $\underset{\sim}{u}_1'\underset{\sim}{A}\underset{\sim}{u}_2 = \lambda_2\underset{\sim}{u}_1'\underset{\sim}{M}\underset{\sim}{u}_2 = 0$, vem que $\mu_2 = 0$, ou seja:

$$\underset{\sim}{A}\underset{\sim}{u}_2 = \lambda_2\underset{\sim}{M}\underset{\sim}{u}_2 \Rightarrow \underset{\sim}{M}^{-1}\underset{\sim}{A}\underset{\sim}{u}_2 = \lambda_2\underset{\sim}{u}_2.$$

Portanto λ_2 é a segunda maior r.c. de $\underset{\sim}{M}^{-1}\underset{\sim}{A}$ e $\underset{\sim}{u}_2$ é o v.c. correspondente.

A demonstração se estende de modo análogo para um vetor unitário $\underset{\sim}{u}_\alpha$ ($\underset{\alpha}{u}'\underset{\sim}{M}\underset{\sim}{u}_\alpha = 1$), M-ortogonal com os vetores $\underset{\sim}{u}_\beta$ precedentes ($\underset{\sim}{u}_\alpha'\underset{\sim}{M}\underset{\sim}{u}_\beta = 0$, $\beta<\alpha$) e que maximiza $\underset{\sim}{u}_\alpha'\underset{\sim}{A}\underset{\sim}{u}_\alpha$. Chegaremos a:

$$\underset{\sim}{M}^{-1}\underset{\sim}{A}\underset{\sim}{u}_\alpha = \lambda_\alpha\underset{\sim}{u}_\alpha \quad \text{e} \quad \underset{\sim}{u}_\alpha$$

será o v.c. correspondente a α-ésima a.c. de $\underset{\sim}{M}^{-1}\underset{\sim}{A}$, $\alpha<p$.

OBSERVAÇÃO - No nosso caso particular tínhamos $\underset{\sim}{A} = \underset{\sim}{X}'\underset{\sim}{X}$ e $\underset{\sim}{M} = \underset{\sim}{I}$. Logo $\underset{\sim}{M}^{-1}\underset{\sim}{A} = \underset{\sim}{X}'\underset{\sim}{X}$.

b) Segunda demonstração:

Daremos um esboço desta demonstração no caso em que $\underset{\sim}{M}$ é positiva definida, o que permite decompor $\underset{\sim}{M}$ na forma clássica

$\underset{\sim}{M} = \underset{\sim}{L}'\underset{\sim}{L}$, onde $\underset{\sim}{L}$ é inversível (Rao,1973).

Seja $\underset{\sim}{y} = \underset{\sim}{L}\underset{\sim}{u}$ ou seja $\underset{\sim}{u} = \underset{\sim}{L}^{-1}\underset{\sim}{y}$. A condição de normalização $\underset{\sim}{u}'\underset{\sim}{M}\underset{\sim}{u} = 1$ fica então

$$u'Mu = \underbrace{\underset{\sim}{y}'\underset{\sim}{L}^{-1'}}_{\underset{\sim}{u}'} \underbrace{\underset{\sim}{L}'\underset{\sim}{L}}_{\underset{\sim}{M}} \underbrace{\underset{\sim}{L}^{-1}\underset{\sim}{y}}_{\underset{\sim}{u}} = \underset{\sim}{y}'\underset{\sim}{y} = 1$$

e a quantidade $\underset{\sim}{u}'\underset{\sim}{A}\underset{\sim}{u}$ que deve ser maximizada torna-se:

$$\underset{\sim}{u}'\underset{\sim}{A}\underset{\sim}{u} = \underset{\sim}{y}'\underset{\sim}{L}^{-1'}\underset{\sim}{A}\underset{\sim}{L}^{-1}\underset{\sim}{y} = \underset{\sim}{y}'\underset{\sim}{A}_0\underset{\sim}{y}$$

onde $\underset{\sim}{A}_0 = \underset{\sim}{L}^{-1'}\underset{\sim}{A}\underset{\sim}{L}^{-1}$.

$\underset{\sim}{A}_0$ é simétrica e pode ser diagonalizada. Seja $\underset{\sim}{T}$ a matriz ortogonal $p \times p$ cujas p colunas são os os v.c. t_α de $\underset{\sim}{A}_0$ normalizados e ordenados segundo λ_α decrescente, onde λ_α são as r. c. de $\underset{\sim}{A}_0$. Se $\underset{\sim}{\Lambda}$ é a matriz diagonal cujo α-ésimo elemento da diagonal vale λ_α, então podemos escrever $\underset{\sim}{A}_0 = \underset{\sim}{T}\underset{\sim}{\Lambda}\underset{\sim}{T}'$.

Seja $\underset{\sim}{c} = \underset{\sim}{T}'\underset{\sim}{y}$. Então $\underset{\sim}{y} = \underset{\sim}{T}\underset{\sim}{c}$ pois, $T' = T^{-1}$.

A quantidade a minimizar $\underset{\sim}{y}'\underset{\sim}{A}_0\underset{\sim}{y}$ fica então:

$$\underset{\sim}{y}'\underset{\sim}{A}_0\underset{\sim}{y} = \underset{\sim}{y}'\underset{\sim}{T}\underset{\sim}{\Lambda}\underset{\sim}{T}'\underset{\sim}{y} = \underset{\sim}{c}'\underset{\sim}{\Lambda}\underset{\sim}{c}$$

e a condição $\underset{\sim}{y}'\underset{\sim}{y} = 1$ fica:

$$\underset{\sim}{y}'\underset{\sim}{y} = \underset{\sim}{c}'\underset{\sim}{T}'\underset{\sim}{T}\underset{\sim}{c} = \underset{\sim}{c}'\underset{\sim}{c} = 1.$$

Temos que:

$$\lambda_1 - \underset{\sim}{c}'\underset{\sim}{\Lambda}\underset{\sim}{c} = \underbrace{\lambda_1\underset{\sim}{c}'\underset{\sim}{c}}_{1} - \underset{\sim}{c}'\underset{\sim}{\Lambda}\underset{\sim}{c} = c'(\lambda_1\underset{\sim}{I} - \Lambda)\underset{\sim}{c} \geq 0$$

pois é uma forma quadrática semi-positiva definida.

Então, como $\lambda_1 - \underset{\sim}{c}'\underset{\sim}{\Lambda}\underset{\sim}{c} \geq 0$ segue que $\lambda_1 \geq \underset{\sim}{c}'\underset{\sim}{\Lambda}\underset{\sim}{c}$. Portanto o má-

ximo de $\underset{\sim}{c}'\underset{\sim}{A}\underset{\sim}{c}$ é alcançado quando $\underset{\sim}{c}'\underset{\sim}{A}\underset{\sim}{c} = \lambda_1$, onde λ_1 é o maior r. c. de $\underset{\sim}{A}_0$.

O máximo λ_1 é alcançado quando $\underset{\sim}{c}' = (1,0,0,\ldots,0)$ pois

$$(1,0,0,\ldots,0)\Lambda\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix} = \lambda_1.$$

Se $\underset{\sim}{c}' = (1,0,0,\ldots,0)$ segue que

$$\underset{\sim}{y} = \underset{\sim}{T}\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix} = \underset{\sim}{t}_1$$

onde $\underset{\sim}{t}_1$ é o v.c. correspondente a λ_1.

Como $\underset{\sim}{u} = \underset{\sim}{L}^{-1}\underset{\sim}{y}$, a solução é $\underset{\sim}{u}_1 = \underset{\sim}{L}^{-1}\underset{\sim}{t}_1$. Da relação $\underset{\sim}{A}_0\underset{\sim}{t}_1 = \lambda_1\underset{\sim}{t}_1$ vem:

$$\underset{\sim}{L}^{-1'}\underset{\sim}{A}\underset{\sim}{L}^{-1}\underset{\sim}{t}_1 = \lambda_1\underset{\sim}{t}_1 = \lambda_1\underset{\sim}{L}\underset{\sim}{u}_1.$$

Multiplicando os dois membros por $\underset{\sim}{L}^{-1}$ vem:

$$\underbrace{\underset{\sim}{L}^{-1}\underset{\sim}{L}^{-1'}}_{\underset{\sim}{M}^{-1}}\underset{\sim}{A}\underbrace{\underset{\sim}{L}^{-1}\underset{\sim}{t}_1}_{\underset{\sim}{u}_1} = \lambda_1\underbrace{\underset{\sim}{L}^{-1}\underset{\sim}{L}}_{\underset{\sim}{I}} \Rightarrow \underset{\sim}{M}^{-1}\underset{\sim}{A}\underset{\sim}{u}_1 = \lambda_1\underset{\sim}{u}_1,$$

ou seja $\underset{\sim}{u}_1$ é o v.c. de $\underset{\sim}{M}^{-1}\underset{\sim}{A}$ correspondente a maior r.c. λ_1.

Notemos que para obtermos a solução λ_1 e conseqüentemente $\underset{\sim}{u}_1$, basta proceder a diagonalização da matriz simétrica $\underset{\sim}{A}_0$, depois de decompor $\underset{\sim}{M} = \underset{\sim}{L}'\underset{\sim}{L}$, uma vez que a matriz precedente $\underset{\sim}{M}^{-1}\underset{\sim}{A}$

será na maioria das vezes não simétrica. Esta propriedade é utilizada nos programas de computadores relativos a técnica de Análise de Correspondência.

1.1.3 - Ajustamento por um sub-espaço vetorial de $\mathbb{R}^n$

Em $\mathbb{R}^n$ cada uma das p colunas da tabela $\underset{\sim}{X}$ pode ser considerada como um vetor ou como um ponto de $\mathbb{R}^n$. Temos então, uma nuvem de p pontos-colunas, cada ponto-coluna com n coordenadas. Procuramos um sub-espaço vetorial com q dimensões, $q << n$.

Começando pelo sub-espaço de uma dimensão, a procura de um vetor unitário $\underset{\sim}{v}$ que se ajusta "melhor" a nuvem de $\mathbb{R}^n$, conduz, de maneira análoga a anterior, a maximização da soma de quadrados das projeções dos p pontos sobre $\underset{\sim}{v}$, que são as p componentes de $\underset{\sim}{X}'\underset{\sim}{v}$.

Devemos então maximizr a quantidade $(\underset{\sim}{X}'\underset{\sim}{v})'(\underset{\sim}{X}'\underset{\sim}{v}) = \underset{\sim}{v}'\underset{\sim}{X}\underset{\sim}{X}'\underset{\sim}{v}$ com a condição $\underset{\sim}{v}'\underset{\sim}{v} = 1$.

Como no caso de $\mathbb{R}^p$, isto nos leva a encontrar os q primeiros v.c. de $\underset{\sim}{X}\underset{\sim}{X}'$ correspondentes as q maiores r.c.

Notamos $\underset{\sim}{v}_\alpha$ o α-ésimo v.c. de $\underset{\sim}{X}\underset{\sim}{X}'$ correspondente a r. c. μ_α.

1.1.4 - Relação entre os ajustamentos de $\mathbb{R}^p$ e $\mathbb{R}^n$

Em $\mathbb{R}^p$ temos:

$$\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{u}_1 = \lambda_1\underset{\sim}{u}_1. \qquad (1.2)$$

Em $\mathbb{R}^n$ temos:

$$\underset{\sim}{X}\underset{\sim}{X}'\underset{\sim}{v}_1 = \mu_1 \underset{\sim}{v}_1. \qquad (1.3)$$

Pré-multiplicando os dois membros de (1.2) por $\underset{\sim}{X}$, vem:

$$\underset{\sim}{X}\underset{\sim}{X}'(\underset{\sim}{X}\underset{\sim}{u}_1) = \lambda_1(\underset{\sim}{X}\underset{\sim}{u}_1).$$

Portanto, para cada v.c. $\underset{\sim}{u}_1$ de $\underset{\sim}{X}'\underset{\sim}{X}$ relativo à r.c. λ_1 não nula, corresponde um v.c. $\underset{\sim}{X}\underset{\sim}{u}_1$ de $\underset{\sim}{X}\underset{\sim}{X}'$ relativo também à r.c. λ_1. Como μ_1 é a maior r.c. de $\underset{\sim}{X}\underset{\sim}{X}'$, temos necessariamente $\lambda_1 \leq \mu_1$.

Pré-multiplicando os dois membros de (1.3) por $\underset{\sim}{X}'$, vem:

$$\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{v}_1) = \mu_1(\underset{\sim}{X}'\underset{\sim}{v}_1).$$

Portanto, para cada v.c. $\underset{\sim}{v}_1$ de $\underset{\sim}{X}\underset{\sim}{X}'$ relativo à r.c. μ_1 não nula, corresponde um v.c. $\underset{\sim}{X}'\underset{\sim}{v}_1$ de $\underset{\sim}{X}'\underset{\sim}{X}$ relativo também à r.c. μ_1. Como λ_1 é a maior r.c. de $\underset{\sim}{X}'\underset{\sim}{X}$, segue que $\mu_1 \leq \lambda_1$.

Portanto, segue-se que $\mu_1 = \lambda_1$.

Analogamente demonstra-se que todas as r.c. não nulas de $\underset{\sim}{X}'\underset{\sim}{X}$ são iguais as de $\underset{\sim}{X}\underset{\sim}{X}'$.

Logo uma simples transformação linear, associada a matriz $\underset{\sim}{X}$, permite obter os v.c. $\underset{\sim}{X}\underset{\sim}{u}_\alpha$ procurados em $\mathbb{R}^n$, não sendo necessaria a diagonalização de $\underset{\sim}{X}\underset{\sim}{X}'$ para encontrar suas r.c. e v. c.

Temos que

$$\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{u}_\alpha = \lambda_\alpha \underset{\sim}{u}_\alpha \Longrightarrow \underset{\sim}{u}_\alpha'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{u}_\alpha = \lambda_\alpha$$

pois $u_\alpha' u_\alpha = 1$. Então a norma de $\underset{\sim}{X}\underset{\sim}{u}_\alpha$ vale λ_α e como o vetor $\underset{\sim}{v}_\alpha$ corresponde a mesma r.c. λ_α deve ser unitário, segue a relação:

$$\underset{\sim}{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \underset{\sim}{X}\underset{\sim}{u}_\alpha . \qquad (1.4)$$

A norma de $\underset{\sim}{X}'\underset{\sim}{v}_\alpha$ também vale λ_α ($\underset{\sim}{X}\underset{\sim}{X}'\underset{\sim}{v}_\alpha = \lambda_\alpha \underset{\sim}{v}_\alpha \Rightarrow \underset{\sim}{v}_\alpha'\underset{\sim}{X}\underset{\sim}{X}'\underset{\sim}{v}_\alpha = \lambda_\alpha$) e como $\underset{\sim}{u}_\alpha$ também deve ser unitário, segue a relação:

$$\underset{\sim}{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \underset{\sim}{X}'\underset{\sim}{v}_\alpha . \qquad (1.5)$$

Chamamos $\underset{\sim}{u}_\alpha$ de α-ésimo "eixo fatorial" de $\mathbb{R}^p$ e $\underset{\sim}{v}_\alpha$ de α-ésimo "eixo fatorial" de $\mathbb{R}^n$.

Sobre o sub-espaço de $\mathbb{R}^p$ formado por $\underset{\sim}{u}_\alpha$, as coordenadas dos pontos-linhas da nuvem são as componentes do vetor $\underset{\sim}{X}\underset{\sim}{u}_\alpha$, ou, da relação (1.4), as componentes de $\underset{\sim}{v}_\alpha\sqrt{\lambda_\alpha}$. Portanto, as coordenadas dos pontos sobre um eixo fatorial de $\mathbb{R}^p$ ($\mathbb{R}^n$) são proporcionais as componentes do eixo fatorial de $\mathbb{R}^n$ ($\mathbb{R}^p$) correspondentes a mesma r.c.

1.1.5 - Reconstituição da tabela original

Usamos sempre a notação:

$\underset{\sim}{u}_\alpha$: α-ésimo v.c. de norma unitária de $\underset{\sim}{X}'\underset{\sim}{X}$ relativo à r.c. λ_α.

$\underset{\sim}{v}_\alpha$: α-ésimo v.c. de norma unitária de $\underset{\sim}{X}\underset{\sim}{X}'$ relativo à r.c. λ_α.

Da relação (1.4) vem $\underset{\sim}{X}\underset{\sim}{u}_\alpha = \sqrt{\lambda_\alpha}\ \underset{\sim}{v}_\alpha$, para os α-ésimos eixos fatoriais de $\mathbb{R}^p$ e $\mathbb{R}^n$.

Pós-multiplicando os dois membros da relação (1.4) por $\underset{\sim}{u}_\alpha'$ e somando para todos os eixos vem:

$$\underset{\sim}{X} \sum_{\alpha=1}^{p} \underset{\sim}{u}_{\alpha}\underset{\sim}{u}'_{\alpha} = \sum_{\alpha=1}^{p} \sqrt{\lambda_{\alpha}}\ \underset{\sim}{v}_{\alpha}\underset{\sim}{u}'_{\alpha}.$$

Seja $\underset{\sim}{U}$ uma matriz $p \times p$ onde cada coluna é um v.c. de $\underset{\sim}{X}'\underset{\sim}{X}$. Como os v.c. são ortogonais e de norma unitária vem que $\underset{\sim}{U}'\underset{\sim}{U}=\underset{\sim}{I}$ e $\underset{\sim}{U}\underset{\sim}{U}' = \underset{\sim}{I}$. Mas

$$\sum_{\alpha=1}^{p} \underset{\sim}{u}_{\alpha}\underset{\sim}{u}'_{\alpha} = \underset{\sim}{U}\underset{\sim}{U}'.$$

A fórmula anterior fica então

$$\underset{\sim}{X} = \sum_{\alpha=1}^{p} \sqrt{\lambda_{\alpha}}\ \underset{\sim}{v}_{\alpha}\underset{\sim}{u}'_{\alpha} \qquad (1.6)$$

onde as r.c. λ_{α} estão sempre ordenadas em ordem decrescente.

A fórmula (1.6) é fórmula de reconstituição da tabela inicial $\underset{\sim}{X}$ a partir das r.c. λ_{α} e dos v.c. $\underset{\sim}{u}_{\alpha}$ e $\underset{\sim}{v}_{\alpha}$ associados.

Esta somatória pode ser limitada às q ($q<p$) maiores r.c. e teríamos:

$$\underset{\sim}{X} \simeq \underset{\sim}{X}^{*} = \sum_{\alpha=1}^{q} \sqrt{\lambda_{\alpha}}\ \underset{\sim}{v}_{\alpha}\underset{\sim}{u}'_{\alpha}. \qquad (1.7)$$

A qualidade da reconstituição usando só os q primeiros eixos é medida pela relação:

$$\tau_{q} = \sum_{i,j} \underset{\sim}{x}^{*2}_{ij} \Big/ \sum_{i,j} \underset{\sim}{x}^{2}_{ij}$$

ou

$$\tau_{q} = \mathrm{tr}\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^{*}/\mathrm{tr}\underset{\sim}{X}'\underset{\sim}{X}$$

onde tr = traço da matriz.

Substituindo $\underset{\sim}{X}$ e $\underset{\sim}{X}^*$ pelos valores obtidos das relações (1.6) e (1.7), vem:

$$\tau_q = \sum_{\alpha \leq q} \lambda_\alpha \Big/ \sum_{\alpha=1}^{p} \lambda_\alpha$$

O coeficiente τ_q, $\tau_q \leq 1$, é chamado "taxa de inércia" ou "porcentagem de variância" relativa aos q primeiros fatores. Sua interpretação como medida da qualidade numérica da reconstituição é clara, mas, como será visto e discutido posteriormente, o problema de seu significado estatístico é delicado.

1.1.6 - Variáveis e indivíduos suplementares

Na prática, é freqüente dispormos de informações complementares que aumentam a tabela inicial $\underset{\sim}{X}$, que é complementada em colunas por uma tabela $\underset{\sim}{X}^+$ com n linhas e p_s colunas suplementares e, em linhas, por uma tabela $\underset{\sim}{X}_+$ com n_s linhas suplementares e p colunas.

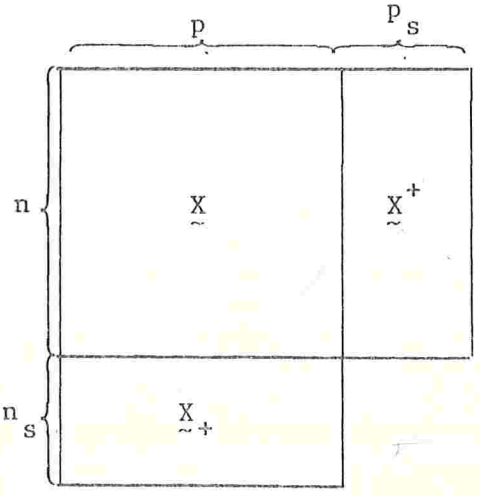


Figura 1.2 - Pontos Suplementares

As abscissas das p_s colunas suplementares sobre o eixo $\underset{\sim}{v}_\alpha$

são as p_s linhas do vetor $\underset{\sim}{X}^{+\prime}\underset{\sim}{v}_\alpha$ e as abscissas das n_s linhas suplementares são as n_s linhas do vetor $\underset{\sim}{X}_+\underset{\sim}{u}_\alpha$.

As variáveis ou indivíduos que realmente participam no cálculo de ajustamento são chamados elementos "ativos" e as variáveis ou indivíduos suplementares são chamados elementos "ilustrativos".

1.2 - ANÁLISE DE CORRESPONDÊNCIA BINÁRIA

1.2.1 - Construção das nuvens de pontos e escolha das distâncias

Vamos supor sempre que a tabela de dados tem n linhas e p colunas, $p \leq n$, e no cruzamento da linha i com a coluna j está o número x_{ij}, que representa o número de elementos pertencentes a linha i e a coluna j, onde i varia de 1 até n e j varia de 1 até p.

Consideremos o espaço $\mathbb{R}^p$: neste espaço temos n vetores cada um com p coordenadas (cada linha constitui um vetor de $\mathbb{R}^p$). Se as componentes dos vetores de $\mathbb{R}^p$ são os próprios valores x_{ij}, as proximidades entre os elementos podem ficar deturpadas pela falta de padronização dos dados. Não são os valores brutos que interessam na análise e sim os perfis das linhas, que serão dados pelas probabilidades condicionais do indivíduo aparecer na coluna j, dado que pertence a linha i, ou seja, x_{ij} deve ser dividido pelo total da linha i.

Se

$$x_{i\cdot} = \sum_{j=1}^{k} x_{ij}$$

(total da linha i) então tomamos como a j-ésima componente do i-ésimo vetor de $\mathbb{R}^p$ o valor

$$\frac{x_{ij}}{x_{i\cdot}}, \quad j = 1,2,\ldots,p.$$

Consideremos agora o espaço $\mathbb{R}^n$: temos p vetores cada um com n coordenadas (cada coluna constitui um vetor de $\mathbb{R}^n$).

Usando o mesmo raciocínio anterior, devemos usar na análise os perfis das colunas, que serão dados pelas probabilidades condicionais do indivíduo aparecer na linha i dado que pertence a coluna j, ou seja x_{ij} deve ser dividido pelo total da coluna j.

Se

$$x_{\cdot j} = \sum_{i=1}^{n} x_{ij}$$

(total da coluna j) então tomamos como i-ésima componente do j-ésimo vetor de $\mathbb{R}^n$ o valor

$$\frac{x_{ij}}{x_{\cdot j}}, \quad i = 1,\ldots,n.$$

Desenvolvemos a análise usando as "freqüências relativas":

$$f_{ij} = \frac{x_{ij}}{x}, \text{ onde } x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \text{ (total da tabela).}$$

$$f_{i\cdot} = \sum_{j=1}^{p} f_{ij} = \frac{x_{i\cdot}}{x}$$

$$f_{\cdot j} = \sum_{i=1}^{n} f_{ij} = \frac{x_{\cdot j}}{x}$$

e temos que

$$\sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{n} f_{i\cdot} = \sum_{j=1}^{p} f_{\cdot j} = 1$$

e

$$\frac{f_{ij}}{f_{\cdot j}} = \frac{x_{ij}}{x_{\cdot j}} \quad \text{e} \quad \frac{f_{ij}}{f_{i\cdot}} = \frac{x_{ij}}{x_{i\cdot}}, \quad i = 1,\ldots,n; \; j = 1,\ldots,p.$$

Uma vez definida a construção das nuvens de pontos, passamos à escolha da distância entre os pontos.

Em $\mathbb{R}^p$ a distância euclidiana clássica entre dois pontos-linhas i e i' será:

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 .$$

Mas se um dos valores, suponhamos o valor na coluna j_0, for muito grande em relação aos outros, o termo na distância relativo a esta coluna

$$\left(\frac{f_{ij_0}}{f_{i\cdot}} - \frac{f_{i'j_0}}{f_{i'\cdot}} \right)^2$$

será muito grande em relação aos outros e poderá deteriorar os resultados.

Usemos então uma expressão ponderada para a distância entre dois pontos:

em $\mathbb{R}^p$ definimos: $d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$;

$d^2(i,i')$ = distância entre os pontos-linhas i e i';

em $\mathbb{R}^n$ definimos: $d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j}} \right)^2$;

$d^2(j,j')$ = distância entre os pontos-colunas j e j'.

Esta distância escolhida verifica uma propriedade chamada "equivalência distribucional", que se exprime assim:

"Juntando dois pontos-linhas (pontos-colunas) com perfis idênticos, as distâncias entre os pontos-colunas (pontos-linhas) não se alteram".

Esta propriedade é importante pois garante uma invariância nos resultados em relação à codificação escolhida para construir as classes das variáveis.

Do ponto de vista técnico, é lógico que dois pontos confundidos no espaço, por estarem muito próximos, possam ser considerados como um só ponto correspondendo a um valor total igual a soma dos valores dos dois pontos. Assim, juntar pontos ou subdividir em mais pontos não provoca perda de informação.

1.2.2 - Ajustamento das nuvens de pontos

A escolha dos perfis de linhas e colunas como coordenadas dos pontos no espaço $\mathbb{R}^p$ e $\mathbb{R}^n$ dão a todos os pontos-linhas e pontos-colunas a mesma importância.

É natural que cada ponto tenha um peso proporcional a sua freqüência, para que não haja uma falsa idéia da repartição real da população. No ajustamento das nuvens de pontos num subespaço vetorial, a quantidade a ser maximizada será uma soma de quadrados ponderada por esses pesos. O peso do ponto i de

$\mathbb{R}^p$ é $f_{i\cdot}$ e peso do ponto j de $\mathbb{R}^n$ é $f_{\cdot j}$.

Resumindo temos:

Em $\mathbb{R}^p$: n pontos-linhas, cada um com p coordenadas do tipo

$$\frac{f_{ij}}{f_{i\cdot}},$$

para $j = 1,\ldots,p$, onde $f_{i\cdot}$ é o peso do ponto-linha i e a distância entre dois pontos-linhas i e i' é definida por:

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 .$$

Em $\mathbb{R}^n$: p pontos-colunas, cada um com n coordenadas do tipo

$$\frac{f_{ij}}{f_{\cdot j}},$$

para $i = 1,\ldots,n$, onde $f_{\cdot j}$ é o peso do ponto-coluna j e a distância entre dois pontos-colunas j e j' é definida por:

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 .$$

Usando notação matricial, designamos por $\underset{\sim}{X}$ a matriz $n \times p$ dos dados e por $\underset{\sim}{F}$ a matriz $n \times p$ das freqüências relativas. Segue a relação

$$\underset{\sim}{F} = \frac{1}{x}\,\underset{\sim}{X} \quad \text{onde} \quad x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}.$$

Chamamos de $\underset{\sim}{D}_p$ a matriz diagonal $p \times p$ cujo j-ésimo elemento da diagonal vale $f_{\cdot j}$, $j = 1,\ldots,p$ e de $\underset{\sim}{D}_n$ a matriz diagonal $n \times n$ cujo i-ésimo elemento da diagonal vale $f_{i\cdot}$, $i = 1,\ldots,n$. Os demais elementos de $\underset{\sim}{D}_p$ e $\underset{\sim}{D}_n$ são nulos.

Então podemos escrever matricialmente:

Em $\mathbb{R}^p$ temos n pontos-linhas cujas p coordenadas são dadas pelas n linhas de matriz $\underset{\sim}{D}_n^{-1}\underset{\sim}{F}$ de ordem n×p e a distância entre pontos-linhas é uma forma quadrática associada a matriz D_p^{-1} e o critério de ajustamento é baseado numa forma quadrática cuja matriz associada é $\underset{\sim}{D}_n$. Em $\mathbb{R}^n$ temos p pontos-colunas cujas n coordenadas são dadas pelas p colunas da matriz $\underset{\sim}{\underset{\sim}{F}}\underset{\sim}{\underset{\sim}{D}}_p^{-1}$ de ordem n×p e a distância entre pontos-colunas é uma forma quadrática associada a matriz D_n^{-1} e o critério de ajustamento é baseado numa forma quadrática cuja matriz associada é $\underset{\sim}{D}_p$.

1.2.3 - Determinação dos eixos fatoriais e dos fatores da análise

O desenvolvimento da análise é baseado na técnica da análise geral, em relação a origem dos eixos.

Análise em $\mathbb{R}^p$

As coordenadas dos n pontos-linhas de $\mathbb{R}^p$ são dados pela matriz $\underset{\sim}{D}_n^{-1}\underset{\sim}{F}$. Um vetor unitário $\underset{\sim}{u}$ de $\mathbb{R}^p$ verifica a relação $\underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{u}=1$ pois $\underset{\sim}{D}_p^{-1}$ é a matriz simétrica positiva definida que define a métrica em $\mathbb{R}^p$ (está associada à distância em $\mathbb{R}^p$). As n projeções dos pontos-linhas sobre o eixo $\underset{\sim}{u}$ são as n linhas do vetor $\underset{\sim}{D}_n^{-1}\underset{\sim}{\underset{\sim}{F}}\underset{\sim}{\underset{\sim}{D}}_p^{-1}\underset{\sim}{u}$ e a soma de quadrados ponderados dessas projeções, que é a quantidade a ser maximizada, é:

$$(\underset{\sim}{D}_n^{-1}\underset{\sim}{\underset{\sim}{F}}\underset{\sim}{\underset{\sim}{D}}_n^{-1}\underset{\sim}{u})'\underset{\sim}{D}_n(\underset{\sim}{D}_n^{-1}\underset{\sim}{\underset{\sim}{F}}\underset{\sim}{\underset{\sim}{D}}_p^{-1}\underset{\sim}{u}) = \underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{\underset{\sim}{F}}\underset{\sim}{\underset{\sim}{D}}_p^{-1}\underset{\sim}{u} \qquad (1.8)$$

Então devemos maximizar (1.8) com a condição $\underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = 1$. Usando o que foi visto na análise geral, $\underset{\sim}{u}$ é v.c. da matriz

$$\underset{\sim}{S} = \underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}$$

correspondente à maior r.c. λ: $\underset{\sim}{S}\underset{\sim}{u} = \lambda\underset{\sim}{u}$.

OBSERVAÇÃO - Usando a notação da análise geral temos que neste caso $\underset{\sim}{M} = \underset{\sim}{D}_p^{-1}$ e $\underset{\sim}{A} = \underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}$ e portanto $\underset{\sim}{u}$ é v.c. da matriz $\underset{\sim}{M}^{-1}\underset{\sim}{A}$, ou seja:

$$\underset{\sim}{M}^{-1}\underset{\sim}{A} = \underset{\sim}{D}_p(\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}) = \underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1} = \underset{\sim}{S}.$$

A matriz $\underset{\sim}{S}$ é de ordem p×p e seu termo geral $s_{jj'}$, se escreve:

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij}f_{ij'}}{f_{i.}f_{.j'}}$$

e portanto $\underset{\sim}{S}$ não é simétrica.

O vetor $\underset{\sim}{u}$ é o primeiro eixo fatorial encontrado e o primeiro fator é dado pelo vetor $\underset{\sim}{h} = \underset{\sim}{D}_p^{-1}\underset{\sim}{u}$ e $\underset{\sim}{h}$ é v.c. da matriz $\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}$. Logo as n projeções dos n pontos-linhas que são as componentes de $\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}$ são também as n componentes de $D_n^{-1}\underset{\sim}{F}\underset{\sim}{h}$.

Dizemos, então, que se $\underset{\sim}{u}_\alpha$ é o α-ésimo v.c. de $\underset{\sim}{S}$ correspondente a r.c. λ_α, $\underset{\sim}{u}_\alpha$ é o α-ésimo eixo fatorial e $\underset{\sim}{h}_\alpha = \underset{\sim}{D}_p^{-1}\underset{\sim}{u}_\alpha$ é o α-ésimo fator, sendo as n projeções dos n pontos-linhas dadas pelas componentes do vetor $\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{h}_\alpha$.

Análise em $\mathbb{R}^n$:

Como existe uma completa simetria entre os índices i e j, as demonstrações em $\mathbb{R}^n$ são deduzidas fazendo-se permutações entre os índices i e j (linhas e colunas), ou seja, transpondo

$\underset{\sim}{F}$ e substituindo $\underset{\sim}{D}_p$ por $\underset{\sim}{D}_n$.

Então se $\underset{\sim}{v}$ é um vetor unitário de $\mathbb{R}^n$ que verifica a relação $\underset{\sim}{v}\underset{\sim}{D}_n^{-1}\underset{\sim}{v} = 1$, por analogia com a análise em $\mathbb{R}^p$, fica fácil de ver que a quantidade a maximizar em $\mathbb{R}^n$ é: $\underset{\sim}{v}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v}$ com a condição $\underset{\sim}{v}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v} = 1$. Usando, novamente, o que foi visto na analise geral, $\underset{\sim}{v}$ é v.c. da matriz $\underset{\sim}{S}_1 = \underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}$, ou seja, $\underset{\sim}{v}$ é o primeiro eixo fatorial correspondente a maior r.c. de

$$\underset{\sim}{S}_1 = \underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}$$

e $\underset{\sim}{w} = \underset{\sim}{D}_n^{-1}\underset{\sim}{v}$ é o primeiro fator da análise. As p projeções dos p pontos-colunas são as componentes de $\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v}$ ou de $\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{w}$.

RELAÇÃO ENTRE OS FATORES DE $\mathbb{R}^p$ E $\mathbb{R}^n$

O eixo fatorial $\underset{\sim}{u}$ de $\mathbb{R}^p$ satisfaz a equação: $\underset{\sim}{S}\underset{\sim}{u} = \lambda\underset{\sim}{u}$, ou seja,

$$\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = \lambda\underset{\sim}{u}.$$

Pré-multiplicando os dois membros desta expressão por $\underset{\sim}{F}\underset{\sim}{D}_p^{-1}$ vem:

$$\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}\ \underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = \lambda\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u} \Rightarrow \underset{\sim}{S}_1(\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}) = \lambda(\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}),$$

ou seja, $\underset{\sim}{v}$ é proporcional a $\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}$.

A $\underset{\sim}{D}_n^{-1}$-norma de $\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}$ vale

$$(\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u})'\underset{\sim}{D}_n^{-1}(\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}) = \underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = \underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{S}\underset{\sim}{u} = \lambda$$

(pois, $\underset{\sim}{S}\underset{\sim}{u} = \lambda\underset{\sim}{u} \Rightarrow \underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{S}\underset{\sim}{u} = \lambda\underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = \lambda$ pois $\underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = 1$).

Como devemos ter $\underset{\sim}{v}$ unitário para a métrica de $\mathbb{R}^n$ ($\underset{\sim}{v}\underset{\sim}{D}^{-1}\underset{\sim}{v} = 1$) segue a relação:

$$\underset{\sim}{v} = \frac{1}{\sqrt{\lambda}} \underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u}. \qquad (1.9)$$

Analogamente, o eixo fatorial $\underset{\sim}{v}$ de $\mathbb{R}^n$ satisfaz a equação $\underset{\sim}{S}_1\underset{\sim}{v} = \lambda\underset{\sim}{v}$, ou seja, $\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v} = \lambda\underset{\sim}{v}$. Pré-multiplicando os dois membros por $\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}$, vem:

$$\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v} = \lambda\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v} \Rightarrow \underset{\sim}{S}_1\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v} = \lambda\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v},$$

ou seja, $\underset{\sim}{u}$ é proporcional a $\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v}$ e a $\underset{\sim}{D}_p^{-1}$-norma de $\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v}$ vale λ. Como $\underset{\sim}{u}$ deve ser unitário para a métrica de $\mathbb{R}^p$ ($\underset{\sim}{u}'\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = 1$) segue a relação:

$$\underset{\sim}{u} = \frac{1}{\sqrt{\lambda}}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v}. \qquad (1.10)$$

As relações (1.9) e (1.10) nos permitem escrever:

$$\underset{\sim}{h} = \underset{\sim}{D}_p^{-1}\underset{\sim}{u} = \frac{1}{\sqrt{\lambda}}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{v} = \frac{1}{\sqrt{\lambda}}\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{w} \qquad (1.11)$$

e

$$\underset{\sim}{w} = \underset{\sim}{D}_p^{-1}\underset{\sim}{v} = \frac{1}{\sqrt{\lambda}}\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{D}_p^{-1}\underset{\sim}{u} = \frac{1}{\sqrt{\lambda}}\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{h}. \qquad (1.12)$$

As relações entre os fatores $\underset{\sim}{h}$ e $\underset{\sim}{w}$ dadas por (1.11) e (1.12) mostram que as coordenadas dos pontos sobre um eixo fatorial num espaço (dadas por $\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{w}$ em $\mathbb{R}^n$ e por $\underset{\sim}{D}_n^{-1}\underset{\sim}{F}\underset{\sim}{h}$ em $\mathbb{R}^p$) são proporcionais às componentes dos fatores do outro espaço (dadas por $\underset{\sim}{h}$ em $\mathbb{R}^p$ e por $\underset{\sim}{w}$ em $\mathbb{R}^n$) correspondentes às mesmas r.c. As relações (1.11) e (1.12) são válidas para o caso geral, $\underset{\sim}{h}_\alpha$ e $\underset{\sim}{w}_\alpha$ correspondentes a r.c. λ_α.

As coordenadas dos n pontos-linhas de $\mathbb{R}^p$ são dadas, usando (1.12), pelas n componentes do vetor $\underset{\sim}{w}\sqrt{\lambda}$; usando (1.11) as

p coordenadas dos p pontos-colunas de $\mathbb{R}^n$ são dadas pela p componentes do vetor $\underset{\sim}{h}\sqrt{\lambda}$.

Esplicitamente, podemos escrever as coordenadas dos pontos como:

$$\begin{cases} \sqrt{\lambda}\, w_{\alpha i} = \sum_{j=1}^{p} f_{ij}/f_{i\cdot} \cdot h_{\alpha j} & (1.13) \\ \text{(projeção do ponto-linha i no eixo } \alpha) & \\ \sqrt{\lambda}\, h_{\alpha j} = \sum_{i=1}^{n} f_{ij}/f_{\cdot j} \cdot w_{\alpha i} & (1.14) \\ \text{(projeção do ponto-coluna j no eixo } \alpha) & \end{cases}$$

1.2.4 - Elementos Suplementares

A tabela $\underset{\sim}{X}$ pode ser completada por n_s linhas e/ou p_s colunas suplementares (ver Figura 1.2).

Suponhamos que temos n_s linhas suplementares. Os perfis das n_s linhas suplementares são:

Seja X_{+ij} = a j-ésima componente de i-ésima linha suplementar e

$$X_{+i\cdot} = \sum_{j=1}^{p} X_{+ij}$$

O perfil da linha i é o vetor com p coordenadas cuja j-ésima componente vale:

$$X_{+ij}/X_{+i\cdot}.$$

A projeção da linha i sobre o eixo α é

$$w_{\alpha i}\sqrt{\lambda} = \sum_{j=1}^{p} X_{+ij}/X_{+i\cdot} \cdot h_{\alpha j}.$$

Analogamente, deduz-se que o perfil de um ponto-coluna suplementar j é dado pelo vetor cuja i-ésima componente vale $X^+_{ij}/X^+_{.j}$ onde

$$X^+_{.j} = \sum_{i=1}^{n} X^+_{ij}$$

e a projeção do ponto j sobre o eixo α é:

$$h_{\alpha j}\sqrt{\lambda} = \sum_{i=1}^{n} (X^+_{.ij}/X^+_{.j}) \cdot w_{\alpha i} .$$

1.2.5 - Análise em relação ao centro de gravidade

Toda a análise desenvolvida até agora foi feita em relação à origem dos eixos. Mostramos nesta seção que a análise pode ser feita em relação ao centro de gravidade (ponto médio da nuvem) e que existe uma equivalência entre as análises em relação à origem dos eixos e em relação ao centro de gravidade.

Vamos mostrar a análise em $\mathbb{R}^p$ e a equivalência da mesma com a análise anteriormente desenvolvida.

Cada ponto-linha i de $\mathbb{R}^p$ tem por coordenadas:

$$\frac{f_{ij}}{f_{i.}}, \quad j = 1,2,\ldots,p.$$

Em $\mathbb{R}^p$ o centro de gravidade G é um ponto com p coordenadas, cada j-ésima coordenada sendo uma "média" dos elementos das linhas da j-ésima coluna, ou seja, a j-ésima componente de G vale

$$g_j = \sum_{i=1}^{n} f_{i.} \frac{f_{ij}}{f_{i.}} = \sum_{i=1}^{n} f_{ij} = f_{.j}, \quad j = 1,\ldots,p.$$

Então, na análise em relação ao centro de gravidade, as coordenadas dos pontos-linhas sofrem uma alteração, passando de

$$\frac{f_{ij}}{f_{i\cdot}}$$

(análise em relação a origem) para

$$\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} = \frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot}}$$

(análise em relação a G).

Como

$$\sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} = 1,$$

para todo i, na verdade a nuvem de pontos-linhas está num sub-espaço com p-1 dimensões, sub-espaço este, que contém o centro de gravidade G e os eixos fatoriais da análise em relação à G.

Na análise em relação à G, a matriz $\underset{\sim}{S}$, definida anteriormente com termo geral

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i\cdot} f_{\cdot j'}},$$

é alterada para uma matriz $\underset{\sim}{S}^*$ cujo termo geral $s^*_{jj'}$, fica:

$$s^*_{jj'} = \sum_{i=1}^{n} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})(f_{ij'} - f_{i\cdot} f_{\cdot j'})}{f_{i\cdot} f_{\cdot j'}}$$

Desenvolvendo, vem:

$$s^*_{jj'} = \sum_{i=1}^{n} \frac{f_{ij}f_{ij'}}{f_{i\cdot}f_{\cdot j'}} - \sum_{i=1}^{n} \frac{f_{ij}f_{i\cdot}f_{\cdot j'}}{f_{i\cdot}f_{\cdot j'}} - \sum_{i=1}^{n} \frac{f_{ij'}f_{i\cdot}f_{\cdot j}}{f_{i\cdot}f_{\cdot j}} +$$

$$+ \sum_{i=1}^{n} \frac{f_{i\cdot}^2 f_{\cdot j}f_{\cdot j'}}{f_{i\cdot}f_{\cdot j'}} = \sum_{i=1}^{n} \frac{f_{ij}f_{ij'}}{f_{i\cdot}f_{\cdot j'}} - \sum_{i=1}^{n} f_{ij'} - \frac{f_{\cdot j}}{f_{\cdot j'}} \sum_{i=1}^{n} f_{ij'} +$$

$$+ f_{\cdot j} \sum_{i=1}^{n} f_{i\cdot} = \sum_{i=1}^{n} \frac{f_{ij}f_{ij'}}{f_{i\cdot}f_{\cdot j'}} - f_{\cdot j} - f_{\cdot j} + f_{\cdot j} =$$

$$= \sum_{i=1}^{n} \frac{f_{ij}f_{ij'}}{f_{i\cdot}f_{\cdot j'}} - f_{\cdot j}.$$

Logo,

$$s^*_{jj'} = s_{jj'} - f_{\cdot j}. \qquad (1.15)$$

Se $\underset{\sim}{u}^*$ é um eixo fatorial da análise em relação a G, então

$$\underset{\sim}{S}^*\underset{\sim}{u}^* = \lambda^*\underset{\sim}{u}^*$$

ou seja,

$$\sum_{j'=1}^{p} s^*_{jj'}u^*_{j'} = \lambda^* u^*_j. \qquad (1.16)$$

Usando (1.15), vem:

$$\sum_{j'=1}^{p} s_{jj'}u^*_{j'} - f_{\cdot j} \sum_{j'=1}^{p} u^*_{j'} = \lambda^* u^*_j.$$

Mas $\underset{\sim}{u}^*$ pertence ao sub-espaço com p-1 dimensões centrado em relação a G, e então a soma de suas componentes é nula:

$$\sum_{j=1}^{p} u^*_{j'} = 0.$$

Segue então que:

$$\sum_{j'=1}^{p} s_{jj'} u^*_{j'} = \lambda^* u^*_j \quad \text{ou} \quad \underset{\sim}{S}\underset{\sim}{u}^* = \lambda^* \underset{\sim}{u}^* \qquad (1.17)$$

As relações (1.16) e (1.17) mostram que todo v.c. de $\underset{\sim}{S}^*$ é v.c. de $\underset{\sim}{S}$ relativo a mesma r.c. A reta que liga a origem dos eixos ao centro de gravidade é um v.c. de $\underset{\sim}{S}$ relativo à r.c. 1 (um) e, é v.c. de $\underset{\sim}{S}^*$ relativo à r.c. 0 (zero). Como este v.c. "explica" uma dispersão nula, não deve ser considerado na análise (a taxa de inércia ou porcentagem de dispersão do fator α é dada pelo coeficiente $\frac{\lambda_\alpha}{\Sigma\lambda_\alpha}$ onde λ_α é a α-ésima r.c.).

Para fazer a análise em relação a G basta, então, diagonalizar a matriz $\underset{\sim}{S}$, da análise em relação a origem dos eixos, e descartar a r.c. igual a 1 (um) e os respectivos eixos em $\mathbb{R}^p$ e $\mathbb{R}^n$.

OBSERVAÇÃO - A matriz $\underset{\sim}{S}$ é não simétrica, o que dificulta os cálculos. Se ao invés de definirmos as coordenadas dos pontos-linhas como $\frac{f_{ij}}{f_{i\cdot}}$, $j = 1,\ldots,p$ e as coordenadas dos pontos-colunas como

$$\frac{f_{ij}}{f_{\cdot j}}, \quad i = 1,\ldots,n$$

definimos estas coordenadas como

$$\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} \quad \text{e} \quad \frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}}$$

respectivamente, recaímos numa matriz $\underset{\sim}{S}_1$ a ser diagonalizada, simétrica, cujo termo geral é:

$$s_{1jj'} = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \frac{f_{ij} \cdot f_{ij'}}{\sqrt{f_{\cdot j} f_{\cdot j'}}}.$$

As distâncias definidas na Seção 1.2.1, com estas novas coordenadas, ficam definidas como distâncias euclidianas usuais:

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot}\sqrt{f_{\cdot j}}} \right)^2$$

e

$$d^2(j,j') = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{\cdot j}\sqrt{f_{i\cdot}}} - \frac{f_{ij'}}{f_{\cdot j'}\sqrt{f_{i\cdot}}} \right)^2$$

A análise usando estas novas coordenadas está desenvolvida no livro de Lebart e Fenelon (1971).

1.2.6 - Interpretação dos resultados

A interpretação dos gráficos representando as nuvens de pontos nos planos de projeção formados pelos primeiros eixos fatoriais tomados dois a dois é o principal e mais difícil objetivo da análise. Temos que verificar quais são os fatores da análise mais representativos e interpretar as proximidades entre elementos de uma mesma nuvem. Salvo em casos particulares é perigoso interpretar a proximidade entre pontos de nuvens diferentes (um de $\mathbb{R}^p$ e outro de $\mathbb{R}^n$).

Definimos a seguir três coeficientes que auxiliam na interpretação dos resultados: taxa de inércia, contribuição absoluta e contribuição relativa.

a) Taxa de Inércia

Taxa de inércia ou porcentagem de variância de um eixo fatorial α é definida pelo quociente entre a α-ésima r.c. e a soma total das r.c. ou seja:

$$I_\alpha = \frac{\lambda_\alpha}{\sum_\alpha \lambda_\alpha}.$$

$100 \cdot \tau_\alpha$ indicada a porcentagem de variação explicada pelo eixo α.

Como será visto no Capítulo 3 a taxa de inércia pode dar uma idéia pessimista da parte de informação dada pelo fator.

b) Contribuição Absoluta

A contribuição absoluta exprime a parte tomada por uma dada variável na inércia (variância) explicada por um fator, ou seja, permite saber quais são as variáveis realmente responsáveis na construção do fator em questão.

Temos que

$$\underset{\sim}{u}'_\alpha \underset{\sim}{D}_p^{-1} \underset{\sim}{u}_\alpha = 1 \quad \text{e} \quad \underset{\sim}{h}_\alpha = \underset{\sim}{D}_p^{-1} \underset{\sim}{u}_\alpha$$

ou seja, $\underset{\sim}{u}_\alpha = \underset{\sim}{D}_p \underset{\sim}{h}_\alpha$, segue então que $\underset{\sim}{h}'_\alpha \underset{\sim}{D}_p \underset{\sim}{h}_\alpha = 1$, ou seja,

$$\sum_{j=1}^{p} f_{\cdot j} h_{\alpha j}^2 = 1. \qquad (1.18)$$

A projeção do ponto j de $\mathbb{R}^n$ no eixo α vale:

$$\sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}} w_{\alpha i} = \sqrt{\lambda_\alpha}\, h_{\alpha j}.$$

A inércia com relação a G da nuvem projetada sobre o eixo α vale então:

$$\sum_{j=1}^{p} f_{\cdot j} (\sqrt{\lambda_\alpha} h_{\alpha j})^2 = \lambda_\alpha$$

(usando a relação (1.18)).

O quociente

$$\frac{\lambda_\alpha \cdot h_{\alpha j}^2 \cdot f_{\cdot j}}{\lambda_\alpha} = f_{\cdot j} h_{\alpha j}^2 = ca_\alpha(j)$$

define a contribuição absoluta da variável j ao eixo α. Segue que

$$\sum_{j=1}^{p} ca_\alpha(j) = 1.$$

Analogamente, a contribuição absoluta de um ponto i de $\mathbb{R}^p$ ao eixo α é definida por: $ca_\alpha(i) = f_{i} . w_{\alpha i}^2$ com

$$\sum_{j=1}^{n} ca_\alpha(i) = 1.$$

c) Contribuição Relativa

A contribuição relativa, ou correlação entre variável e fator, exprime a parte tomada por um fator na explicação da dispersão de uma variável. Exibe quais são as características exclusivas do fator.

Os eixos fatoriais formam em cada espaço uma base ortogonal. O quadrado da distância de uma variável j ao centro de gravidade G se decompõe na soma de quadrados da projeção do

ponto sobre cada um dos eixos.

Em $\mathbb{R}^n$, o quadrado da projeção do ponto j sobre o eixo α vale: $d_\alpha^2(j,G) = (\sqrt{\lambda_\alpha} h_{\alpha j})^2$ e o quadrado das distâncias da variável j ao centro de gravidade vale:

$$d^2(j,G) = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 .$$

Então, $\sum_\alpha d_\alpha^2(j,G) = d^2(j,G)$.

A contribuição relativa do fator α na explicação da variância da variável j é então definida como

$$cr_\alpha(j) = \frac{d_\alpha^2(j,G)}{d^2(j,G)}$$

e temos que $\sum_\alpha cr_\alpha(j) = 1$.

Analogamente, em $\mathbb{R}^p$ a contribuição relativa do fator α na explicação da variância da variável i é:

$$cr_\alpha(i) = \frac{d_\alpha^2(i,G)}{d^2(i,G)}$$

onde

$$d_\alpha^2(i,G) = (\sqrt{\lambda_\alpha} w_{\alpha i})^2 \quad \text{e} \quad d^2(i,G) = \sum_{i=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 .$$

A seguir, apresentamos uma generalização da análise de correspondência binária.

CAPÍTULO 2

ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA

A análise de correspondência múltipla é uma extensão do caso binário, visto anteriormente, quando mais de dois conjuntos de variáveis são postos em correspondência.

2.1 - APRESENTAÇÃO DE DADOS

Na análise múltipla temos uma série de variáveis cada uma com um certo número de respostas possíveis (modalidades de respostas) e, estas variáveis são respondidas por um conjunto de indivíduos. No caso em que temos somente duas variáveis e estas são cruzadas formando uma tabela onde as linhas correspondem às modalidades de uma das variáveis e as colunas correspondem às modalidades da outra variável recaimos na análise de correspondência binária. Mostramos aqui o caso em que o número de variáveis é maior do que dois.

A notação adotada é a seguinte:

Q é o conjunto de variáveis;

q é uma variável, $q = 1,2,\ldots,Q$.

J_q é o número de modalidades de respostas da variável q, $q = 1,\ldots,Q$.

J é o número total de modalidades de resposta do conjunto Q, ou seja,

$$J = \sum_{q=1}^{Q} J_q .$$

H é o conjunto-produto, ou seja, o conjunto cujos elementos são as séries de Q modalidades, cada modalidade de uma variável diferente. É o conjunto de respostas possíveis considerando-se as Q variáveis. Se considerarmos a tabela cruzando todas as variáveis, cada casela será um elemento de H.

$\underset{\sim}{Z}$ é a tabela cujas n linhas são os n indivíduos que respondem as variáveis e as J colunas são as modalidades de respostas de todas as variáveis. $\underset{\sim}{Z}$ é formada por Q sub-tabelas

$$\underset{\sim}{Z} = [\underset{\sim}{Z}_1, \underset{\sim}{Z}_2, \ldots, \underset{\sim}{Z}_Q]$$

onde $\underset{\sim}{Z}_q$ é a tabela com n linhas e J_q colunas correspondentes às modalidades da variável q.

É muito comum que as respostas às modalidades apareçam na forma de zeros ou uns, isto é, para cada variável o indivíduo escolhe uma modalidade de resposta e esta recebe o valor 1 (um) e as demais modalidades da variável recebem o valor 0 (zero). Este tipo de codificação chamamos de codificação "disjuntiva completa": as modalidades de respostas se excluem mutuamente e somente uma das modalidades é escolhida.

A tabela $\underset{\sim}{Z}$ é tabelada na forma disjuntiva completa, ou seja, para cada variável q com J_q modalidades, J_q-1 modalidades têm resposta "zero" e uma modalidade tem resposta "um".

Então cada linha de $\underset{\sim}{Z}_q$, q = 1,...,Q, contém apenas um valor "um" e J_q-1 zeros, fazendo com que a soma de cada uma das linhas da tabela $\underset{\sim}{Z}$ seja sempre Q.

A matriz $\underset{\sim}{B} = \underset{\sim}{Z}'\underset{\sim}{Z}$ é chamada "Tabela de Contingência de Burt" associada à tabela $\underset{\sim}{Z}$. $\underset{\sim}{B}$ é uma matriz quadrada de ordem J × J e é formada por Q^2 blocos, como mostra a Figura 2.1.

$$\underset{\sim}{B} = \begin{array}{c} \\ \underset{\sim}{Z}_1 \\ \underset{\sim}{Z}_2 \\ \vdots \\ \underset{\sim}{Z}_Q \end{array} \begin{array}{c} \overbrace{}^{Z_1} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overbrace{}^{Z_2} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \cdots \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overbrace{}^{Z_Q} \\ \\ \\ \\ \\ \end{array} \left[\begin{array}{ccccc} B11 & B12 & \cdots & B1Q \\ B1Q+1 & B1Q+2 & \cdots & \\ \vdots & \vdots & & \\ & & \cdots & B1Q^2 \end{array}\right]$$

Figura 2.1 - Matriz de Burt

Os blocos da diagonal, ou seja, os blocos formados por $Z_q'Z_q$, q = 1,...,Q são matrizes diagonais $J_q \times J_q$ e os elementos da diagonal de cada uma destas matrizes indicam quantas pessoas responderam a cada modalidade da variável q, q=1,...,Q.

Os demais blocos de $\underset{\sim}{B}$ são formados pelo cruzamento de duas variáveis diferentes q e q', ou seja, cada bloco $\underset{\sim}{Z}_q' \underset{\sim}{Z}_{q'}$, de ordem $J_q \times J_{q'}$, é uma tabela de contingência das respostas das variáveis q e q' cruzadas.

Vamos definir também a matriz $\underset{\sim}{D}$ de ordem J×J, diagonal, cujos elementos da diagonal são os mesmos elementos da diagonal de $\underset{\sim}{B}$ e os demais elementos são nulos.

Vamos mostrar a seguir, como fica a análise de correspon-

dência dessas tabelas para o caso em que temos somente duas variáveis e depois faremos a generalização para o caso de mais de duas variáveis.

2.2 - ANÁLISE NO CASO PARTICULAR DE DUAS VARIÁVEIS

No caso em que temos somente duas variáveis a tabela $\underset{\sim}{Z}$ fica: $\underset{\sim}{Z} = [\underset{\sim}{Z}_1 \ \underset{\sim}{Z}_2]$ e $\underset{\sim}{Z}$ tem n linhas e $J = J_1 + J_2$ colunas.

Vamos mostrar que existe uma equivalência entre as análises de correspondência das tabelas $\underset{\sim}{Z}$ de ordem $n \times J$, $\underset{\sim}{B}$ de ordem $J \times J$ e $\underset{\sim}{Z}_1' \ \underset{\sim}{Z}_2$ de ordem $J_1 \times J_2$.

a) Equivalência da análise nas tabelas $\underset{\sim}{Z}$ e $\underset{\sim}{B}$

Como estamos num caso de análise de correspondência binária (só duas variáveis), vamos fazer a ligação entre a notação na Seção 1.2 e a notação usada agora em 2.1. A matriz $\underset{\sim}{F}$, definida em 1.2, em função de $\underset{\sim}{Z}$ se escreve: $\underset{\sim}{F} = \frac{1}{nQ} \underset{\sim}{Z}$ pois nQ é o total da tabela $\underset{\sim}{Z}$ e $\underset{\sim}{F}$ é matriz de freqüências. A matriz $\underset{\sim}{D}_p$ em função de $\underset{\sim}{D}$ se escreve: $\underset{\sim}{D}_p = \frac{1}{nQ} \underset{\sim}{D}$ e $\underset{\sim}{D}_n$ se escreve $\underset{\sim}{D}_n = \frac{1}{n} \underset{\sim}{I}_n$ onde $\underset{\sim}{I}_n$ é a matriz identidade $n \times n$.

Na análise binária em $\mathbb{R}^p$ vimos que α-ésimo fator $\underset{\sim}{h}_\alpha$ é v. c. da matriz $\underset{\sim}{D}_p^{-1} \underset{\sim}{F}' \underset{\sim}{D}_n^{-1} \underset{\sim}{F}$. Reescrevendo esta matriz em função de $\underset{\sim}{Z}$ e $\underset{\sim}{D}$ temos:

$$\underset{\sim}{D}_p^{-1} \underset{\sim}{F}' \underset{\sim}{D}_n^{-1} \underset{\sim}{F} = (nQ\underset{\sim}{D}^{-1}) \left(\frac{1}{nQ} \underset{\sim}{Z}'\right) (n\underset{\sim}{I}_n) \left(\frac{1}{nQ} \underset{\sim}{Z}\right) = \frac{1}{Q} \underset{\sim}{D}^{-1} \underset{\sim}{Z}' \underset{\sim}{Z}.$$

Logo o α-ésimo fator da análise da tabela $\underset{\sim}{Z}$ será v.c. de $\frac{1}{Q} \underset{\sim}{D}^{-1} \underset{\sim}{Z}' \underset{\sim}{Z}$. Então, se $\underset{\sim}{\Phi}_\alpha$ é este fator vale a relação:

$$\frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{Z}'\underset{\sim}{Z}\underset{\sim}{\Phi}_\alpha = \mu_\alpha \underset{\sim}{\Phi}_\alpha \qquad (2.1)$$

onde μ_α é a α-ésima r.c.

No caso da tabela $\underset{\sim}{B}$ temos que a matriz $\underset{\sim}{F}$ se escreve:

$$\underset{\sim}{F} = \frac{1}{nQ^2}\underset{\sim}{B}$$

pois o total da tabela $\underset{\sim}{B}$ vale nQ^2 e a matriz $\underset{\sim}{D}_p$ e $\underset{\sim}{D}_n$ se escrevem, em função de $\underset{\sim}{D}$: $\underset{\sim}{D}_p = \underset{\sim}{D}_n = \frac{1}{nQ}\underset{\sim}{D}$.

Então,

$$\underset{\sim}{D}_p^{-1}\underset{\sim}{F}'\underset{\sim}{D}_n^{-1}\underset{\sim}{F} = nQ\underset{\sim}{D}^{-1}\frac{1}{nQ^2}\underset{\sim}{B}'nQ\underset{\sim}{D}^{-1}\frac{1}{nQ^2}\underset{\sim}{B} = \frac{1}{Q^2}\underset{\sim}{D}^{-1}\underset{\sim}{B}'\underset{\sim}{D}^{-1}\underset{\sim}{B}.$$

Logo o α-ésimo fator da análise da tabela $\underset{\sim}{B}$ será v.c. de

$$\frac{1}{Q^2}\underset{\sim}{D}^{-1}\underset{\sim}{B}'\underset{\sim}{D}^{-1}\underset{\sim}{B} \quad \text{ou} \quad \frac{1}{Q^2}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{D}^{-1}\underset{\sim}{B}$$

pois $\underset{\sim}{B} = \underset{\sim}{Z}'\underset{\sim}{Z}$ é simétrico ($\underset{\sim}{B} = \underset{\sim}{B}'$).

Pré-multiplicando os dois membros da relação (2.1) por $\frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{B}$ e lembrando que $\underset{\sim}{Z}'\underset{\sim}{Z} = \underset{\sim}{B}$, vem:

$$\frac{1}{Q^2}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{\Phi}_\alpha = \mu_\alpha \underbrace{\frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{\Phi}_\alpha}_{\mu_\alpha\Phi_\alpha \text{ pela relação (2.1)}} \Rightarrow \frac{1}{Q^2}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{\Phi}_\alpha =$$

$$= \mu_\alpha^2\underset{\sim}{\Phi}_\alpha,$$

ou seja, Φ_α também é v.c. da matriz $\frac{1}{Q^2}\underset{\sim}{D}^{-1}\underset{\sim}{B}\underset{\sim}{D}^{-1}\underset{\sim}{B}$ só que em relação a r.c. μ_α^2. Logo Φ_α é fator da análise da tabela $\underset{\sim}{Z}$ e também é fator da análise da tabela $\underset{\sim}{B}$.

Generalizando, temos que os fatores da análise de $\underset{\sim}{Z}$ são

os mesmos da análise de $\underset{\sim}{B}$. Portanto, as análises de $\underset{\sim}{Z}$ e $\underset{\sim}{B}$ são equivalentes.

b) Equivalência da análise nas tabelas $\underset{\sim}{B}$ e $\underset{\sim}{Z}_1' \underset{\sim}{Z}_2$

Como $\underset{\sim}{Z}_1' \underset{\sim}{Z}_2$ é uma tabela cruzada do mesmo tipo que as tabelas da Seção 1.2, $\underset{\sim}{Z}_1' \underset{\sim}{Z}_2$ com J_1 linhas e J_2 colunas, vamos dizer, que $\underset{\sim}{h}_\alpha$ é o α-ésimo fator extraído da análise em $\mathbb{R}^{J_2}$ e $\underset{\sim}{w}_\alpha$ é o α-ésimo fator extraído da análise em $\mathbb{R}^{j_1}$ relativos a r.c. α.

Vamos mostrar que para todo par de fatores $\underset{\sim}{h}_\alpha$ e $\underset{\sim}{w}_\alpha$ devido a análise de $\underset{\sim}{Z}_1' \underset{\sim}{Z}_2$ existe um fator $\underset{\sim}{\Phi}_\alpha$ da análise de $\underset{\sim}{Z}$ (onde $\underset{\sim}{B}$) correspondente tal que

$$\underset{\sim}{\Phi}_\alpha = \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix},$$

já que a dimensão da tabela $\underset{\sim}{Z}$ (ou tabela $\underset{\sim}{B}$) é $J = J_1 + J_2$ e a dimensão de $\underset{\sim}{h}_\alpha$ é J_2 e a de $\underset{\sim}{w}_\alpha$ é J_1.

Vamos decompor a matriz $\underset{\sim}{D}$ em $\underset{\sim}{D}_1 = \underset{\sim}{Z}_1'\underset{\sim}{Z}_1$ e $\underset{\sim}{D} = \underset{\sim}{Z}_2'\underset{\sim}{Z}_2$ tal que

$$\underset{\sim}{D} = \begin{bmatrix} \underset{\sim}{D}_1 & 0 \\ 0 & \underset{\sim}{D}_2 \end{bmatrix}.$$

Os elementos das diagonais de $\underset{\sim}{D}_1$ e $\underset{\sim}{D}_2$ são respectivamente os totais em linha e coluna da tabela $\underset{\sim}{Z}_1' \underset{\sim}{Z}_2$, ou seja, $\underset{\sim}{D}_1 = \underset{\sim}{D}_n$ e $\underset{\sim}{D}_2 = \underset{\sim}{D}_p$.

Usando as relações (1.11) e (1.12) da Seção 1.2.3, podemos escrever para o α-ésimo fator:

$$\begin{cases} \underset{\sim}{h}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \underset{\sim}{D}_2^{-1} \underset{\sim}{Z}_2' \underset{\sim}{Z}_1 \underset{\sim}{w}_\alpha \\ \text{e} \\ \underset{\sim}{w}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} \underset{\sim}{D}_1^{-1} \underset{\sim}{Z}_1' \underset{\sim}{Z}_2 \underset{\sim}{h}_\alpha \end{cases}$$

Podemos escrever este sistema como:

$$\begin{cases} \underset{\sim}{D}_2^{-1} (\underset{\sim}{D}_2 \underset{\sim}{h}_\alpha + \underset{\sim}{Z}_2' \underset{\sim}{Z}_1 \underset{\sim}{w}_\alpha) = (1+\sqrt{\lambda_\alpha}) \underset{\sim}{h}_\alpha \\ \text{e} \\ \underset{\sim}{D}_1^{-1} (\underset{\sim}{D}_1 \underset{\sim}{w}_\alpha + \underset{\sim}{Z}_1' \underset{\sim}{Z}_2 \underset{\sim}{h}_\alpha) = (1+\sqrt{\lambda_\alpha}) \underset{\sim}{w}_\alpha \end{cases}$$

ou ainda, na forma matricial:

$$\begin{bmatrix} \underset{\sim}{D}_1^{-1} & 0 \\ 0 & \underset{\sim}{D}_2^{-1} \end{bmatrix} \begin{bmatrix} \underset{\sim}{D}_1 & \underset{\sim}{Z}_1' \underset{\sim}{Z}_2 \\ \underset{\sim}{Z}_2' \underset{\sim}{Z}_1 & \underset{\sim}{D}_2 \end{bmatrix} \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix} = (1+\sqrt{\lambda_\alpha}) \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix}.$$

Então, temos:

$$\underset{\sim}{D}^{-1} \underset{\sim}{Z}' \underset{\sim}{Z} \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix} = (1+\sqrt{\lambda_\alpha}) \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix}.$$

Multiplicando os dois membros desta expressão por $\frac{1}{Q}$, que no caso é $\frac{1}{2}$ (Q=2), vem:

$$\frac{1}{2} \underset{\sim}{D}^{-1} \underset{\sim}{Z}' \underset{\sim}{Z} \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix} = \frac{1+\sqrt{\lambda_\alpha}}{2} \begin{bmatrix} \underset{\sim}{w}_\alpha \\ \underset{\sim}{h}_\alpha \end{bmatrix}$$

que é a expressão (2.1) no caso de $\mu_\alpha = \dfrac{1+\sqrt{\lambda_\alpha}}{2}$ e

$$\underset{\sim}{\Phi}_{\alpha} = \begin{bmatrix} \underset{\sim}{w}_{\alpha} \\ \\ \underset{\sim}{h}_{\alpha} \end{bmatrix}.$$

Em resumo, quando fazemos a análise da tabela $\underset{\sim}{Z}_1'\underset{\sim}{Z}_2$ vamos encontrar a α-iésima r.c. λ_α e os fatores correspondentes $\underset{\sim}{w}_{\alpha}$ e $\underset{\sim}{h}_{\alpha}$. Baseado nisto podemos encontrar a r.c.

$$\mu_\alpha = \frac{1+\sqrt{\lambda_\alpha}}{2} \text{ e o fator } \underset{\sim}{\Phi}_{\alpha} = \begin{bmatrix} \underset{\sim}{w}_{\alpha} \\ \\ \underset{\sim}{h}_{\alpha} \end{bmatrix}$$

da análise da tabela $\underset{\sim}{Z}$ e, a r.c. μ_α^2 é o mesmo fator $\underset{\sim}{\Phi}_{\alpha}$ da análise da tabela $\underset{\sim}{B}$.

Apesar das análises das três tabelas ($\underset{\sim}{Z}$, $\underset{\sim}{B}$ e $\underset{\sim}{Z}_1'\underset{\sim}{Z}_2$) serem equivalentes, produzindo resultados similares, as r.c. nas três análises são diferentes e consequentemente as taxas de inércia são diferentes. De um modo geral tabelas do tipo disjuntivas completa como a $\underset{\sim}{Z}$ dão taxas de inércia mais fracas que as das tabelas do tipo cruzado como a $\underset{\sim}{Z}_1'\underset{\sim}{Z}_2$. No caso da tabela $\underset{\sim}{Z}$, temos da relação (2.1) que $\frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{Z}'\underset{\sim}{Z}\underset{\sim}{\Phi}_{\alpha} = \mu_\alpha \underset{\sim}{\Phi}_{\alpha}$. Pela propriedade do traço de uma matriz, que diz que se $\underset{\sim}{A}\underset{\sim}{x}_{\alpha} = \lambda_\alpha \underset{\sim}{x}_{\alpha}$, onde λ_α é a r.c. de $\underset{\sim}{A}$ e $\underset{\sim}{x}_{\alpha}$ é o v.c. correspondente, então tr $A = \sum_\alpha \lambda_\alpha$, temos que a soma das r.c. não triviais (a r.c. trivial é $\lambda_1 = 1$) de $\underset{\sim}{Z}$ valem:

$$\sum_\alpha \lambda_\alpha = \text{tr } \frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{Z}'\underset{\sim}{Z}.$$

Como $\underset{\sim}{Z}'\underset{\sim}{Z}$ é igual a $\underset{\sim}{B}$ e $\underset{\sim}{D}$ é a matriz diagonal com os mesmos elementos da diagonal de $\underset{\sim}{B}$ temos que os elementos da dia-

gonal de $\underset{\sim}{D}^{-1}\underset{\sim}{B}$ são unitários. Então, o $\mathrm{tr}\,\frac{1}{Q}\,\underset{\sim}{D}^{-1}\underset{\sim}{Z}'\underset{\sim}{Z} = \frac{J_1+J_2}{Q}$. Neste caso particular, $Q = 2$, e portanto a soma das r.c. não triviais de $\underset{\sim}{Z}$ vale

$$\frac{J_1 + J_2}{2} - 1 = \frac{J_1 + J_2 - 2}{2}.$$

A taxa de inércia do α-iésimo eixo é definida por

$$\frac{\lambda_\alpha}{\frac{J_1+J_2-2}{2}} = \frac{2\lambda_\alpha}{J_1+J_2-2}$$

Como para qualquer α, λ_α é sempre menor ou igual a 1 (um), nenhum fator tem taxa de inércia superior, em porcentagem a $\frac{2\times 100}{J_1+J_2-2}$ na análise de uma tabela disjuntiva completa com duas variáveis com J_1 e J_2 modalidades respectivamente. Para este tipo de tabela, a taxa de inércia dá uma idéia muito pessimista da informação extraída pelos eixos fatoriais.

2.3 - ANÁLISE NO CASO GERAL

No caso geral com Q variáveis, $Q > 2$, a tabela $\underset{\sim}{Z}$ é da forma $\underset{\sim}{Z} = [\underset{\sim}{Z}_1, \underset{\sim}{Z}_2, \ldots, \underset{\sim}{Z}_Q]$ e possui $J = J_1+J_2+\ldots+J_Q$ colunas e n linhas. Cada coluna corresponde a um ponto de $\mathbb{R}^n$. Em $\mathbb{R}^n$ cada subtabela $\underset{\sim}{Z}_q$ é uma variedade linear V_q com J_q dimensões. Todas as Q variedades lineares têm em comum pelo menos a primeira bissetriz (a soma das J_q colunas de V_q dá uma coluna só de valores 1 (um)). O posto da tabela $\underset{\sim}{Z}$ é então no máximo igual a $J-(Q-1)$.

Seja $\underset{\sim}{\phi}_q$ o vetor cujas J_q componentes são as coordenadas de um ponto $\underset{\sim}{m}_q$ de V_q na base definida pelas colunas de $\underset{\sim}{Z}_q$. As coordenadas de $\underset{\sim}{m}_q$ em $\mathbb{R}^n$ são as n componentes de $\underset{\sim}{m}_q = \underset{\sim}{Z}_q \underset{\sim}{\phi}_q$. O quadrado da distância de $\underset{\sim}{m}_q$ à origem, usando a distância euclidiana usual, vale:

$$\underset{\sim}{\phi}'_q \underset{\sim}{Z}'_q \underset{\sim}{Z}_q \underset{\sim}{\phi}_q = \underset{\sim}{\phi}_q \underset{\sim}{D}_q \underset{\sim}{\phi}_q .$$

Fazer a análise de correspondência da tabela de contingência que cruza duas variáveis q e q' é o mesmo que estudar as posições respectivas das variedades lineares V_q e $V_{q'}$, ou seja, o mesmo que fazer a análise canônica da tabela $[Z_q / Z_{q'}]$. Das relações (1.11) e (1.12) da Seção 1.2.3 podemos escrever, omitindo o índice do eixo α para simplificar a notação, as seguintes relações:

$$\begin{cases} \underset{\sim}{\phi}_q = \dfrac{1}{\sqrt{\lambda}} \underset{\sim}{D}_q^{-1} \underset{\sim}{Z}'_q \underset{\sim}{Z}_{q'} \underset{\sim}{\phi}_{q'} \\ \\ \underset{\sim}{\phi}_{q'} = \dfrac{1}{\sqrt{\lambda}} \underset{\sim}{D}_{q'}^{-1} \underset{\sim}{Z}'_{q'} \underset{\sim}{Z}_q \underset{\sim}{\phi}_q \end{cases}$$

ou seja:

$$\begin{cases} \underset{\sim}{Z}_q \underset{\sim}{\phi}_q = \dfrac{1}{\sqrt{\lambda}} \underset{\sim}{Z}_q \underset{\sim}{D}_q^{-1} \underset{\sim}{Z}'_q \underset{\sim}{Z}_{q'} \underset{\sim}{\phi}_{q'} \\ \\ \underset{\sim}{Z}_{q'} \underset{\sim}{\phi}_{q'} = \dfrac{1}{\sqrt{\lambda}} \underset{\sim}{Z}_{q'} \underset{\sim}{D}_{q'}^{-1} \underset{\sim}{Z}'_{q'} \underset{\sim}{Z}_q \underset{\sim}{\phi}_q \end{cases} \Rightarrow \begin{cases} \underset{\sim}{m}_q = \dfrac{1}{\sqrt{\lambda}} \underset{\sim}{P}_q \underset{\sim}{m}_{q'} \quad (2.2) \\ \\ \underset{\sim}{m}_{q'} = \dfrac{1}{\sqrt{\lambda}} \underset{\sim}{P}_{q'} \underset{\sim}{m}_q \quad (2.3) \end{cases}$$

onde

$$\underset{\sim}{P}_q = \underset{\sim}{Z}_q (\underset{\sim}{Z}'_q \underset{\sim}{Z}_q)^{-1} \underset{\sim}{Z}'_q \quad \text{e} \quad \underset{\sim}{P}_{q'} = \underset{\sim}{Z}_{q'} (\underset{\sim}{Z}'_{q'} \underset{\sim}{Z}_{q'})^{-1} \underset{\sim}{Z}'_{q'},$$

são as matrizes n×n que representam os operadores de projeção

sobre V_q e $V_{q'}$.

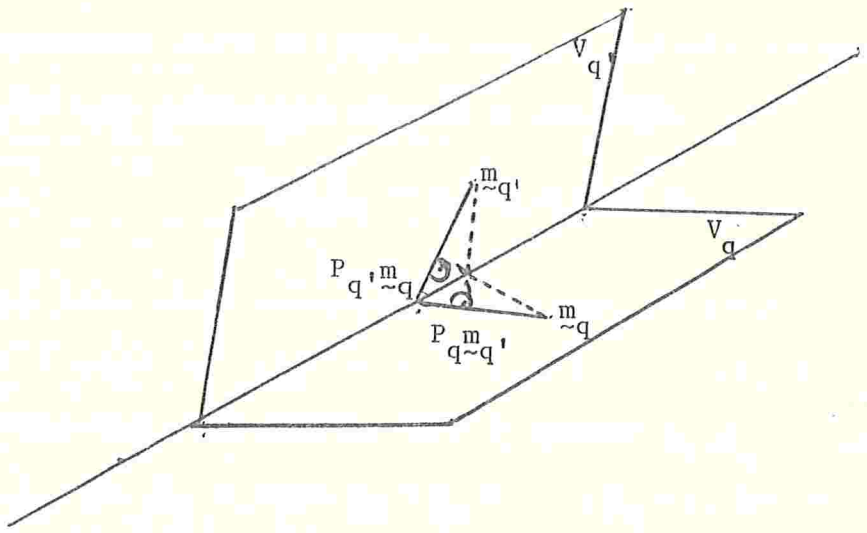


Figura 2.2 - Projeções sobre V_q e $V_{q'}$

Pelas relações (2.2) e (2.3) podemos ver que existe uma relação linear entre a projeção ortogonal de $\underset{\sim}{m}_q$ em $V_{q'}$ e $\underset{\sim}{m}_{q'}$ e também entre a projeção de $\underset{\sim}{m}_{q'}$ em V_q e $\underset{\sim}{m}_q$.

A análise canônica de $[Z_q/Z_{q'}]$ pode ser expressa da seguinte maneira: "encontrar dois pontos $\underset{\sim}{m}_q$ e $\underset{\sim}{m}_{q'}$ tal que a média dos quadrados de suas distâncias à origem seja constante, ou seja,

$$\underset{\sim}{\phi}_q' \underset{\sim}{D}_q \underset{\sim}{\phi}_q + \underset{\sim}{\phi}_{q'}' \underset{\sim}{D}_{q'} \underset{\sim}{\phi}_{q'} = 2n \qquad (2.4)$$

e tal que a distância à origem do ponto $\underset{\sim}{m} = \underset{\sim}{m}_q + \underset{\sim}{m}_{q'}$ seja máxima".

A distância do ponto $\underset{\sim}{m}$ à origem vale:

$$\|\underset{\sim}{m}\|^2 = \underset{\sim}{\phi}_q' \underset{\sim}{D}_q \underset{\sim}{\phi}_q + \underset{\sim}{\phi}_{q'}' \underset{\sim}{D}_{q'} \underset{\sim}{\phi}_{q'} + 2\underset{\sim}{\phi}_q' \underset{\sim}{Z}_q' \underset{\sim}{Z}_{q'} \underset{\sim}{\phi}_{q'} =$$

$$= 2n + 2\underset{\sim}{\phi}_q'\underset{\sim}{Z}_q'\underset{\sim}{Z}_q\underset{\sim}{\phi}_q' =$$

$$= 2n(1 + \frac{1}{n}\underset{\sim}{\phi}_q'\underset{\sim}{Z}_q'\underset{\sim}{Z}_q\underset{\sim}{\phi}_q').$$

Maximizar esta distância com a condição expressa por (2.4) é o mesmo que maximizar usando a condição

$$\underset{\sim}{\phi}_q'\underset{\sim}{D}_q\underset{\sim}{\phi}_q = \underset{\sim}{\phi}_q'\underset{\sim}{D}_q\underset{\sim}{\phi}_q' = n.$$

Com esta nova condição o problema se generaliza facilmente para o caso de mais de duas variáveis. Sejam $\phi_1, \phi_2, \ldots, \phi_Q$ os vetores das componentes dos Q pontos $\underset{\sim}{m}_1, \underset{\sim}{m}_2, \ldots, \underset{\sim}{m}_Q$ nas bases $\underset{\sim}{Z}_1, \underset{\sim}{Z}_2, \ldots, \underset{\sim}{Z}_Q$ e seja $\underset{\sim}{m}_1 + \underset{\sim}{m}_2 + \ldots + \underset{\sim}{m}_Q = \underset{\sim}{m}$. Devemos maximizar a quantidade:

$$\|m\|^2 = \sum_{q \in Q} \sum_{q' \in Q} \underset{\sim}{\phi}_q'\underset{\sim}{Z}_q'\underset{\sim}{Z}_q\underset{\sim}{\phi}_q'$$

com a condição

$$\sum_{q \in Q} \underset{\sim}{\phi}_q'\underset{\sim}{D}_q\underset{\sim}{\phi}_q = Qn.$$

Se $\underset{\sim}{\Phi}$ é o vetor com J componentes definido por

$$\underset{\sim}{\Phi}' = [\underset{\sim}{\phi}_1' \ \underset{\sim}{\phi}_2' \ \cdots \ \underset{\sim}{\phi}_Q']$$

devemos maximizar $\underset{\sim}{\Phi}'\underset{\sim}{B}\underset{\sim}{\Phi}$ com a condição $\underset{\sim}{\Phi}'\underset{\sim}{D}\underset{\sim}{\Phi} = Qn$.

Os fatores procurados são então v.c. de $\underset{\sim}{D}^{-1}\underset{\sim}{B}$ relativos às maiores r.c. Eles são proporcionais àqueles da análise de $\underset{\sim}{Z}$ e coincidem com os fatores da Tabela $\underset{\sim}{B}$ (Burt), considerando $\underset{\sim}{B}$ como a tabela de dados.

2.4 - TAXA DE INÉRCIA DE UMA MODALIDADE DE RESPOSTA E DE UMA VARIÁVEL

Como já vimos, os fatores da análise de $\underset{\sim}{Z}$ são v.c. da matriz $\frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{B}$. Logo, a soma das r.c., ou seja, a inércia total vale:

$$\operatorname{tr}\frac{1}{Q}\underset{\sim}{D}^{-1}\underset{\sim}{B} - 1 = \frac{J_1 + J_2 + \ldots + J_Q}{Q} - 1 = \frac{J}{Q} - 1.$$

Quando definimos contribuição relativa, na Seção 1.2.6, vimos que o quadrado da distância de uma variável j ao centro de gravidade G vale:

$$d^2(j,G) = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}}\left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot}\right)^2 .$$

Agora no nosso caso, podemos definir então o quadrado da distância ao centro de gravidade de um ponto modalidade j de $\mathbb{R}^n$, $j \in J$, como:

$$d^2(j,G) = \sum_{i=1}^{n} \frac{nQ}{Q}\left(\frac{Z_{ij}}{d_{jj}} - \frac{Q}{nQ}\right)^2 \quad \left(f_{i\cdot} = \frac{Q}{nQ}\right)$$

ou seja,

$$d^2(j,G) = n\sum_{i=1}^{n}\left(\frac{Z_{ij}}{d_{jj}} - \frac{1}{n}\right)^2 = n\sum_{i=1}^{n}\left(\frac{Z_{ij}^2}{d_{jj}^2} - \frac{2Z_{ij}}{nd_{jj}} + \frac{1}{n^2}\right),$$

onde d_{jj} é o j-ésimo elemento da diagonal de $\underset{\sim}{D}$.

Como

$$\sum_{i=1}^{n} Z_{ij} = d_{jj}$$

e Z_{ij} só assume valor zero ou um,

$$\sum_{i=1}^{n} Z_{ij} = \sum_{i=1}^{n} Z_{ij}^2 = d_{jj} \Rightarrow$$

$$\Rightarrow d^2(j,G) = n\left(\frac{1}{d_{jj}} - \frac{2}{n} + \frac{1}{n}\right) = n\left(\frac{1}{d_{jj}} - \frac{1}{n}\right).$$

A contribuição à inércia total da modalidade j vale então:

$$c(j) = \frac{d_{jj}}{nQ}\, d^2(j,G) = \frac{1}{Q}\left(1 - \frac{d_{jj}}{n}\right).$$

Logo quanto menor for o número de pessoas que respondem a modalidade j (ou seja, quanto menor for d_{jj}) maior será a contribuição desta modalidade para a inércia total. O máximo é alcançado quando $d_{jj} = 0$, ou seja, ninguém responde a modalidade j. Para haver um equilíbrio do sistema de respostas é necessário, então, evitar modalidades com poucas respostas.

A soma das contribuições de cada modalidade da variável q à inércia total dá a contribuição da variável q à inércia total, ou seja:

$$C(q) = \sum_{j \in J_q} c(j) = \sum_{j \in J_q} \frac{1}{Q}\left(1 - \frac{d_{jj}}{n}\right) \Rightarrow C(q) = \frac{1}{Q}(J_q - 1).$$

Logo a parte da inércia total devido a variável q cresce quando cresce o número de modalidades de resposta da variável (cresce quando J_q cresce). O mínimo é alcançado quando temos apenas duas modalidades de resposta ($J_q = 2$).

A seguir, discutimos a validade dos resultados da análise de correspondência binária e múltipla.

CAPÍTULO 3

VALIDADE DOS RESULTADOS DA ANÁLISE DE CORRESPONDÊNCIA

Neste capítulo mostramos que tipo de cuidados devemos ter para melhorar a qualidade dos resultados, o que podemos esperar desses resultados e como avaliamos a validade das representações obtidas quando aplicamos a técnica de análise de correspondência, bem como outras técnicas similares de análise de dados.

3.1 - CUIDADOS NA ELABORAÇÃO DAS TABELAS DE DADOS

Os dados submetidos à análise devem possuir características como a homogeneidade e a exaustividade.

A *homogeneidade* é obtida através de codificação das variáveis que deve permitir que as linhas e as colunas da tabela possam ser comparadas entre si ou através de transformações analíticas. Devemos, por exemplo, nos preocupar com o fato de que unidades de medidas diferentes não sejam consideradas simultaneamente na análise, pois caso contrário temos que recorrer a transformação de padronização das variáveis. A codificação disjuntiva, por exemplo, permite que a análise de correspondência trate simultaneamente variáveis nominais e variáveis contínuas divididas em classes.

A distinção entre as variáveis ativas e ilustrativas da tabela também é fundamental. As variáveis ilustrativas,têm um caracter de prova, pois como não participam na construção dos fatores a interpretação de sua correlação com os fatores é feita com mais segurança

O critério de *exaustividade* diz que todas as situações ou aspectos de um fenômeno (variável) devem estar representados na tabela. Na prática, obtemos a exustividade da tabela fazendo um levantamento vasto de dados, e definindo categorias ou classes das variáveis de forma a permitir que sejam alocados todos os valores de cada variável.

Um fator importante na qualidade dos resultados é que eles sejam estáveis, ou seja, não sofram modificações sensíveis quando ocorrem modificações na tabela original de dados.Existem três principais elementos que podem causar modificações na tabela original, mas que não devem afetar as configurações obtidas como resultado da análise, se estas são supostas estáveis:

a) ERROS DE MEDIDAS

A ordem de grandeza dos erros de medida, assim como a distribuição aproximada dos mesmos na população devem ser especificados pelo próprio pesquisador em função de seu conhecimento do campo estudado;

b) ESCOLHA E PESO DAS VARIÁVEIS

O problema da escolha das variáveis aparece quando o estatístico tem a possibilidade de colher uma amostra de variáveis dentro do espaço das variáveis. Ele pode se basear nos critérios de homogeneidade e exaustividade para fazer esta amos-

tragem corretamente, mas isto nem sempre é suficiente. O problema do peso das variáveis aparece sobretudo quando a análise de correspondência é aplicada em tabelas de medidas ou notas e não de contagens;

c) CODIFICAÇÃO DAS VARIÁVEIS

Designamos por codificação a transformação preliminar a que são submetido os dados brutos, antes de iniciarmos a análise. A codificação é feita com a preocupação de não se perder informação e de se tornar mais fácil o uso do algoritmo da análise. A codificação é um processo fundamentalmente empírico, que supõe um conhecimento simultâneo dos dados, através de estatísticas discritivas, do método da análise e suas exigências. É importante se fazer uma codificação capaz de conservar as configurações realmente observadas.

3.2 - CONTRIBUIÇÕES DOS RESULTADOS DA ANÁLISE

A aplicação da técnica de análise de correspondência bem como de outras técnicas de análise de dados nos traz algumas contribuições importantes no campo da pesquisa usando grandes tabelas. Citaremos algumas destas contribuições:

a) GANHO DE PRODUTIVIDADE E DETECÇÃO DO ERRO

Em certas fases do desenvolvimento da pesquisa,o caso das técnicas de análise de dados é indispensável, pois estas técnicas controlam, com as representações visuais (gráficos) que fornecem, a maior parte das etapas do trabalho, e dão acesso a informação antes inascessíveis. Permitem também a detecção

de valores enormes ou valores aberrantes que são facilmente reveladas nos eixos fatoriais. Assim, erros de medidas ou anomalias devidas ao método de medida são facilmente detectadas.

O conjunto de operações da técnica de análise de dados implica num ganho de produtividade e num melhoramento da qualidade dos resultados através da correção dos erros detectados.

b) CONSTRUÇÃO DE ÍNDICES SINTÉTICOS

O primeiro fator, encontrado pela análise de correspondência ou outra análise de dados, é uma combinação linear das variáveis e tem a maior variância, constituindo assim um excelente índice discriminatório entre as observações. Estes índices, chamados índices sintéticos, têm grande poder descritivo e possuem interpretações interessantes.

c) ACESSO A NOVOS CAMPOS DE OBSERVAÇÃO E NOVOS MATERIAIS

As técnicas de análise de dados permitem tratar simultaneamente as numerosas informações contidas nas grandes tabelas, tornando possível observar os universos multidimensionais complexos. A análise de correspondência, principalmente, permite que se estude os dados sobre as duas dimensões das tabelas. As duas dimensões (de linhas e colunas) são formadas por variáveis que não apresentam o caracter repetitivo das observações como acontece na aplicação das outras técnicas de dados, quando uma das dimensões (geralmente as linhas da tabela) é formada por observações e não variáveis. A apresentação dos resultados na forma de gráficos, onde os primeiros fatores extraídos da análise são cruzados dois a dois, constitui

uma inovação metodologica que facilita muito a interpretação e apresentação dos resultados.

3.3 - AVALIAÇÃO DA VALIDADE DOS RESULTADOS

Vamos fazer agora algumas considerações sobre as raízes características extraídas na análise de correspondência, sua distribuição, e as taxas de inércia ou porcentagens de variância. Discutimos o quanto a taxa de inércia pode ser considerada como medida da informação obtida pelo eixo fatorial. Vamos construir também intervalos de confiança para os pontos nos gráficos dos pontos fatoriais, para ajudar a interpretar, a proximidade dos pontos à origem.

3.3.1 - A distribuição das raízes características em análise de correspondência

A hipótese de independência das linhas e das colunas de uma tabela, em geral, é muito severa para ser realista. É improvável que uma tabela submetida à análise de correspondência possa ser análoga a uma tabela de números aleatórios. Entretanto a hipótese de independência permite que se defina "níveis de significância" para as raízes características (r. c.) e a porcentagem de variância. No caso da análise de correspondência para tabelas de contingência, as r.c. seguem leis não paramétricas e assim sendo é possível fazer tabulações aproximadas das mesmas e apresentá-las em forma de gráficos.

A distribuição das r.c. produzidas pela análise de correspondência deu margem a algumas publicações errôneas: Ken-

dall & Stuart (1961), concluíram que as r.c., assim como a inércia total, seguem as leis do qui-quadrado (χ^2): Lancaster (1963) mostrou que a esperança matemática da primeira r. c. é sempre superior aos valores encontrados pelo método de Kendall & Stuart; posteriormente Kshirsagar (1972) sugeriu outra distribuição para as r.c.

Vamos mostrar aqui, que a distribuição das r.c. pode ser aproximada pela das r.c. de uma matriz de Wishart, cuja lei é conhecida. A densidade de probabilidade das r.c. de uma matriz de Wishart pode ser encontrada no livro de Anderson (1958).

Usamos, aqui, a mesma notação dos Capítulos 1 e 2.

Seja p_{ij} a probabilidade da casela (i,j) e $p_{i.}$ e $p_{.j}$ as probabilidades marginais de linhas e colunas respectivamente. A hipótese de independência das linhas e das colunas é expressa pela relação $p_{ij} = p_{i.} \cdot p_{.j}$ para todo i e j. Então x_{ij} é uma das n_p componentes de um vetor multinomial cuja esperança matemática vale $E(x_{ij}) = xp_{i.}p_{.j}$, sob a hipótese de independência. Supomos que X é suficientemente grande para que se possa utilisar a aproximação normal da lei multinomial e que $f_{i.}$ e $f_{.j}$ são os estimadores de $p_{i.}$ e $p_{.j}$ respectivamente.

Então um vetor $\underset{\sim}{k}$ com np componentes

$$k_{ij} = \frac{\sqrt{X}(f_{ij} - f_{i.}f_{.j})}{\sqrt{f_{i.}f_{.j}}}$$

tem distribuição normal com média 0 ($E(k_{ij})=0$, para todo i e j) e matriz de covariâncias definida por: (Rao, 1973)

$$V_k(i,j,i',j') = \delta_{ij,i'j'} - \sqrt{f_{i.}f_{.j}f_{i'.}f_{.j'}}$$

onde

$$\delta_{ij,i'j'} = \begin{cases} 1 \text{ se } i=i' \text{ e } j=j' \\ 0 \text{ caso contrário} \end{cases}$$

Seja $\underset{\sim}{A}$ uma matriz $p \times p$ ortogonal tal que a primeira coluna tem como j-ésimo elemento o valor $\sqrt{f_{.j}}$, $j=1,\ldots,p$, e as p-1 outras colunas formam junto com a primeira, uma base ortonormal de $\mathbb{R}^p$.

Seja $\underset{\sim}{B}$ uma matriz $n \times p$ ortogonal tal que na primeira linha o i-ésimo elemento vale $\sqrt{f_{i.}}$, $i=1,\ldots,n$ e as outras n - 1 linhas formam, junto com a primeira, uma base ortonormal de $\mathbb{R}^n$.

A matriz $\underset{\sim}{B} \otimes \underset{\sim}{A}'$ de ordem $np \times np$ definida pelo produto direto ou produto de Kronecker de $\underset{\sim}{B}$ e $\underset{\sim}{A}'$ é também ortogonal.

Seguem as relações:

$$\sum_j \sqrt{f_{.j}}\, k_{ij} = 0 \text{ para todo } i$$

$$\sum_i \sqrt{f_{i.}}\, k_{ij} = 0 \text{ para todo } j$$

$$\sum_m b_{rm} \sqrt{f_{m.}} = 0 \text{ para todo } r,\ 1 < r \leq n$$

$$\sum_d a_{ds} \sqrt{f_{.d}} = 0 \text{ para todo } s,\ 1 < s \leq p.$$

Essas relações fazem com que o vetor $\underset{\sim}{y}$ de $\mathbb{R}^{np}$ definido por $\underset{\sim}{y} = \underset{\sim}{B} \otimes \underset{\sim}{A}'\underset{\sim}{k}$ tenha só (n-1)(p-1) componentes não nulas, pois $y_{rs} = 0$ se $r = 1$ ou se $s = 1$. A matriz de covariâncias de $\underset{\sim}{y}$ é definida por:

$$V\underset{\sim}{y} = (\underset{\sim}{B} \otimes \underset{\sim}{A}')\underset{\sim}{V}_k(\underset{\sim}{B}' \otimes \underset{\sim}{A})$$

e para todo par de componentes não nulos temos

$$V\underset{\sim}{y}(r,s;r',s') = \delta_{rr'}\delta_{ss'}$$

onde

$$\delta_{rr'} = \begin{cases} 1 \text{ se } r = r' \\ 0 \text{ caso contrário} \end{cases}$$

e

$$\delta_{ss'} = \begin{cases} 1 \text{ se } s = s' \\ 0 \text{ caso contrário.} \end{cases}$$

Seja $\underset{\sim}{Y}$ a matriz de ordem $n \times p$ definida por

$$\underset{\sim}{Y} = \underset{\sim}{B}\underset{\sim}{K}\underset{\sim}{A}$$

onde $\underset{\sim}{K}$ é a matriz cujo termo geral vale k_{ij}, definido anteriormente. A primeira linha e a primeira coluna de $\underset{\sim}{Y}$ são nulas. A matriz $\hat{\underset{\sim}{Y}}$ de ordem $n-1 \times p-1$ formada pelas linhas e colunas não nulas $\underset{\sim}{Y}$ tem seus elemento independentemente distribuídos segundo a normal padrão. A matriz

$$\underset{\sim}{S} = \hat{\underset{\sim}{Y}}'\hat{\underset{\sim}{Y}}$$

é distribuída, então, segundo a lei de Wishart com parâmetros $n-1$ e $p-1$ e temos que $\underset{\sim}{S}$ tem as mesmas r.c. não nulas de

$$\underset{\sim}{Y}'\underset{\sim}{Y} = \underset{\sim}{A}'\underset{\sim}{K}'\underset{\sim}{B}'\underset{\sim}{B}\underset{\sim}{K}\underset{\sim}{A} = \underset{\sim}{A}'\underset{\sim}{K}'\underset{\sim}{K}\underset{\sim}{A}$$

e portanto as mesmas r.c. de $\underset{\sim}{K}'\underset{\sim}{K}$ pois $\underset{\sim}{A}$ é ortogonal. A matriz simétrica $\underset{\sim}{S}^*$, definida do Capítulo 1, que temos que diagonalizar na análise de correspondência da tabela $\underset{\sim}{X}$ não é outra se-

não a matriz

$$\underset{\sim}{S}^* = \frac{1}{x}\underset{\sim}{K}'\underset{\sim}{K}.$$

Isto nos leva a seguinte conclusão: "Se λ_α é a α-ésima r.c. da análise de correspondência da tabela $\underset{\sim}{X}$ de ordem $n \times p$, cujo total vale x, então a distribuição de $x\,\lambda_\alpha$ é aproximadamente a distribuição da α-ésima r.c. da matriz de Wishart com parâmetros n-1 e p-1".

Temos então que a distribuição das r.c. extraídas da análise de correspondência de uma tabela de contingência só depende das dimensões da tabela, não importando os pesos de cada linha ou cada coluna da tabela. Esta propriedade permite que sejam feitas tabulações dos r.c. só com as dimensões das tabelas. Entretanto, é uma propriedade assintótica e deve ser sempre verificado o realismo da mesma.

A densidade da lei conjunta das r.c. $\lambda_1,\ldots,\lambda_p$ de uma matriz de Wishart é: (Anderson, 1958)

$$w(\Lambda) = C(n,p) \prod_{\alpha=1}^{p} \lambda_\alpha^{(n-p-1)/2} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{p}\lambda_\alpha\right\} \prod_{\alpha<\beta}(\lambda_\alpha-\lambda_\beta)$$

onde

$$C(n,p) = \left(\pi^{p/2}/2^{np/2}\right)\cdot\left[\prod_{\alpha=1}^{p}\Gamma\left(\frac{n+1-\alpha}{2}\right)\right]\Gamma\left(\frac{p+1-k}{2}\right).$$

A Figura 3.7 compara a lei teórica das r.c. com a lei empírica obtida por simulação no caso de uma tabela com 8 colunas e o número de linhas variando de 3 a 100.

Mostramos no final deste capítulo algumas tabelas e grá-

ficos:

— As Figuras 3.1 a 3.5 representam a taxa de inércia das cinco primeiras r.c., para dimensões variadas da tabela de dados. Assim, por exemplo na Figura 3.1, para uma tabela de dimensão 10 × 8, a taxa de inércia correspondente à primeira r. c. está em torno de 50%, ao nível de significância de 5%, sob a hipótese de independência das linhas e colunas da tabela.

— A Figura 3.6 é uma tabela aproximada dando as estimações das médias, medianas e desvios-padrões, a um nível de significância de 5%, das cinco primeiras r.c., das cinco porcentagens de inércia correspondentes e do traço. A dimensão da tabela varia de 6 × 6 até 6 × 20 e o total da tabela é fixado em 1.000. Assim, por exemplo, para uma tabela 6 × 8 a média da segunda r.c. é 0,0096, a média da porcentagem da inércia desta mesma r.c. vale 27,93% e a média do traço vale 0,0346.

— Finalmente a Figura 3.7 compara a lei teórica das r.c. com a lei empírica obtida por simulação, no caso de uma tabela com 8 colunas (largura) e o número de linhas (comprimento) variando de 3 a 100 e considerando apenas a primeira r.c.

3.3.2 - A independência entre a taxa de inércia e o traço

Seja t a soma das r.c.:

$$t = \sum_{\alpha=1}^{p} \lambda_\alpha ,$$

ou seja t é o traço da matriz e τ_α a taxa de inércia do fator

α: $\tau_\alpha = \lambda_\alpha / t$ ou seja, $\lambda_\alpha = \tau_\alpha \cdot t$, para $\alpha < p$ e

$$\lambda_p = (1 - \tau_1 - \tau_2 - \ldots - \tau_{p-1})t.$$

A densidade conjunta $w(\Lambda)$ pode ser fatorada na forma:

$$w(\Lambda) = w_1(t) w_2(\tau_1, \ldots, \tau_{p-1}).$$

O jacobiano desta transformação vale t^{p-1}.

A distribuição do traço t, $w_1(t)$, segue a lei do qui-quadrado (χ^2) com np graus de liberdade. Logo,

$$w_1(t) = \frac{1}{2\Gamma\left(\frac{np}{2}\right)} \left(\frac{t}{2}\right)^{(np/2)-1} e^{-t/2}.$$

A distribuição conjunta $w(\Lambda)$ pode então ser fatorada em duas funções com domínios de integração independentes. Então podemos concluir que as taxas de inércia (ou porcentagens de variância) e o traço são independentes.

Com este fato, chegamos a duas conclusões importantes:

— Mesmo se o traço (ou o χ^2) aceita a hipótese de independência, a primeira porcentagem de variância pode ser significativamente elevada e neste caso a análise de correspondência deve ser usada, pois apesar de aceitarmos a independência entre as linhas e as colunas da tabela, estas podem ser rica em informação.

— Por outro lado, podemos ter um traço elevado que rejeita a hipótese de independência mas a porcentagem de variância não ser significante. Neste caso a análise de correspondência poderia não ser o melhor método para obter as informações das

relações entre as linhas e as colunas da tabela.

3.3.3 - Taxa de inércia como medida de informação

O problema do uso da taxa de inércia como medida da informação representada por um fator extraído da análise de correspondência é bastante delicado. No caso da análise de correspondência para tabelas de contingência, o uso da taxa de inércia é satisfatório e dá uma idéia boa da informação obtida, mas já no caso de tabelas disjuntivas completas as taxas de inércia são bem mais fracas que as que seriam obtidas se cruzássemos as variáveis e aplicássemos a análise à tabela cruzada. Então ocorre que para tabelas disjuntivas completas as taxas de inércia dão uma idéia muito pessimista da parte da informação representadas pelos eixos fatoriais.

A escolha das variáveis também influencia na taxa de inércia. Se completarmos uma tabela $n \times p$, na qual foi feita a análise de correspondência com mais q novas colunas formadas por números tirados de uma tabela de números aleatórios e fizermos novamente a análise de correspondência a esta nova tabela $n \times p+q$, os primeiros eixos obtidos serão praticamente os mesmos que os obtidos na tabela $n \times p$. Mas as taxas de inércia da nova tabela serão mais fracas pois o traço aumentou já que temos mais r.c. (passamos de p para $p+q$). Logo, os eixos não mudam, ou seja continuam representando a mesma informação, e as taxas de inérica diminuem, ficando mais uma vez provado que as taxas de inércia não devem medir a parte de informação representada pelo eixo fatorial. A escolha das variáveis tem en-

tão influência negativa na taxa de inércia e não influencia significativamente nos fatores da análise.

Vamos agora ver a *teoria da informação de Shannon-Wiener* onde a taxa de inércia não aparece como medida de não-esferacidade de uma nuvem. O desenvolvimento desta teoria é baseado na medida da distância entre duas hipóteses dada pela "divergência de Jeffreys" (Jeffreys, 1946). Sejam H_1 e H_2 duas hipóteses no caso de uma amostra multidimensional X, saída de um dos dois esquemas relativos às leis normais de $\mathbb{R}^p$:

H_1: Hipótese de Independência:

Média teórica = $\underset{\sim}{\mu}_1$

Matriz de covariância teórica = $\sigma^2 \underset{\sim}{I}$

H_2: Caso Geral:

Média teórica = $\underset{\sim}{\mu}_2$

Matriz de covariância teórica = $\underset{\sim}{S}$ (suposta positiva definida)

A divergência de Jerffreys permite expressar a distância entre H_1 e H_2 em função das r.c. de $\underset{\sim}{S}$ e considera as r.c. pequenas, enquanto que a análise de correspondência só se preocupa com as r.c. "grandes".

Essa divergência é definida por:

$$J(H_1,H_2) = \int \log \frac{P(H_1/\underset{\sim}{X})}{P(H_2/\underset{\sim}{X})} \, d\gamma_1(\underset{\sim}{X}) - \int \log \frac{P(H_2/\underset{\sim}{X})}{P(H_1/\underset{\sim}{X})} \, d\gamma_2(\underset{\sim}{X})$$

onde

$P(H_i/\underset{\sim}{X})$ = probabilidade condicional de H_i ser verdadeira dado $\underset{\sim}{X}$; e

γ_1 e γ_2 são as medidas associadas a H_1 e H_2.

No caso de densidades contínuas $f_1(\underset{\sim}{X})$ e $f_2(\underset{\sim}{X})$ temos:

$$J(H_1,H_2) = \int (f_1(\underset{\sim}{X})+f_2(\underset{\sim}{X}))\log \frac{f_1(\underset{\sim}{X})}{f_2(\underset{\sim}{X})}\, d\underset{\sim}{X}.$$

Quando a matriz de covariância é $\underset{\sim}{S}_i$ e o vetor médio é $\underset{\sim}{\mu}_i$, a densidade $f_i(\underset{\sim}{X})$ é:

$$f_i(\underset{\sim}{X}) = \frac{1}{|2\pi\underset{\sim}{S}_1|^{1/2}} \exp -\left\{\frac{1}{2}(X-\underset{\sim}{\mu}_i)'\underset{\sim}{S}_i^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}_i)\right\}$$

e assim:

$$\log \frac{f_1(\underset{\sim}{X})}{f_2(\underset{\sim}{X})} = \frac{1}{2}\log\frac{|\underset{\sim}{S}_2|}{|\underset{\sim}{S}_1|} - \frac{1}{2}\operatorname{tr}(\underset{\sim}{S}_1^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}_1)(\underset{\sim}{X}-\underset{\sim}{\mu}_1)') +$$

$$+ \frac{1}{2}\operatorname{tr}(\underset{\sim}{S}_2^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu}_2)(\underset{\sim}{X}-\underset{\sim}{\mu}_2)')$$

O primeiro termo de $J(H_1,H_2)$ é a informação média trazida pela amostra $\underset{\sim}{X}$ sob a hipótese H_1 contra H_2:

$$I(1;2) = \int f_1(\underset{\sim}{X})\log\frac{f_1(\underset{\sim}{X})}{f_2(\underset{\sim}{X})}\, d\underset{\sim}{X} = \frac{1}{2}\log\frac{|\underset{\sim}{S}_1|}{|\underset{\sim}{S}_2|} +$$

$$+ \frac{1}{2}\operatorname{tr}(\underset{\sim}{S}_1(\underset{\sim}{S}_2^{-1}-\underset{\sim}{S}_1^{-1})) + \frac{1}{2}\operatorname{tr}(\underset{\sim}{S}_2^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)')$$

(colocamos $x - \mu_2 = \underset{\sim}{x}-\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2$).

Então,

$$J(H_1,H_2) = I(1;2) + I(2;1) = \frac{1}{2}\operatorname{tr}((\underset{\sim}{S}_1-\underset{\sim}{S}_2)(\underset{\sim}{S}_2^{-1}-\underset{\sim}{S}_1^{-1})) +$$

$$+ \frac{1}{2}\operatorname{tr}(\underset{\sim}{S}_1^{-1}+\underset{\sim}{S}_2^{-1})(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)')$$

O caso que nos interessa é quando $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ e

$$\underset{\sim}{S}_1 = \sigma^2 \underset{\sim}{I} \quad e \quad \underset{\sim}{S}_2 = \underset{\sim}{S}$$

ou seja,

$$J(\sigma^2 \underset{\sim}{I}; \underset{\sim}{S}) = \frac{1}{2} \operatorname{tr} \left((\sigma^2 \underset{\sim}{I} - \underset{\sim}{S})(\underset{\sim}{S}^{-1} - \frac{1}{\sigma^2} \underset{\sim}{I}) \right) = \frac{1}{2} \operatorname{tr} \left(\frac{1}{\sigma^2} \underset{\sim}{S} + \sigma^2 \underset{\sim}{S}^{-1} \right) - p.$$

Em função das r.c. podemos escrever:

$$J(\sigma^2 \underset{\sim}{I}, \underset{\sim}{S}) = \frac{1}{2} \left(\frac{1}{\sigma^2} \sum_{\alpha=1}^{p} \lambda_\alpha + \sigma^2 \sum_{\alpha=1}^{p} \frac{1}{\lambda_\alpha} \right) - p.$$

Se as inércias totais teóricas são iguais sob as hipóteses H_1 e H_2 temos:

$$\sum_{\alpha=1}^{p} \lambda_\alpha = p\sigma^2$$

e o único termo variável de $J(\sigma^2 \underset{\sim}{I}; \underset{\sim}{S})$ é

$$\sum_{\alpha=1}^{p} \frac{1}{\lambda_\alpha}.$$

Temos então que a divergência entre as duas hipóteses será grande quando algumas r.c. forem próximas do zero. Logo, uma r.c. de $\underset{\sim}{S}$ muito pequena, será muito mais importante na determinação da divergência do que, por exemplo, as primeiras r.c. explicando 80% da inércia total. Vamos introduzir agora o conceito de *informação mútua* no caso da análise de correspondência.

Seja (i,j) um elemento aleatório de um conjunto I×J. P_{IJ} designará a lei de probabilidades deste conjunto e P_I e P_J as leis marginais correspondentes, isto é,

$$P_{IJ} = \{P_{ij}, i \in I \text{ e } j \in J\}$$

$$P_I = \{p_{i.},\ i \in I\}$$

$$P_J = \{p_{.j},\ j \in J\}.$$

Chamamos de grau de indeterminação, ou informação sobre I (sobre J) a quantidade

$$H(P_I) = -\sum_{i \in I} p_{i.} \log p_{i.}; (H(P_j) = -\sum_{j \in J} p_{.j} \log p_{.j}).$$

A informação mútua entre I e J é definida por:

$$H(P_{IJ}; P_I P_J) = H(P_I) + H(P_J) - H(P_{IJ}),$$

que não é outra senão a informação $I(P_I P_J;\ P_{IJ})$ de Shannon-Wiener.

$$H(P_{IJ}; P_I P_J) = \sum_{ij}\sum p_{ij} \log \frac{p_{ij}}{p_{i.} p_{.j}}.$$

Na vizinhança da hipótese de independência, temos a aproximação:

$$H(P_{IJ}; P_I P_J) \simeq \frac{1}{2}\left\{\sum_{ij}\sum \frac{(p_{ij} - p_{i.} p_{.j})^2}{p_{i.} p_{.j}}\right\}.$$

Assim, a soma das r.c. não triviais da análise de correspondência representa uma aproximação da informação mutua entre linhas e colunas da tabela, quando não nos afastamos muito da hipótese de independência.

A parte da inércia de um sub-espaço pode ser significativa sem que a informação global o seja.

3.3.4 - Intervalos de Confiança para pontos dos gráficos dos eixos fatoriais

A análise de correspondência fornece os gráficos com as nuvens de pontos nos planos de projeção formados pelos eixos fatoriais tomados dois a dois. Vamos calcular agora os intervalos de confiança para estes pontos dos gráficos, a fim de verificar se eles estão significantemente próximos da origem ou não.

Na Tabela $\underset{\sim}{X}$ de dados com n linhas e p colunas, cada uma das n linhas forma um vetor com p coordenadas $f_{ij}/f_{i.}$, ou seja, forma um ponto de $\mathbb{R}^p$. O ponto médio G dos n pontos-linhas têm como coordenadas os valores $f_{.j}$, $j = 1,\ldots,p$. A "distância do qui-quadrado" entre cada ponto-linha i e G vale:

$$d^2(i,G) = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 .$$

Se x é a soma total dos elementos da tabela a quantidade $c_i^2 = xf_{i.}d^2(i,G)$ vale:

$$c_i^2 = x \sum_{j=1}^{p} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

e temos que c_i^2 é aproximadamente um qui-quadrado com p-1 graus de liberdade,

Se χ^2 é não significante, o ponto i só difere de G devido a flutuações amostrais.

Considerada a projeção dos pontos no sub-espaço formado pelos dois primeiros eixos fatoriais da análise de correspon-

dência, $d^2(i,G)$ é proporcional a um qui-quadrado com dois graus de liberdade (basta multiplicar por $xf_{i.}$ para dar um qui-quadrado). Isto nos leva a um procedimento simples para testar a significância das posições dos pontos no gráfico em relação a origem. Podemos construir círculos centrados na origem com raio r igual a

$$\sqrt{\frac{\chi^2_{2;\alpha}}{xf_{i.}}}$$

onde $\chi^2_{2;\alpha}$ = qui-quadrado com dois graus de liberdade e nível de significância α, cujo valor é encontrado na Tabela de distribuição do qui-quadrado. Se a projeção do ponto i, i=1,...,n, está fora deste círculo, com uma probabilidade α, podemos afirmar que o i-ésimo ponto da tabela é significantemente diferente da origem e, portanto, traz informação sobre os dados.

Na prática, invés de construirmos círculos centrados na origem, é mais fácil contrui-los em torno de cada ponto e olhar a sua posição em relação à origem, ou seja, se esta está dentro ou fora do círculo.

No capítulo a seguir, apresentamos alguns exemplos de aplicação onde discutimos a validade dos resultados da análise, aplicando o que foi visto neste capítulo.

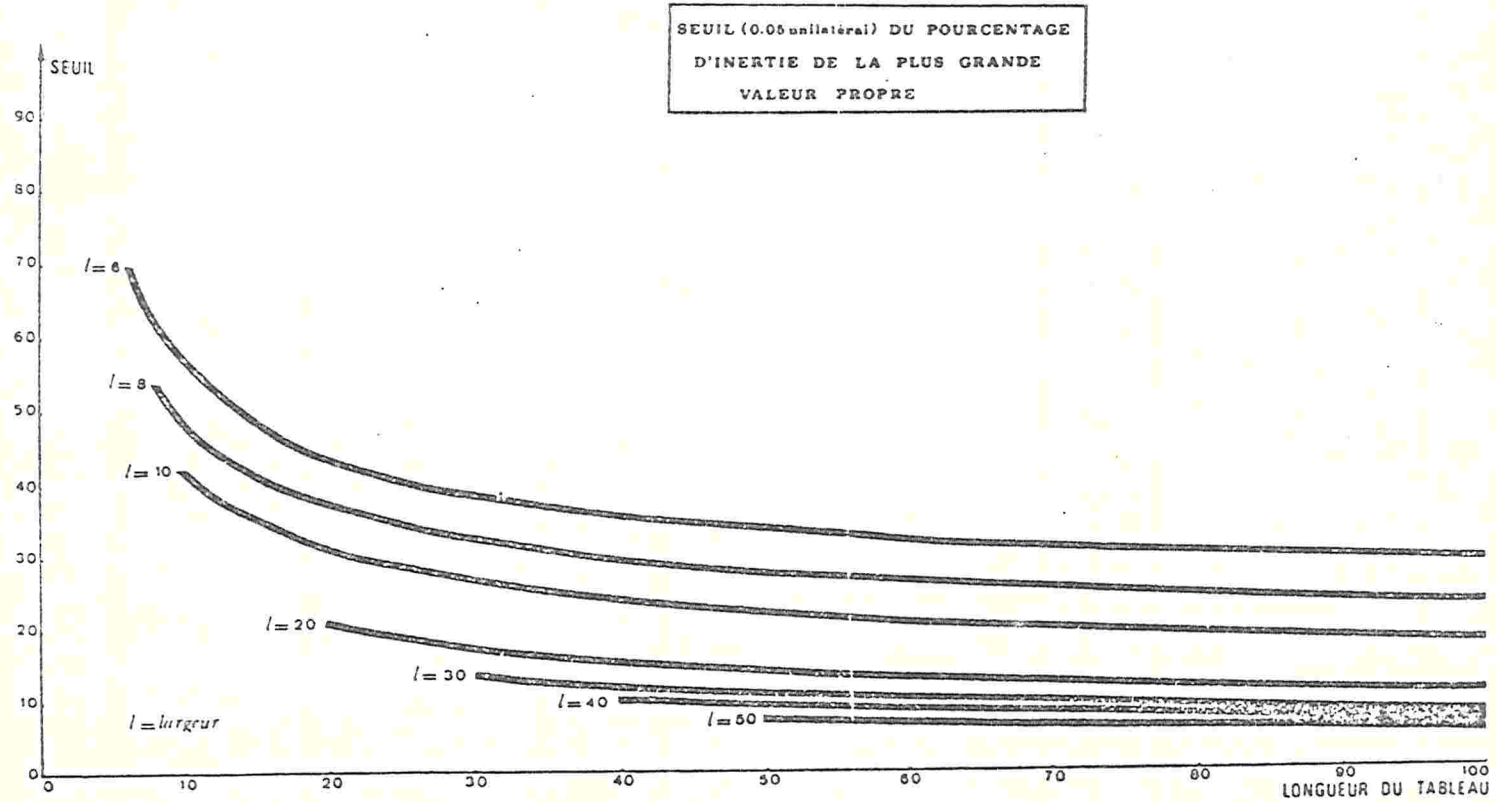


Figura 3.1 - Taxas de inércia da primeira raiz característica para tabelas de diferentes tamanhos.

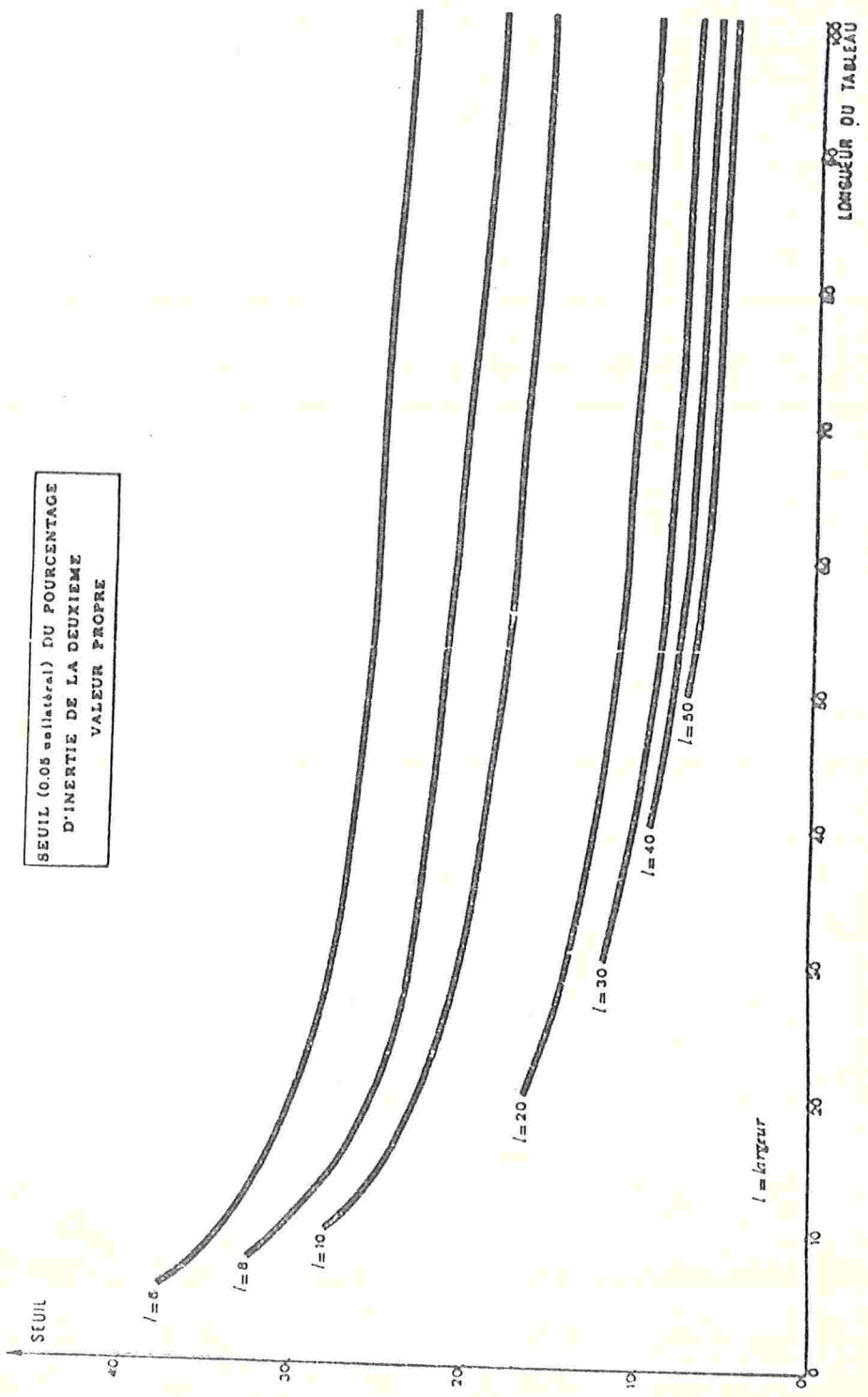


Figura 3.2 - Taxas de inércia da segunda raiz característica para tabelas de diferentes tamanhos.

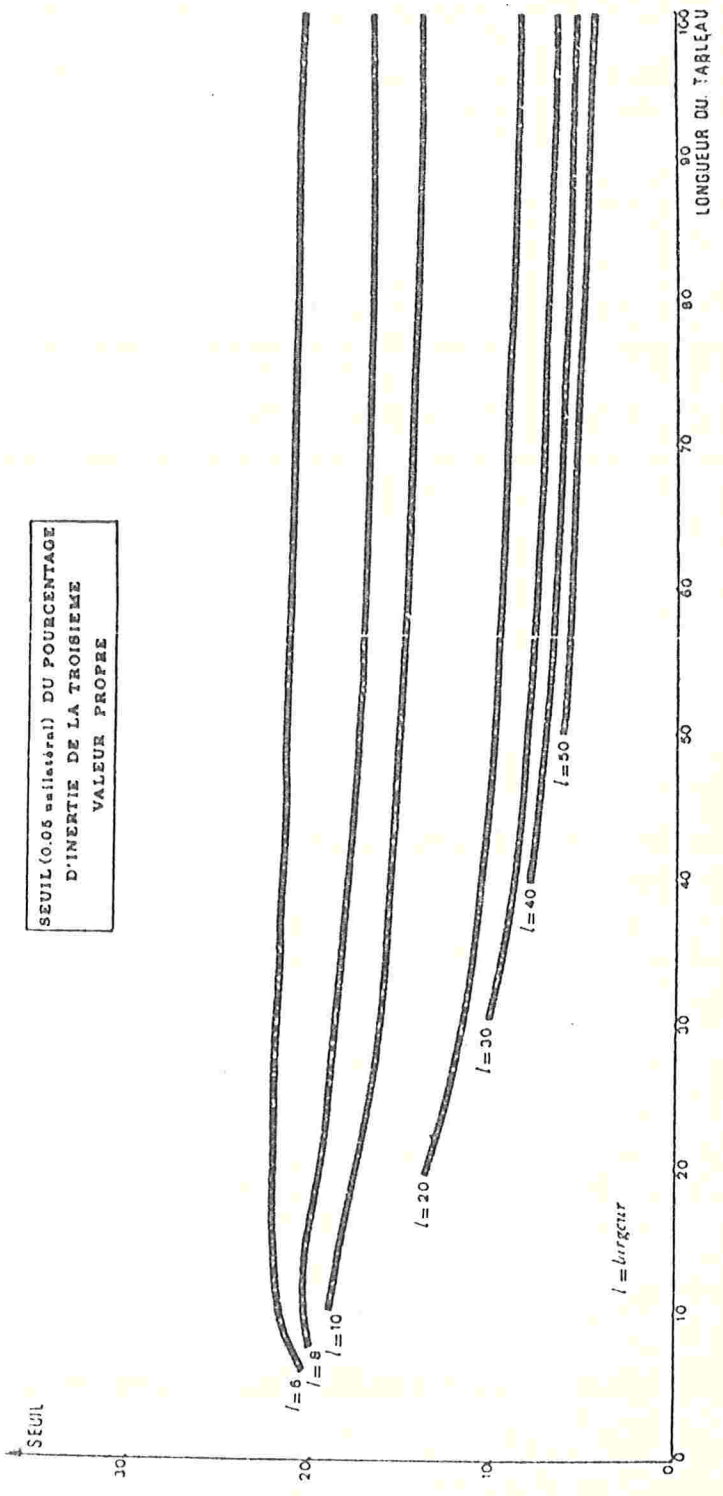


Figura 3.3 - Taxas de inércia da terceira raiz característica para tabelas de diferentes tamanhos.

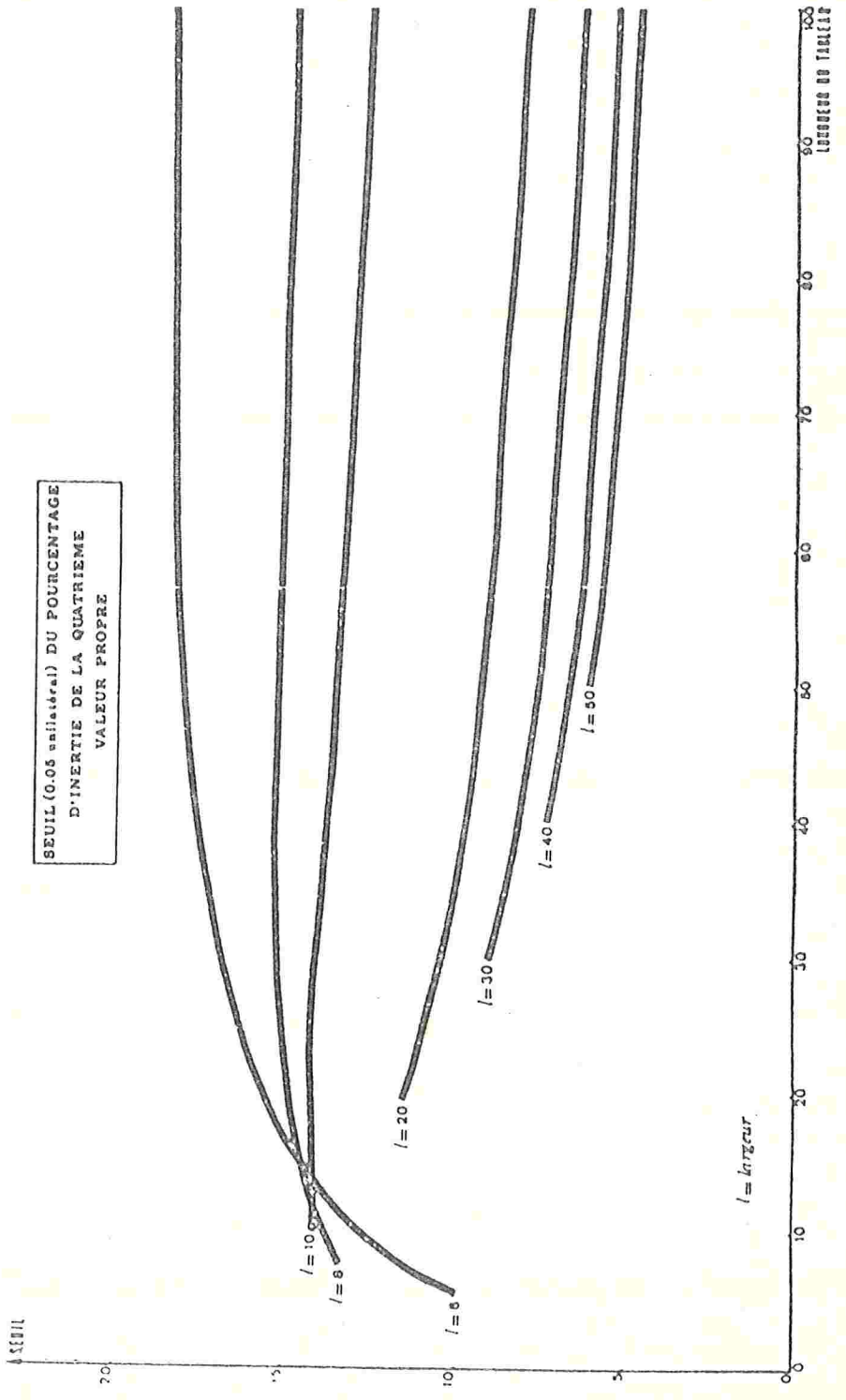


Figura 3.4 - Taxas de inércia da quarta raiz característica para tabelas de diferentes tamanhos.

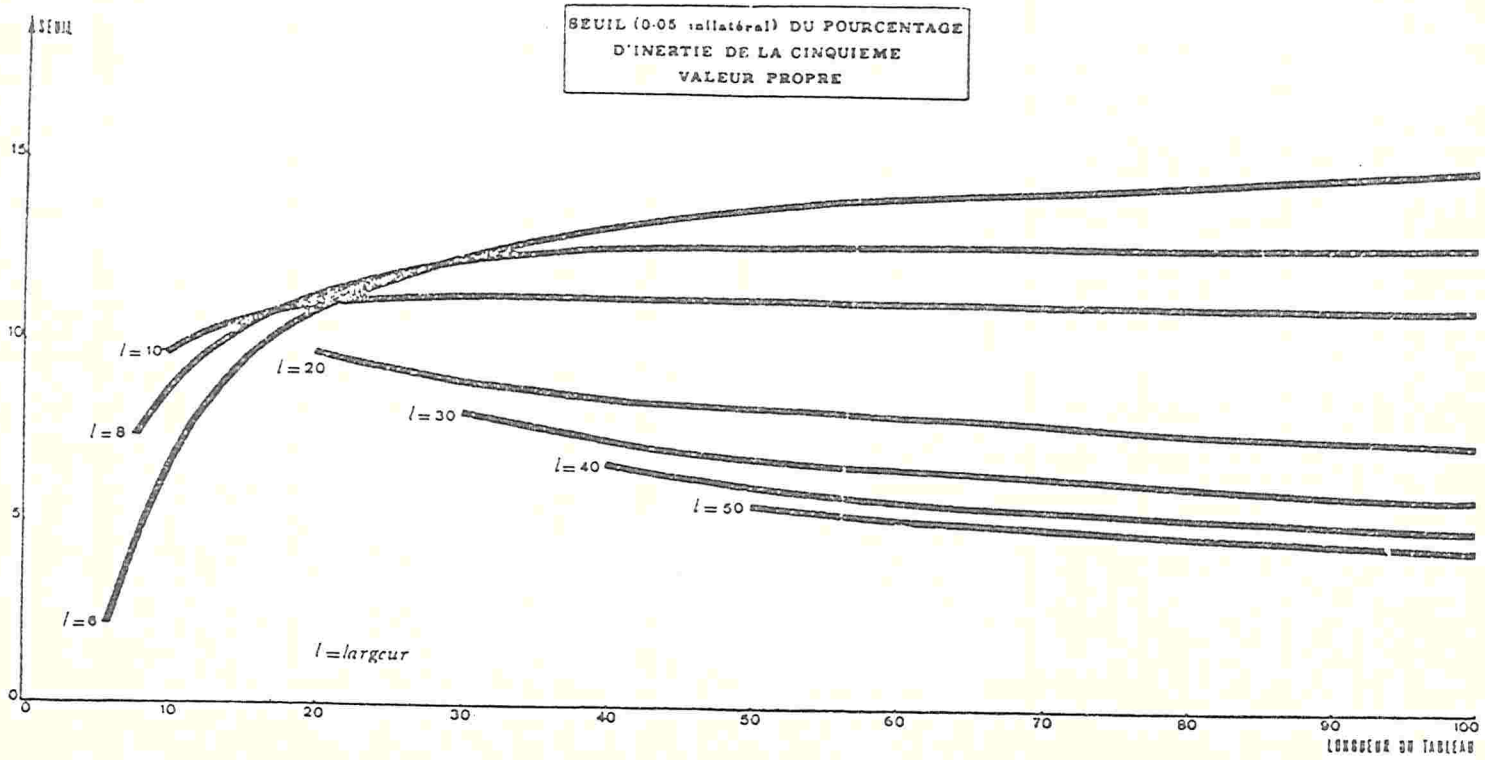


Figura 3.5 - Taxas de inércia da quinta raiz característica para tabelas de diferentes tamanhos.

SIGNIFICATION DES VALEURS PROPRES (VP),ET DES POURCENTAGES (PC) EN ANALYSE DES CORRESPONDANCES (EFFECTIFS= 1000.)

LARGEUR DU TABLEAU= 6

LONG.= 6	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0139	0.0071	0.0031	0.0012	0.0002	54.17	27.97	12.44	4.84	0.59	0.0255
MEDIANES	0.0138	0.0065	0.0028	0.0010	0.0001	53.21	28.77	11.79	4.45	0.25	0.0253
ECARTS-TYPES	0.0047	0.0025	0.0015	0.0007	0.0002	8.46	6.18	4.67	2.72	0.71	0.0073
SEUIL 0.05	0.0211	0.0116	0.0062	0.0027	0.0006	70.10	37.67	20.58	10.12	2.12	0.0380

LONG.= 8	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0169	0.0096	0.0051	0.0024	0.0006	48.72	27.93	14.62	6.93	1.63	0.0346
MEDIANES	0.0162	0.0095	0.0049	0.0021	0.0006	47.58	27.57	14.52	6.55	1.77	0.0344
ECARTS-TYPES	0.0049	0.0030	0.0020	0.0012	0.0005	7.34	5.22	4.09	2.56	1.24	0.0066
SEUIL 0.05	0.0251	0.0150	0.0083	0.0043	0.0016	60.74	36.67	20.95	11.20	3.75	0.0486

LONG.= 10	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0197	0.0120	0.0073	0.0039	0.0015	44.58	27.00	16.37	8.76	3.29	0.0444
MEDIANES	0.0190	0.0116	0.0067	0.0037	0.0013	44.23	27.09	16.12	8.61	3.01	0.0433
ECARTS-TYPES	0.0049	0.0034	0.0024	0.0015	0.0009	6.15	4.17	3.49	2.80	1.83	0.0047
SEUIL 0.05	0.0273	0.0181	0.0113	0.0061	0.0031	55.52	33.68	22.04	13.08	6.59	0.0600

LONG.= 12	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0228	0.0147	0.0092	0.0053	0.0025	42.05	26.98	16.80	9.66	4.51	0.0545
MEDIANES	0.0224	0.0141	0.0088	0.0049	0.0021	41.62	26.76	16.45	9.23	4.40	0.0516
ECARTS-TYPES	0.0053	0.0038	0.0030	0.0019	0.0013	5.14	3.85	3.10	2.42	1.90	0.0121
SEUIL 0.05	0.0321	0.0208	0.0145	0.0091	0.0049	50.35	33.08	22.39	13.57	8.08	0.0777

LONG.= 14	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0269	0.0172	0.0110	0.0067	0.0035	41.12	26.28	16.88	10.38	5.35	0.0654
MEDIANES	0.0257	0.0167	0.0111	0.0068	0.0035	40.59	26.37	17.06	10.02	5.13	0.0641
ECARTS-TYPES	0.0071	0.0046	0.0033	0.0021	0.0015	5.33	3.52	2.98	2.37	1.94	0.0146
SEUIL 0.05	0.0390	0.0250	0.0169	0.0099	0.0060	49.73	32.54	22.67	14.08	9.26	0.0910

LONG.= 16	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0310	0.0205	0.0136	0.0090	0.0045	39.46	26.19	17.28	11.35	5.72	0.0786
MEDIANES	0.0297	0.0202	0.0134	0.0088	0.0042	38.24	26.09	17.18	11.23	5.28	0.0765
ECARTS-TYPES	0.0063	0.0042	0.0030	0.0025	0.0019	5.21	3.73	2.67	2.47	2.13	0.0120
SEUIL 0.05	0.0412	0.0281	0.0180	0.0134	0.0080	48.97	33.16	21.48	15.60	9.60	0.0951

LONG.= 18	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0330	0.0219	0.0151	0.0102	0.0059	38.31	25.46	17.57	11.91	6.65	0.0862
MEDIANES	0.0318	0.0213	0.0149	0.0098	0.0057	37.61	25.39	17.67	11.84	6.88	0.0824
ECARTS-TYPES	0.0071	0.0044	0.0031	0.0025	0.0021	4.59	3.09	2.52	2.15	2.07	0.0143
SEUIL 0.05	0.0457	0.0302	0.0202	0.0143	0.0097	46.02	30.32	21.82	14.92	10.40	0.1140

LONG.= 20	VP1	VP2	VP3	VP4	VP5	PC1	PC2	PC3	PC4	PC5	TRA
MOYENNES	0.0349	0.0246	0.0174	0.0117	0.0070	36.51	25.69	18.24	12.24	7.33	0.0955
MEDIANES	0.0340	0.0236	0.0172	0.0110	0.0068	35.48	25.61	18.14	12.23	7.09	0.0937
ECARTS-TYPES	0.0072	0.0054	0.0034	0.0029	0.0022	4.32	3.10	2.25	2.17	2.14	0.0158
SEUIL 0.05	0.0456	0.0345	0.0231	0.0169	0.0112	43.59	30.74	22.15	15.33	10.92	0.1220

Figura 3.6 - Estimação das médias, medianas e desvios padrões das cinco primeiras raízes características, taxas de inércia e do traço (nível de 5%).

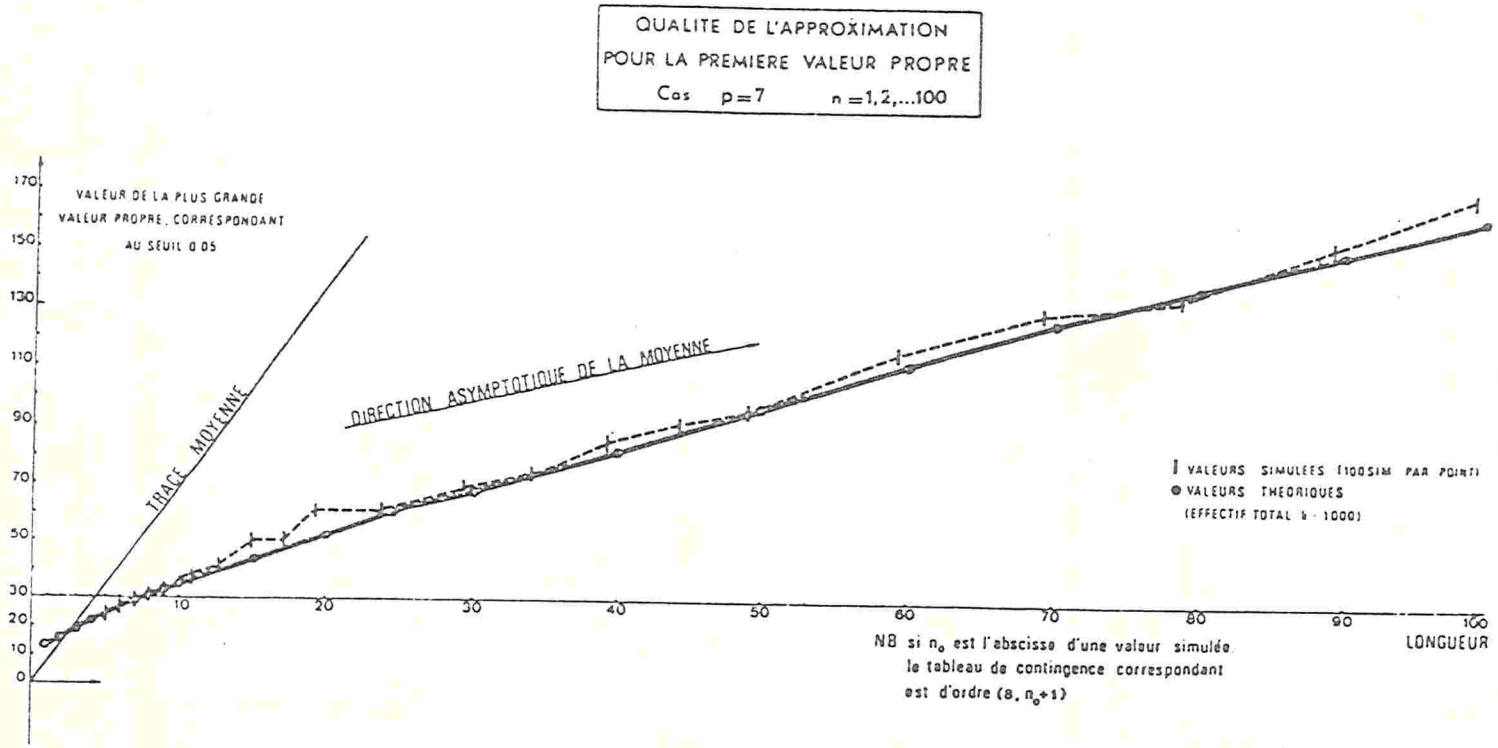


Figura 3.7 - Distribuição teórica e distribuição empírica obtida por simulação da primeira raiz característica.

CAPÍTULO 4

APLICAÇÕES NUMÉRICAS

Neste capítulo, apresentamos aplicações das técnicas de Análise de Correspondência Binária e Múltipla. Fazendo também comparações entre as duas técnicas , discutimos a validade dos resultados obtidos com o uso dessas técnicas.

Os dados aos quais aplicamos as duas técnicas foram tirados da Tese de Doutoramento do Professor Carlos Roberto Azzoni, da Faculdade de Economia e Administração da Universidade de São Paulo, sob o título: "Motivações dos Empresários e Processo de Decisão Locacional (a Experiência das Indústrias Paulistas)". Esta tese foi baseada no trabalho apresentado em 1981, em uma publicação do Governo do Estado de São Paulo – Secretaria do Interior – Coordenadoria de Ação Regional: "Fatores Locacionais da Indústria e o Desenvolvimento Regional no Estado de São Paulo".

Os resultados foram obtidos através do uso de programas de computador destinados à Análise de Correspondência e implantados no Centro de Computação Eletrônica da Universidade de São Paulo.

4.1 - APRESENTAÇÃO DAS VARIÁVEIS

Os dados da pesquisa foram colhidos no Estado de São Paulo em 1980. O objetivo do trabalho é fazer um estudo do pro-

cesso de escolha do local em que são instaladas as empresas pesquisadas, relacionando este processo com as características das empresas, tentando avaliar em que medida empresas de diferentes características apresentam comportamentos distintos. Aplicamos a técnica de Análise de Correspondência para descrever, de uma maneira global, as associações existentes entre tipos de comportamento na escolha locacional e características das empresas.

No total foram pesquisadas 566 empresas e as variáveis de estudo foram:

Variáveis locacionais:

Y1 - Número de regiões consideradas: refere-se ao número de regiões consideradas durante o processo da escolha locacional.

Categorias: Y11 - zero
Y12 - 1 a 2
Y13 - 2, 5 e mais.

Y2 - Número de alternativas consideradas: refere-se ao número de alternativas consideradas pelas empresas dentro da região escolhida.

Categorias: Y21 - zero
Y22 - 1 a 3
Y23 - 3,5 e mais.

Y3 - Elaborou perfil de localização?: refere-se aos requisitos que deveriam ser preenchidos pela localização, uma espécie de perfil de localização ideal.

Categorias: Y30 - Não

Y31 - Sim-

Y4 - Duração do processo de escolha: refere-se ao período gasto pela empresa para se decidir por um local.

Categorias: Y41 - 0 a 5 meses

Y42 - 6 a 11 meses

Y43 - 12 meses e mais.

Y5 - Fez consulta a Instituições oficiais de desenvolvimento?: refere-se às consultas feitas pelas empresas para obter orientação na escolha do local.

Categorias: Y50 - Não

Y51 - Sim.

Características das Empresas:

X1 - Propriedade do imóvel: espera-se que para imóveis próprios a imobilização de recursos é maior e para imóveis alugados é menor.

Categorias: X10 - Alugado

X11 - Próprio.

X2 - Tamanho (em número de empregados): espera-se decisões mais criteriosas de empresas maiores.

Categorias: X20 - 0 a 39

X21 - 40 a 79

X22 - 80 a 149

X23 - 150 e mais.

X3 - Grau de vinculação: empresas pertencentes a companhias ou grupos com várias unidades já contam com a esperiência do grupo na escolha do local.

Categorias: X30 - Não Independente

X31 - Independente

X4 - Área construída: espera-se que quanto maior a área construída mais acurada deva ser a análise.

Categorias: X41 - menos de 2.000 m^2

X42 - 2.000 m^2 ou mais.

X5 - Localização: há razões para se esperar um comportamento locacional conservador por parte das empresas.

Categorias: X51 - Área A (até 50 km da Capital)

X52 - Área B (entre 50 e 150 km da Capital)

X53 - Outras.

X6 - Tipo de Edifício: o fato de se utilizar um prédio já existente ou se construir um edifício especialmente para a empresa indica o grau de comprometimento de recursos.

Categorias: X60 - Já existente

X61 - Especialmente construído para a empresa.

X7 - Evolução passada das vendas do setor: evolução das vendas do ramo em que opera a empresa nos cinco anos que antecederam a entrevista.

Categorias: X70 - Abaixo da Média

X71 - Acima da Média

X8 - Tecnologia: espera-se que as empresas que utilizam tecnologia mais moderna sejam as que têm comportamento locacional mais apurado.

Categorias: X80 - Abaixo da Média

X81 - Acima da Média.

X9 - Tipos de estabelecimento: também influencia no comportamento locacional.

Categorias: X91 - Novo

X92 - Mudança

X93 - Instalação de Filial.

A seguir, primeiramente aplicamos a técnica de análise de correspondência binária e posteriormente a técnica múltipla, em ambos os casos tendo como objetivo descobrir associações existentes entre o processo de escolha locacional das empresas e suas características.

4.2 - ANÁLISE DE CORRESPONDÊNCIA BINÁRIA

Consideramos a tabela de contingência em que as características das empresas aparecem nas colunas e as variáveis locacionais aparecem nas linhas. Considerando-se que cada categoria de cada variável corresponde a uma linha (ou coluna) temos uma tabela com 13 linhas e 22 colunas (ver Tabela 4.1).

A Tabela 4.2 mostra que o primeiro fator, extraído pela análise, explica 68,79% da variância total das variáveis e, que o segundo fator explica 14,61% desta mesma variância. Assim sendo, os dois primeiros fatores explicam 83,40% da variância total. Vamos considerar, aqui, somente estes dois primeiros fatores.

A Tabela 4.3 mostra as coordenadas dos pontos-colunas nos dois primeiros fatores, bem como as contribuições absolutas e relativas dos pontos para estes fatores, enquanto que a Tabe-

la 4.4 mostra as coordenadas e as contribuições absolutas e relativas dos pontos-linhas.

A Tabela 4.3 mostra que somando-se as contribuições absolutas das categorias X20, X21, X22 e X23 da variável X2-tamanho temos uma contribuição de 30,4% desta variável na construção do fator 1. A seguir vem a variável X4 - àrea construída — com uma contribuição absoluta de 16,0% na construção do fator 1, ou seja, quase a metada da contribuição da variável X2. Em relação ao fator 2, a variável X5 - Localização — é a que mais se destaca, dando uma contribuição absoluta de 63,7%, sendo seguida pela variável X3 - Grau de vinculação — com uma contribuição de 19,1%.

Na Tabela 4.4 temos que a variável que mais contribui para a construção do fator 1 é a variável Y4 - Duração do processo de escolha — com 41,5%, sendo seguida pela variável Y1-Número de regiões consideradas — com 38,5%. Para o fator 2,as variáveis que mais se destacaram são Y2 - número de alternativas consideradas - com 64,6% e Y1 - número de regiões consideradas com 25%.

Quanto as contribuições relativas temos nas duas Tabelas 4.3 e 4.4 que o primeiro fator é muito mais importante que o segundo na explicação da dispersão das variáveis. Só para as variáveis X3 - Grau de vinculação e X5 - Localização — é que o fator 2 se destaca mais que o fator 1.

Através do gráfico, representado 4.1, onde os dois primeiros fatores F1 e F2 aparecem nos eixos horizontal e vertical respectivamente e todos os pontos-linhas e pontos-colunas estão plotados, vamos analisar o significado dos dois fatores.

Tabela 4.1 - Tabela de Contingência cruzando variáveis locacionais e características das empresas.

		X1		X2			X3		X4			X5			X6		X7		X8		X9		
		0	1	0	1	2	3	0	1	1	2	3	1	2	3	0	1	0	1	0	1	1	2	3
Y1	1	157	279	124	121	95	96	101	335	233	203	0	263	90	83	164	272	252	184	302	134	105	271	60
	2	22	64	11	11	22	42	29	57	24	62	0	51	29	6	25	61	50	36	48	38	10	62	14
	3	8	36	2	7	13	22	22	22	11	33	0	23	20	1	8	36	23	21	20	24	5	26	13
Y2	1	98	186	82	73	61	68	85	199	145	139	0	139	68	77	102	182	163	121	194	90	64	169	51
	2	54	92	33	36	32	45	36	110	67	79	0	81	52	13	53	93	87	59	104	42	31	97	18
	3	35	101	22	30	37	47	31	105	56	80	0	117	19	0	42	94	75	61	72	44	25	93	18
Y3	0	69	127	54	51	44	47	50	146	95	101	0	110	48	38	69	127	112	84	126	70	40	125	31
	1	118	252	83	89	86	113	102	268	173	197	0	227	91	52	128	242	213	157	244	126	80	234	56
Y4	1	136	180	96	86	67	67	84	232	173	143	0	171	81	64	136	180	185	131	210	106	90	180	46
	2	35	95	27	31	36	36	35	95	59	71	0	85	31	14	39	91	73	57	84	46	22	84	24
	3	16	104	14	22	27	57	33	87	36	84	0	81	27	12	22	98	67	53	76	44	8	95	17
Y5	0	172	334	126	129	118	134	131	375	246	260	0	302	127	77	183	323	297	209	335	171	110	322	74
	1	15	45	11	11	12	26	21	39	22	38	0	35	12	13	14	46	28	32	35	25	10	37	13

Tabela 4.2 - Raízes Características, porcentagens de variância acumuladas na análise de correspondência binária.

EDITION DES VALEURS-PROPRES

LA PREMIERE VALEUR-PROPRE (PARASITE) EST ELIMINEE 1.00000000

SOMME DES VALEURS-PROPRES ACTIVES 0.01948323

HISTOGRAMME DES PREMIERES VALEURS-PROPRES

	VALEUR-PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	0.01340289	68.79	68.79	**********
2	0.00284723	14.61	83.41	**********
3	0.00163200	8.38	91.78	*********
4	0.00097926	5.03	96.81	******
5	0.00036243	1.86	98.67	**
6	0.00014594	0.75	99.42	*
7	0.00007870	0.40	99.82	*
8	0.00003510	0.18	100.00	*
9	0.00000000	0.00	100.00	*
10	0.00000000	0.00	100.00	*

EDITION SOMMAIRE DES VALEURS-PROPRES DE 11 A 21

0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000

Tabela 4.3 - Coordenadas e Contribuições das Características das Empresas

Variável (pontos-colunas)	Coordenadas		Contribuição Absoluta		Contribuição Relativa	
	F1	F2	F1	F2	F1	F2
X10	-0,18	0,00	8,9	0,0	0,86	0,00
X11	0,09	0,00	4,4	0,0	0,86	0,00
X20	-0,23	0,00	10,7	0,0	0,98	0,00
X21	-0,12	0,03	3,2	0,9	0,85	0,05
X22	0,06	0,03	0,6	0,6	0,42	0,09
X23	0,26	-0,04	15,9	2,0	0,92	0,02
X30	0,07	-0,12	1,1	14,0	0,23	0,62
X31	-0,03	0,04	0,4	5,1	0,23	0,62
X41	-0,15	0,02	8,4	1,1	0,96	0,03
X42	0,13	-0,02	7,6	1,0	0,96	0,03
X51	0,07	0,10	2,1	24,5	0,27	0,67
X52	0,04	-0,12	0,3	12,8	0,04	0,33
X53	-0,30	-0,21	12,1	26,4	0,57	0,26
X60	-0,15	0,02	6,2	0,8	0,85	0,02
X61	0,08	-0,01	3,3	0,4	0,85	0,02
X70	-0,02	0,01	0,2	0,1	0,27	0,05
X71	0,02	-0,01	0,2	0,2	0,27	0,05
X80	-0,06	0,00	1,7	0,0	0,58	0,00
X81	0,11	0,01	3,2	0,0	0,58	0,00
X91	-0,21	0,02	7,6	0,4	0,86	0,01
X92	0,06	0,02	1,7	1,1	0,59	0,09
X93	0,05	-0,12	0,4	8,5	0,11	0,52

Tabela 4.4 - Coordenadas e Contribuições das Variáveis Locacionais

Variável (pontos-linhas/colunas)	Coordenadas		Contribuição Absoluta		Contribuição Relativa	
	F1	F2	F1	F2	F1	F2
Y11	-0,09	0,03	8,6	4,5	0,84	0,09
Y12	0,25	-0,05	14,2	3,2	0,85	0,04
Y13	0,37	-0,18	15,7	17,3	0,70	0,16
Y21	-0,08	-0,08	5,3	22,4	0,46	0,41
Y22	0,01	0,01	0,0	0,1	0,01	0,00
Y23	0,17	0,16	9,8	42,1	0,48	0,44
Y30	-0,04	-0,01	0,8	0,4	0,53	0,06
Y31	0,02	0,01	0,4	0,2	0,53	0,06
Y41	-0,14	-0,01	15,2	0,3	0,89	0,00
Y42	0,07	0,02	1,8	0,9	0,57	0,06
Y43	0,28	0,00	24,5	0,0	0,87	0,00
Y50	-0,02	0,01	0,4	0,9	0,46	0,23
Y51	0,15	-0,10	3,4	7,8	0,46	0,23

Inicialmente vamos estudar o 1º fator:começando da esquerda para a direita, ou seja, dos valores negativos para os positivos, vemos que as características aparecem na seguinte ordem: X53 - localização no interior; X20 - número de empregados menor que 40; X91 - empresa nova; X10 - imóvel alugado; X41 - área construída menor que 2.000 m^2; X60 - imóvel já existente e X21 - número de empregados entre 40 a 79. Chegando no extremo direito do gráfico encontramos: X61 - edifício novo; X11 - imóvel próprio; X81 - tecnologia acima da média; X42 - área construída maior que 2.000 m^2 e X23 - número de empregados maior que 150. Sendo assim, o primeiro fator pode ser interpretado como representando o "porte" da empresa, ou seja, quanto mais seus valores, mais complexa é a organização da empresa.

Considerando-se agora as variáveis locacionais,também da esquerda para a direita temos: Y41 - duração do processo da escolha menor que 6 meses; Y11 - nenhuma região considerada além da escolhida; Y21 - nenhuma alternativa considerada além da escolhida e Y30 - não foi elaborado perfil da localização. Chegando no extremo direito do gráfico encontramos: Y51 - consulta a instituições oficiais; Y-23 - grande número de alternativas consideradas; Y12 - número médio de regiões consideradas; Y43 - duração da escolha maior que um ano e Y13 - grande número de regiões consideradas. Logo, quanto maior é o valor no fator 1, mais elaborado é o processo de escolha.

Baseado na análise feita acima, chegamos à conclusão de que existe uma associação entre o porte da empresa e o cuida-

do na escolha locacional, no sentido de que existe um maior cuidado nas empresas de maior porte ou mais complexas e modernas e vice-versa. Esta associação entre as variáveis pode ser notada observando-se a proximidade, no gráfico, dos pontos que representam as variáveis: pontos que representam empresas cujas características são de empresa de grande porte, estão próximos de pontos que representam escolha locacional mais apurada e vice-versa. Encontrar associação entre conjuntos de variáveis é uma das características mais importantes da análise de correspondência.

Analisemos, agora, o 2º fator: considerando de baixo para cima (ordem dos valores crescentes) as características aparecem na ordem: X53 - localização no interior; X93 - filial ou membro de grupo; X30 - empresa não-independente; X52 - localização na área B e X23 - número de empregados a partir de 150. Nos valores mais altos encontram-se: X31 - empresa independente e X51 - localização na área A. Assim, concluímos que quanto maior o valor do 2º fator, maior é a proximidade da área A e mais acentuada é a independência das empresas.

Considerando-se as variáveis locacionais, também na ordem crescente dos valores do 2º fator temos: Y13 - grande número de regiões consideradas, Y51 - consultou instituições oficiais e Y21 - nenhuma alternativa considerada. E no valor mais alto do 2º fator encontramos Y23 - grande número de alternativas consideradas.

Analisando conjuntamente as características e as variá-

veis locacionais, observamos uma associação entre localização na área A e análise de muitas alternativas, enquanto que localização no interior, que está ligado com alto grau de vinculação, associa-se com grande número de regiões consideradas e poucas alternativas consideradas. Logo, empresas da área A não pesquisam outras regiões mas analisam muitas alternativas dentro da região, enquanto que empresas do interior se preocupam mais com a variedade das regiões possíveis não importando em analisar muitas alternativas na região escolhida.

Vemos, então, que através da análise do gráfico dos dois primeiros fatores cruzados conseguimos uma interpretação lógica e racional do significado destes fatores que juntos representam 83,40% da variância total de todas as variáveis em estudo. E, através destes fatores encontramos uma "linha de comportamento geral" das empresas em relação ao processo de escolha do local do estabelecimento.

4.3 - ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA

As variáveis consideradas, aqui, são as mesmas da análise binária, ou seja, 5 variáveis locacionais (de Y1 a Y5) e 9 variáveis de características das empresas(de X1 a X9). Considerando todas as modalidades de respostas possíveis, temos um total de 35 modalidades. Na análise múltipla estas 35 modalidades entram como variáveis coluna e as linhas são as 566 empresas. Temos então uma tabela do tipo disjuntiva completa com 566 linhas e 35 colunas. A tabela de contingência de Burt tem,

então, 35 linhas e 35 colunas, ou seja, cruzamos as 35 modalidades. Os resultados obtidos são apresentados a seguir.

A Tabela 4.5 mostra que o primeiro fator extraído pela análise, explica 14,46% da variância total das variáveis enquanto que o segundo fator explica 8,69% desta mesma variância. Assim, os dois primeiros fatores explicam 23,15% da variância total.

A Tabela 4.6 mostra as coordenadas e as contribuições relativas e absolutas de todas as 35 modalidades para os dois primeiros fatores. Considerando-se as contribuições absolutas, temos que, para o 1º fator, as variáveis que mais se destacam são X2 - Tamanho com 17,3%, X4 - Área construída, com 15,2% e X1 - Propriedade do imóvel com 14,2%, e para o 2º fator, destacam-se: X3 - Grau de vinculação com 37,6% e X9 - Tipo de Estabelecimento com 37,3%.

Agora, com o gráfico representado pela Figura 4.2, onde todas as 35 modalidades estão plotadas no plano formado pelos dois primeiros fatores (eixo horizontal - F1 e eixo vertial - F2), vamos interpretar os dois fatores.

Começando pelo 1º fator temos que da esquerda para a direita aparecem as características X23 - número de empregados maior que 150; X42 - área construída maior que 2.000 m^2; X81- tecnologia acima da média; X11 - imóvel próprio e X61 - edifício novo. No extremo direito do gráfico aparecem X41 - área construída menor que 2.000 m^2; X60 - imóvel já existente; X10 - imóvel alugado; X91 - empresa nova e X20 - número de empre-

gados menor que 40. Assim, quanto maior foi o valor do 1º fator, mais simples e menor é a empresa.

Para as variáveis locacionais temos no extremo esquerdo: Y13 - grande número de regiões consideradas; Y43 - duração da escolha maior que um ano; Y51 - consulta a instituições oficiais; Y12 - número médio de regiões consideradas e Y23 - grande número de alternativas consideradas. No extremo direito destacam-se Y11 - nenhuma região considerada; Y30 - não foi elaborado perfil da localização e Y41 - duração do processo de escolha menor que 6 meses. Logo, quanto maior é o valor no 1º fator, menos elaborado é o processo da escolha.

Concluímos então que existe uma associação entre o porte da empresa e o cuidado na escolha locacional, sendo que empresas de menor porte ou menos complexas têm menor cuidado na escolha locacional e vice-versa.

Analisando o 2º fator temos que para as características da empresa, no extremo inferior gráfico destacam-se: X31 - empresa independente; X92 - mudança de estabelecimento e X51 - localização na área A, enquanto que no extremo superior do gráfico destacam-se X52 - localização na área B; X53 - localização no interior; X81 - tecnologia acima da média; X30 - empresa não-independente e X93 - filial ou membro de grupo. Logo, quanto maior o valor do 2º fator maior é a proximidade da área B e interior e mais dependente é a empresa.

Para as variáveis locacionais temos no extremo inferior Y23 - grande número de alternativa consideradas no extremo su-

Tabela 4.5 - Raizes características, porcentagens de variância e porcentagens de variância acumulada da análise de correspondência múltipla

EDITION DES VALEURS-PROPRES

SOMME DES VALEURS-PROPRES ACTIVES 1.50000000

HISTOGRAMME DES PREMIERES VALEURS-PROPRES

	VALEUR-PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	0.21694858	14.46	14.46	**********
2	0.13035574	8.69	23.15	**********
3	0.12622223	8.41	31.57	**********
4	0.09235812	6.16	37.73	**********
5	0.08997783	6.00	43.72	**********
6	0.08122086	5.41	49.14	**********
7	0.07651872	5.10	54.24	**********
8	0.07348092	4.91	59.15	**********
9	0.07116869	4.74	63.90	**********
10	0.06575846	4.38	68.28	**********

EDITION SOMMAIRE DES VALEURS-PROPRES DE 11 A 21

0.06335271 0.06053831 0.05926627 0.05396201 0.05202057 0.04799660 0.03988969 0.03547215 0.02682567 0.01845754
0.01800632

Tabela 4.6 - Coordenadas e Contribuições das 35 modalidades de respostas

Variáveis (modalidades)	Coordenadas		Contribuição Absoluta		Contribuição Relativa	
	F1	F2	F1	F2	F1	F2
X10	0,94	0,39	9,5	2,8	0,43	0,08
X11	-0,46	-0,19	4,7	1,4	0,43	0,08
X20	1,01	0,19	8,2	0,5	0,33	0,01
X21	0,27	-0,11	0,6	0,2	0,02	0,00
X22	-0,20	-0,27	0,3	0,9	0,01	0,02
X23	-0,94	0,15	8,2	0,4	0,35	0,01
X30	-0,55	1,37	2,7	27,5	0,11	0,69
X31	0,20	-0,50	1,0	10,1	0,11	0,69
X41	0,72	-0,04	8,0	0,0	0,46	0,00
X42	-0,64	0,04	7,2	0,0	0,46	0,00
X51	0,04	-0,13	0,0	0,6	0,00	0,03
X52	-0,31	0,21	0,8	0,6	0,03	0,01
X53	0,33	0,17	0,6	0,3	0,02	0,01
X60	0,85	0,38	8,3	2,7	0,39	0,08
X61	-0,45	-0,20	4,4	1,4	0,39	0,08
X70	0,17	-0,11	0,5	0,4	0,04	0,02
X71	-0,22	0,14	0,7	0,5	0,04	0,02
X80	0,22	-0,19	1,0	1,3	0,09	0,07
X81	-0,41	0,36	2,0	2,4	0,09	0,07
X91	1,00	-0,18	7,0	0,4	0,27	0,01
X92	-0,23	-0,41	1,1	5,8	0,09	0,29
X93	-0,45	1,92	1,0	31,2	0,04	0,67
Y11	0,26	-0,04	1,7	0,1	0,22	0,01
Y12	-0,67	0,01	2,2	0,0	0,08	0,00
Y13	-1,24	0,41	4,0	0,7	0,13	0,01
Y21	0,23	0,25	0,9	1,8	0,05	0,07
Y22	0,02	-0,12	0,0	0,2	0,00	0,01
Y23	-0,49	-0,40	1,9	2,1	0,08	0,05
Y30	0,25	0,16	0,7	0,5	0,03	0,01
Y31	-0,13	-0,09	0,4	0,3	0,03	0,01
Y41	0,41	0,19	3,1	1,1	0,21	0,05
Y42	-0,22	-0,11	0,4	0,1	0,01	0,00
Y43	-0,85	-0,39	5,0	1,8	0,19	0,04
Y50	0,09	-0,01	0,2	0,0	0,06	0,00
Y51	-0,72	0,07	1,8	0,0	0,06	0,00

perior Y21 - nenhuma alternativa considerada e Y13 - grande número de regiões consideradas.

Vemos então que o fator 2 associa localização no interior com grande número de regiões consideradas e poucas alternativas consideradas e localização na área A com muitas alternativas consideradas na região. Vemos então que os dois primeiros fatores, que juntos representam 23,15% da variância total, dão a mesma "linha de comportamento geral" das empresas, na escolha locacional, encontrada na análise de correspondência binária.

A discussão e comparação dos resultados das análises de correspondência binária e múltipla serão apresentadas a seguir na seção de validade dos resultados.

4.4 - A VALIDADE DOS RESULTADOS

4.4.1 - Taxa de inércia como medida de informação

Pelo que foi discutido nas Seções 4.2 e 4.3, notamos que as conclusões, em relação as associações existentes entre as características da empresa e a escolha locacional, são as mesmas. Isto significa que os dois primeiros fatores da análise binária e os dois primeiros fatores da análise múltiplanos dão a *mesma informação* a respeito das variáveis em estudo. Observando as Figuras 4.1 e 4.2 temos que elas seriam semelhantes se trocássemos os sinais dos dois eixos, ou seja, as coordenadas dos pontos na Figura 4.1 são próximas das coordenadas dos pontos na Figura 4.2 se considerarmos só o valor absoluto. Isso significa que o 1º

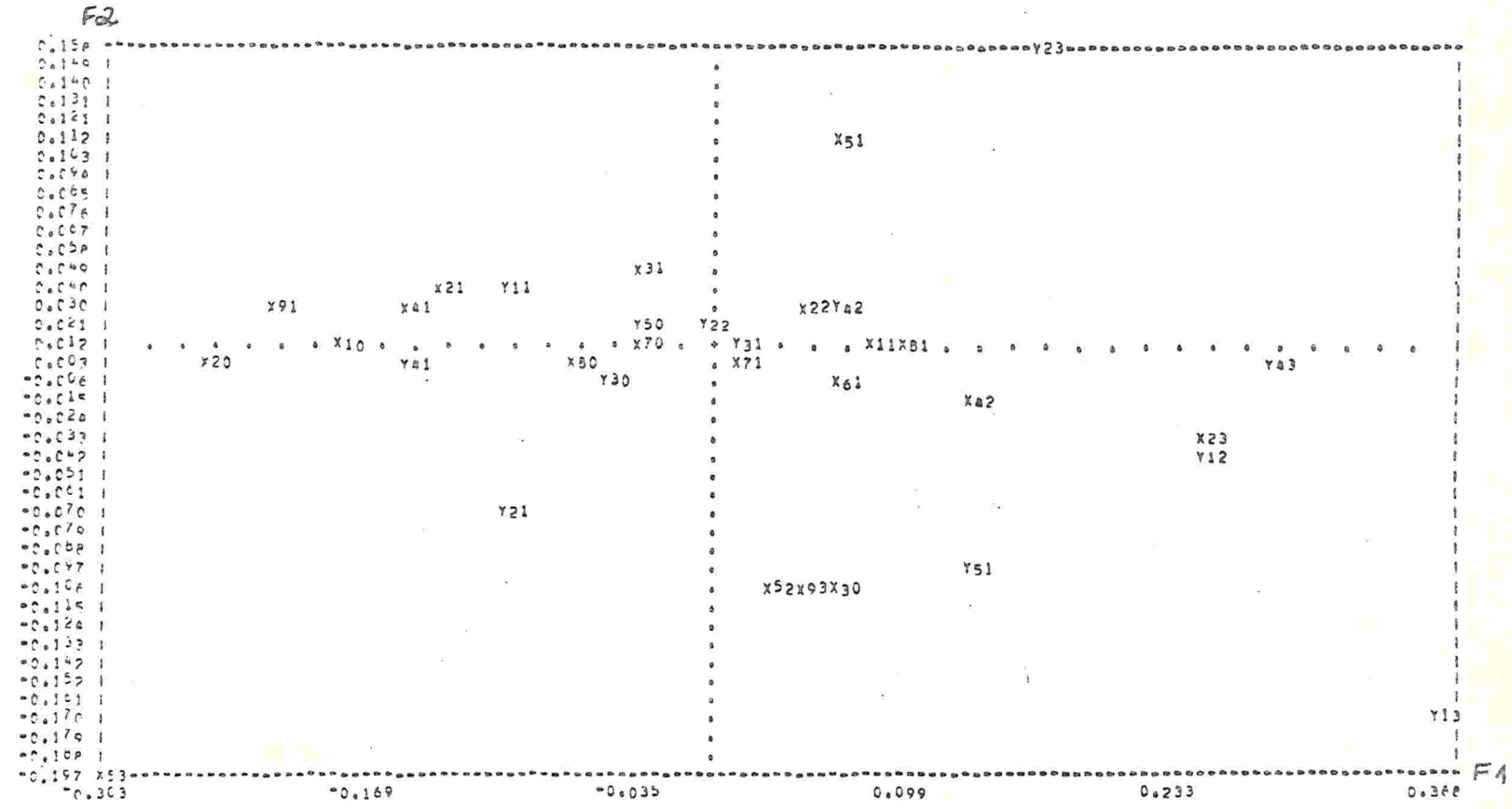
fator (F1) da análise binária é equivalente ao 1º fator da análise multipla trocando-se o sinal (-F1), o mesmo ocorrendo em relação ao 2º fator (F2).

Apesar de concluirmos que as duas análises (binária e múltipla) fornecem a mesma informação das variáveis, temos que as taxas de inércia (ou porcentagens de variância) representada pelos dois primeiros fatores na análise binária (83,40%) é muito superior à taxa de inércia representada pelos dois primeiros fatores na análise múltipla (23,15%). Temos então a mesma informação nas duas análises e taxa de inércia da análise binária quase quatro vezes maior que a da análise múltipla. Este fato demonstra o que foi discutido no Capítulo 3, quando mostramos que a taxa de inércia dá uma idéia muito pessimista da informação contida pelos fatores, principalmente no caso de tabelas com codificação disjuntiva completa, como é o caso da tabela da análise múltipla.

Considerando-se todas estas observações, podemos concluir que taxas de inércia altas implicam em muita informação das variáveis contidas pelos fatores e taxas de inércia baixas não implicam necessariamente que os fatores contêm pouca informação sobre as variáveis, mas quando a informação sobre as variáveis contida pelos fatores é pouca, a taxa de inércia é baixa.

4.4.2 - Intervalos de confiança para pontos do gráfico

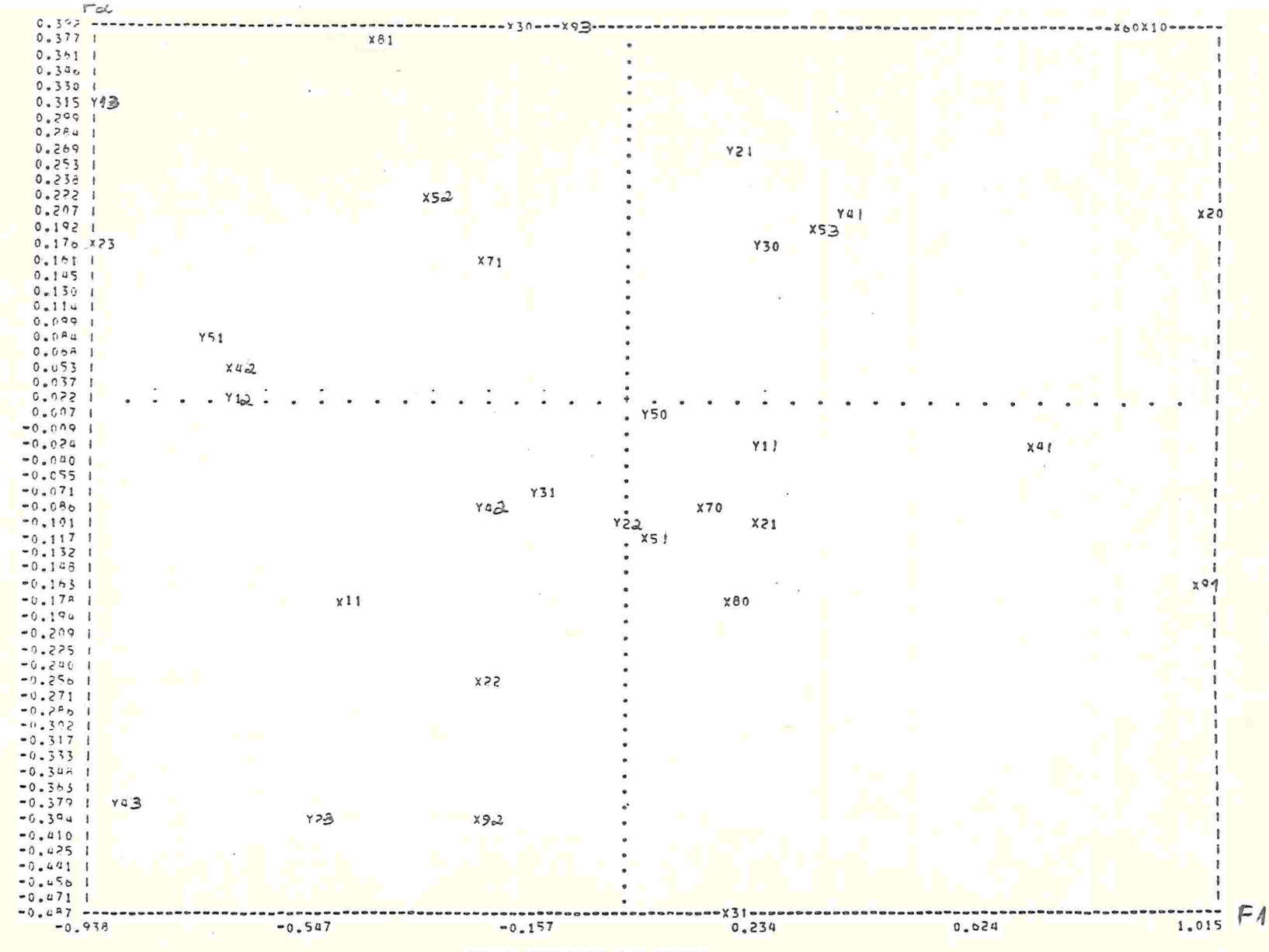
A fim de verificar quais os pontos dos gráficos que são significantemente diferentes da origem e portanto trazem infor-



POINTS MULTIPLES PAR COUPLE

POINT VU	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE	POINT CACHE	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
X41	-0.13970	0.02135	X60	-0.13970	0.02135
X22	0.05665	0.02135	X92	0.05665	0.02135

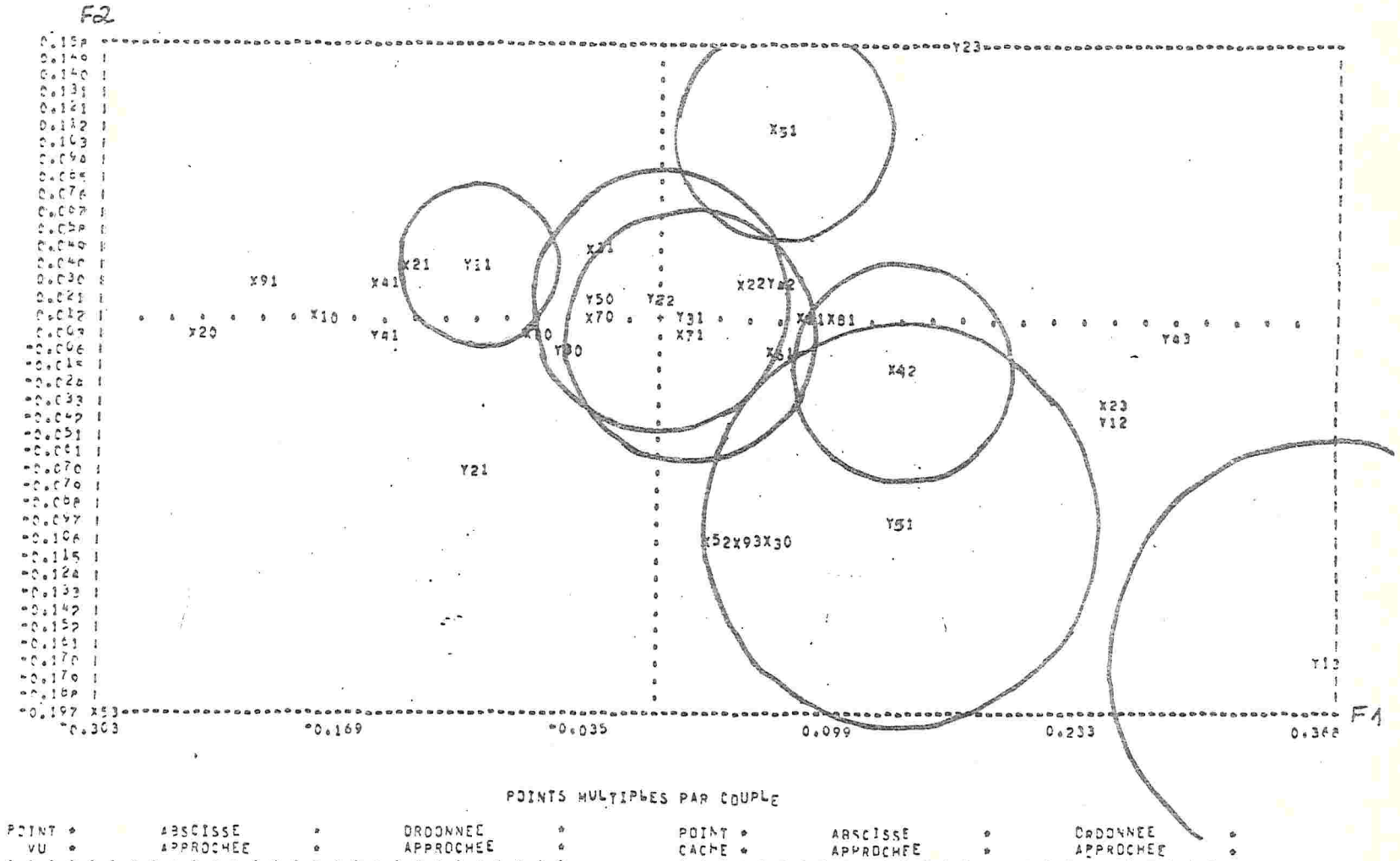
Figura 4.1 - Gráfico cruzando os dois primeiros fatores da análise de correspondência binária.



POINTS MULTIPLES PAR COUPLE

POINT VU	ABSCISSE APPROCHEF	ORDONNEE APPROCHEF	POINT CACHE	ABSCISSE APPROCHEE	ORDONNEE APPROCHEF
X11	-0.46133	-0.19389	X61	-0.46133	-0.19389

Figura 4.2 - Gráfico cruzando os dois primeiros fatores da análise de correspondência múltipla.



POINTS MULTIPLES PAR COUPLE

POINT VU	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE	POINT CACHE	ABSCISSE APPROCHEE	ORDONNEE APPROCHEE
X41	-0.13970	0.02135	X60	-0.13970	0.02135
X22	0.05665	0.02135	X92	0.05665	0.02135

Figura 4.3 - Intervalos de confiança para pontos dosgráfico cruzando os dois primeiros fatores da análise de correspondência binária

mações sobre os dados, construímos intervalos de confiança para os pontos dos gráficos conforme foi apresentado no Capítulo 3.

A título de exemplo, vamos construir estes intervalos para alguns pontos da Figura 4.1. Para isto, devemos construir um círculo de raio

$$r = \sqrt{\frac{\chi^2_{2,\alpha}}{x f_{i.}}},$$

no caso de ponto linha, ou

$$r = \sqrt{\frac{\chi^2_{2,\alpha}}{x f_{.j}}},$$

no caso de ponto-coluna, centrado no próprio ponto em questão e ver se a origem está dentro ou fora deste círculo.

Temos $\chi^2_{2,\alpha=5\%} = 5,991$.

x = total da Tabela 4.1 = 25.470.

$f_{i.}$ = freqüência relativa da linha i.

$f_{.j}$ = freqüência relativa da coluna j.

A Tabela 4.7 abaixo mostra alguns pontos com os respectivos raios:

Tabela 4.7 - Raios dos círculos para construção dos intervalos de confiança

Ponto	Raio	Ponto	Raio
Y11	0,04	X51	0,06
Y13	0,12	X42	0,06
Y51	0,11	X71	0,07
Y22	0,07		

Na Figura 4.3 vemos os círculos traçados nos pontos citados na Tabela 4.7 e notamos que só os círculos centrados em Y22 e X71 contêm a origem dos eixos, ou seja, não são pontos significantemente diferentes da origem ao nível de 5% de significância. Os outros pontos citados acima são todos significantes, ao nível de 5%, ou seja, são importantes na interpretação dos eixos pois trazem informações significantes sobre os dados.

CAPÍTULO 5

CONSIDERAÇÕES FINAIS

Este trabalho foi desenvolvido baseado nos estudos do Professor J.P.Benzécri, da Universidade de Paris VII, e seus colaboradores, a partir de 1973. Merecem destaque, entre outros, os trabalhos de Lebart, Morineau, Tabard e Fénelon, desenvolvidos entre 1973 e 1981.

A técnica de análise de correspondência foi desenvolvida exaustivamente só nos últimos anos e pesquisas nesta área estão sendo realizadas pelos professores referidos acima.

Devemos destacar também, as várias ligações entre as análises canônicas, discriminante e correspondência. A análise discriminante pode ser apresentada como um caso particular da análise canônica e a análise de correspondência como um caso particular de análise discriminante. Além disso, a análise de correspondência também pode ser vista como um caso particular da análise de componentes principais quando são feitas transformações adequadas nas variáveis e com a condição de se tratar cada um dos espaços $\mathbb{R}^n$ e $\mathbb{R}^p$ separadamente (ver Lebart, 1973).

Para finalizar, destacamos o amplo domínio de aplicação da análise de correspondência, pelo fato de não exigir qualquer restrição nos dados. É aplicável a qualquer tabela de valores numéricos positivos, e é capaz de detectar, entre outras coisas, qualquer estrutura que exista, a priori, na tabela, como

por exemplo: escalas de Guttmann e matrizes associadas a grá-
ficos particulares.

- II -

REFERENCIAS BIBLIOGRÁFICAS

[1] - ANDERSON,T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, Willey & Sons, New York,

[2] - AZZONI,C.R. (1982), *Motivações dos Empresários e Processo de Decisão Locacional (A Experiência das Indústrias Paulistas)*,Tese de Doutorado, FEA - USP, São Paulo.

[3] - BENZÉCRI,J.P. et Collaborateurs, (1973), *L'Analyse des Données, Tome II L'Analyse des Correspondances*, Dunod, Paris.

[4] - BENZÉCRI,J.P. et Collaborateurs, (1976), *L'Analyse des Données, Tome I La Taxinomie*, Dunod, Paris.

[5] - CAZES,P., SOLETY,P. & VUILLAUME,Y. (1970), Exemple du traitement statistique de donnée hydrochimiques, Extrait du *Bulletin do B. R.G.M. 2ème*, nº 4.

[6] - CORDIER,B. (1965), *L'Analyse Factorielle des Correspondances*, Thèse 3ème cycle, Paris.

[7] - ESCOUFIER,Y. (1981), *L'Analyse des Tableaux de Contingence Simples et Multiples*, Roma.

[8] - FERNANDEZ,P. & YOHAI,V. (1980), *Analisis de datos Multivariados*, Rio de Janeiro.

[9] - FLORES Jr.,R.G. (1975), Análise de Correspondência: Uma Introdução, *Rev. Bras. de Estat.*, Rio de Janeiro, Abril/Junho, pp. 177-256.

[10] - GOVERNO DO ESTADO DE SÃO PAULO (1981), *Fatores Locacionais da Indústria e o Desenvolvimento Regional no Estado de São Paulo*,Secretaria do Interior, Coordenadoria de Ação Social, São Paulo.

[11] - HILL,M.O. (1974), Correspondence Analysis: A Neglected Multivariate Method, *Applied Statistics*, nº 3, pp. 340-354.

[12] - JEFFREYS,H. (1946) - An Invariant Form for the Prior Probability in Estimation Problems, *Proc. Roy. Soc.* (A), vol. 186, pp.453-461.

[13] - JÖRESKOG,K.G., KLOVAN,J.E. & REYMENT,R.A. (1976), *Geological Factor Analysis*, Elsevier Scientific Publishing Company, Amsterdam.

[14] - KENDALL,M.G. & STUART,A. (1961), *The Advanced Theory of Statistics*, vol. 2, Griffins, Londres.

[15] - KSHIRSAGAR,A.M. (1972), *Multivariate Analysis*, Marcel Dekker Inc., New York.

[16] - LANCASTER,H.O. (1963), Canonical Correlation and Partition of χ^2, *Quart. J. Math.*, vol. 14, pp. 220-224.

[17] - LEBART,L. & FÉNELON,J.P. (1973), *Statistique et Informatique Appliquées*, 2ª ed., Dunod, Paris.

[18] - LEBART,L. (1975), *Validite des Resultats en Analyse des Données*, Paris.

[19] - LEBART,L., MORINEAU,A. & TABARD,N. (1977), *Techniques de la Description Statistique*, Dunod, Paris.

[20] - LEBART,L.& MORINEAU,A. (1981), Statistical Significance Criteria in Multiple Choice Data. Reduction and Visualization, *Psychometric Society Meeting*, Chapel Hill.

[21] - RAO,C.R. (1973), *Linear Statistical Inference and its Applications*, Wiley & Sons, New York.

[22] - VERDINELLI,M.A. (1980), *Análise Incercial em Ecologia*, Tese, USP, São Paulo.