Acta Scientiarum 21(4):861-868, 1999.
ISSN 1415-6814.

Seleção de variáveis categóricas utilizando análise de correspondência e análise *procrustes*

Terezinha Aparecida Guedes*, Ivan Ludgero Ivanqui, Ana Beatriz Tozzo Martins e Etelvina Barreto Rodrigues Cochia

Departamento de Estatística, Universidade Estadual de Maringá, Av. Colombo, 5790, 87020-900, Maringá-Paraná, Brazil.
*Author for correspondence. e-mail: taguedes@maringa.com.br e ilivanqui@uem.br

RESUMO. A análise de correspondência é uma técnica da análise multivariada (em particular, um método de análise fatorial para variáveis categóricas) que permite obter uma representação gráfica através da distribuição dos *score*s das categorias de linhas e/ou colunas em um sistema de coordenadas. Krzanowski (1987) apresentou uma metodologia que combina a análise de componentes principais e a análise *procrustes* para determinar o quanto o novo subconjunto de variáveis representa a estrutura dos dados originais Esse trabalho tem como objetivo propor um procedimento que aplica a análise *procrustes* combinada com a análise de correspondência para encontrar um *ranking* de importância para as colunas (atributos) de uma tabela de contingência. Esse procedimento foi aplicado num exemplo de Krzanowski(1993) e algumas conclusões quanto ao comportamento do procedimento são apresentadas.

Palavras-chave: análise *procrustes*, análise de correspondência, seleção de variáveis.

ABSTRACT. Selection of categorical variables using correspondence and *procrustes* analysis. Correspondence analysis is a technique of multivaried analysis (particularly, a method of factorial analysis to categorical variables) that allows us to obtain a graphic representation through the distribution of the scores from the categories of lines and/or columns in a system of coordinates. Krzanowski (1987) presented a methodology that combines the analysis of principal components with *procrustes* analysis to determine how much the new subset of variables represents the structure of the original data. The aim of this study is to propose a procedure that applies to the *procrustes* analysis combined with the correspondence analysis to find a rank of importance to the columns (variables) of a contingency table. That procedure has been applied according to an example by Krzanowski (1993). Some conclusions are presented about the behavior of the procedure.

Key words: *procrustes* analysis, correspondence analysis, selection of variables.

A análise de correspondência é uma técnica descritiva/exploratória da análise multivariada, que permite obter uma representação gráfica multidimensional da dependência entre as linhas e/ou colunas de uma tabela de contingência de duas entradas, onde as linhas e as colunas representam categorias, modalidades, de variáveis categóricas. Análise de correspondência é um método de análise fatorial para variáveis categóricas, isto é, não contínuas ou discretizadas. A representação gráfica é obtida pela distribuição de *scores* das categorias de linhas e colunas e marcando estas categorias como pontos, onde os *scores* são utilizados como as coordenadas destes pontos.

Nessa análise, uma decomposição dos dados é obtida para se estudar a estrutura dos dados sem que um modelo seja hipotetizado ou que uma distribuição de probabilidade tenha sido assumida. O objetivo principal é a representação ótima da estrutura dos dados observados. Assim sendo, as conclusões obtidas não podem ser generalizadas para a população, embora venha sendo feito na prática. Outro aspecto, discutido por Von der Heizden *et al.* (1989), é que a análise de correspondência, geralmente, é introduzida sem qualquer tratamento estatístico prévio, para dados categóricos, o que prova sua utilidade e flexibilidade. Entretanto, é aconselhável em uma análise de correspondência realizar um estudo preliminar através das marginais categóricas, diagramas de barra, por exemplo, pois a própria marginal pode revelar ausência de inércia ou riqueza de variabilidade que deveriam ser exploradas,

sempre que a freqüência de uma das modalidades supere as demais que apresentam freqüências insignificantes.

Considere uma tabela de duas entradas, em que as linhas são os agrupamentos das unidades amostrais, as colunas são as categorias ou modalidades das variáveis de interesse e as entradas são as freqüências das variáveis observadas para cada uma das unidades amostrais do agrupamento. O interesse nessa análise de dados é determinar se as categorias agrupadas (linhas da tabela) podem ser distinguidas de cada outra com base nas variáveis observadas (colunas da tabela).

Após ter sido estabelecida as diferenças entre as linhas da tabela, freqüentemente é de interesse determinar quais das variáveis medidas são responsáveis por essas diferenças, ou seja, quais das variáveis são importantes e quais são ignoráveis. Tal informação pode ser de grande valor para o entendimento do sistema ou para investigações futuras. Para a seleção de variáveis, existem vários métodos.

A seleção de variáveis, para variáveis contínuas, já foi explorado por vários autores. Jolliffe (1972 e 1973) utiliza duas variações do método do coeficiente de correlações múltiplas, quatro variações de componentes principais e duas de análise de *cluster*. Os métodos foram testados para vários conjuntos de dados artificiais e reais e nenhum se apresentou superior aos outros.

Pack e Jolliffe (1992) introduziram os conceitos de influência para investigar as mudanças na análise de correspondência quando uma única observação é adicionada, quando uma linha inteira é adicionada e quando uma linha inteira é retirada do conjunto de dados.

A redução da dimensão do conjunto de variáveis envolve distorções no relacionamento entre amostras, que podem ser medidas entre as semelhanças originais e as semelhanças no espaço reduzido. Essas medidas podem ser obtidas, principalmente, por meio da análise *procrustes,* conforme apresentada nos artigos de Krzanowski (1987, 1993) e Guedes e Ivanqui (1998).

Para enfatizar as diferenças numa tabela de contingência, Krzanowski (1993) sugeriu a utilização da análise *procrustes* na identificação de quais colunas ou variáveis do conjunto que mais contribuem. Nessa análise, procura-se o subconjunto de variáveis que melhor represente a estrutura das variáveis originais medindo a importância de cada variável.

As análises, acima propostas, envolvem a utilização de complicados algoritmos computacionais, requerendo o conhecimento de uma linguagem computacional para a elaboração de uma rotina que solucione o problema.

Nesse trabalho, será dado ênfase ao aspecto computacional da aplicação da análise *procrustes* para a criação de um *ranking* de importância para as colunas (variáveis) de uma tabela de contingência na análise de correspondência. Será utilizado um exemplo, retirado de Krzanowski (1993), e será apresentado um algoritmo para a solução através do programa estatístico SAS.

Aspecto teórico da análise de correspondência

A análise de correspondência é realizada sobre uma matriz de probabilidades ou freqüências relativas determinada a partir de uma matriz de dados, não negativos, ou tabela de contingência. A representação multidimensional da dependência entre as linhas e colunas da tabela é obtida distribuindo *scores* para as categorias das linhas e colunas e utilizando as categorias como pontos, onde os *scores* são utilizados como coordenadas desses pontos. Esses *scores* devem ser normalizados de tal forma que as distâncias entre pontos linhas e/ou pontos colunas no espaço Euclidiano sejam iguais a distância chamada qui-quadrado.

Considere uma tabela de contingência de duas entradas, $N_{IxJ} = (n_{ij})$, i = 1, ..., I e j =1, ..., J, descrevendo uma nuvem de pontos de dimensão IxJ. Seja $n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ a freqüência total absoluta. As freqüências absolutas n_{ij} devem ser transformadas em probabilidades ou freqüências relativas da seguinte forma: $p_{ij} = \frac{n_{ij}}{n}$, e a nova matriz, chamada matriz de correspondência, será denotada por: $P_{IxJ} = (p_{ij})$, i = 1, ..., I e j = 1, ..., J, cujas probabilidades marginais são calculadas por: $p_{i.} = \sum_{j} p_{ij}$ e $p_{.j} = \sum_{i} p_{ij}$.

A partir das probabilidades e das probabilidades marginais é possível definir, sobre R^J, as características de cada ponto linha na nuvem de pontos, situação que é dada pelos seguintes elementos:

perfil da linha i = $\frac{p_{ij}}{p_{i.}}$, j=1, ..., J, e

massa = $p_{i.}$.

Observe que o perfil da linha é a probabilidade condicional p(j/i) e o perfil médio da linha é

equivalente à probabilidade marginal da matriz de freqüências P_{IxJ} onde

$$\text{perfil médio da linha} = \sum_i \frac{p_{ij}}{p_{i.}} = p_{.j}.$$

Em R^I, cada ponto da coluna vem definido pelos seguintes elementos:

$$\text{perfil da coluna j} = \frac{p_{ij}}{p_{.j}}, \; i=1, ..., I, \text{ emassa} = p_{.j},$$

$$\text{perfil médio da coluna} = \sum_j \frac{p_{ij}}{p_{.j}} = p_{i.}.$$

A massa de uma linha ou coluna deve ser entendida como a importância relativa da linha ou coluna, respectivamente, dentro da tabela de dados e serve para atenuar a preponderância de alguma linha ou coluna e também para identificar cada modalidade quanto à sua importância relativa.

As distâncias qui-quadrado podem ser calculadas entre linhas bem como entre colunas. Serão consideradas as distâncias qui-quadrado entre linhas que serão calculadas dos perfis das linhas da matriz.

A distância qui-quadrado entre os perfis das linhas i e i′ é definida por:

$$\chi^2_{(i,i')} = \sum_j \frac{1}{p_{.j}} \left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}} \right)^2 \qquad (2.1)$$

Na equação (2.1), o termo $1/p_{.j}$ tem a função de diminuir a influência das colunas que têm grandes perfis marginais. A configuração dos I pontos linha fica localizada em um espaço Euclidiano de dimensão (I-1). Neste espaço, as coordenadas de N são utilizadas de forma que $d^2(x_i, x_{i'}) = \chi^2_{(i,i')}$. Os perfis das colunas sendo o perfil médio da linha é a média ponderada dos pontos linha, onde as marginais das linhas são utilizadas como pesos. Esta média ponderada está situada na origem.

Como a análise de correspondência é simétrica no sentido de resultados similares para linhas e colunas, as distâncias acima podem ser calculadas em termos das colunas da matriz.

Para testar a independência entre linhas e colunas da tabela, a estatística de teste χ^2 é calculada por:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - np_{i.}\, p_{.j})^2}{np_{i.}\, p_{.j}} \qquad (2.2)$$

A Equação (2.2) pode ser reescrita como:

$$\frac{\chi^2}{n} = \sum_i \sum_j \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_i p_{i.} \left\{ \sum_j \frac{1}{p_{.j}} \left(\frac{p_{ij}}{p_{i.}} - p_{.j} \right)^2 \right\} = \sum_i p_{i.} D^2_{i.}, \qquad (2.3)$$

onde $D^2_{i.}$ é chamada distância χ^2 (qui-quadrado) entre o i-ésimo perfil da linha e o perfil médio da linha.

Assim sendo, χ^2/n é uma soma ponderada das distâncias qui-quadrado. Nos casos em que linhas e colunas são independentes, o valor de χ^2/n será pequeno e, conseqüentemente, os valores de $p_{i.} D^2_{i.}$ também serão. Entretanto, quando existir associação significativa entre linhas e colunas, alguns dos valores $p_{i.} D^2_{i.}$ serão grandes, incluindo a parte da tabela que é responsável pela não independência.

Como já foi discutido, a análise de correspondência é um meio de se criar configurações representando as linhas da tabela por pontos no espaço, tal que a distância Euclidiana entre os pontos na configuração seja igual a distância qui-quadrado calculada entre as linhas da tabela. A equação (2.3) tem um dual em termos das distâncias $D^2_{.j}$ entre as colunas, e isto leva a uma representação dual das colunas por pontos. Em uma análise de correspondência completa é necessário construir ambas as configurações, representando linhas e colunas.

As coordenadas dos pontos nas representações podem ser obtidas como segue. Considere as matrizes N_{IxJ} e P_{IxJ} definidas no início desta seção. Seja D_l e D_c matrizes diagonais cujas entradas são, respectivamente, as proporções marginais das linhas $p_{i.}$ e colunas $p_{.j}$, assumindo que $p_{i.}>0$ e $p_{.j}>0$.

Seja $K = D_l\, tt^t\, D_c$, onde t é um vetor unitário cuja dimensão vai depender das circunstâncias. Faz-se, então, a decomposição em valores singulares da matriz

$$E = D_l^{-1/2}(P - K)D_c^{-1/2}. \qquad (2.4)$$

Os elementos dessa matriz são proporcionais aos resíduos padronizados e têm a forma:

$$e_{ij} = (p_{ij} - p_{i.}p_{.j}) / \sqrt{p_{i.}p_{.j}} \quad i = 1, ..., I \text{ e } j = 1, ..., J$$

e podem ser decompostos como:

$$E = D_l^{-1/2}(P - K)D_c^{-1/2} = U\Lambda V^t \qquad (2.5)$$

onde $U^tU = I = V^tV$ e Λ é a matriz diagonal cujos elementos são os valores singulares λ_k, em ordem decrescente, de tal forma que λ_k^2 são os autovalores de E^tE ou de EE^t e as colunas u_k de U e v_k de V são os autovetores correspondentes de EE^t e E^tE, respectivamente. Quando for definido que $D \leq \min(I-1, J-1)$ U será de ordem IxD, V é de ordem JxD e Λ de ordem DxD e é não-singular. Os valores

das linhas e colunas podem ser normalizados como segue:

$$L = D_l^{-1/2} U \qquad (2.6)$$

e

$$C = D_c^{-1/2} V . \qquad (2.7)$$

Assim, $L^t D_l L = I = C^t D_c C$ e reflete o fato de que as somas das linhas e colunas de (P-K) desaparecem; $t^t D_l L = 0 = t^t D_c C$, isto é, os *scores* das linhas e os *scores* das colunas são não correlacionadas, enquanto que para cada dimensão, ambos os *scores* têm média zero e variância unitária.

A relação entre os pontos linhas e os pontos colunas é especificada pela fórmula de transição:

$$\tilde{L} = D_l^{-1} P^t C = L\Lambda \qquad (2.8)$$

e

$$\tilde{C} = D_c^{-1} P^t L = C\Lambda , \qquad (2.9)$$

respectivamente.

Das relações (2.8) e (2.9) pode-se observar que os pontos das linhas $\tilde{L}$ são médias ponderadas dos pontos da colunas C, enquanto os pontos das colunas $\tilde{C}$ são médias ponderadas dos pontos das linhas L.

Quando a matriz $\tilde{L}$ é utilizada como as coordenadas dos pontos das linhas e a matriz $\tilde{C}$ como as coordenadas dos pontos das colunas, as distâncias entre os pontos das linhas e as distâncias entre os pontos das colunas são as distâncias qui-quadrado como na fórmula (2.2) e seu dual para as colunas.

As equações (2.8) e (2.9) são utilizadas para interpretar as distâncias entre linhas e colunas. Nos casos em que um perfil linha é igual ao perfil médio da linha, a equação (2.8) mostra que o ponto linha será a média ponderada das colunas, isto é, a origem. Se para alguma coluna o valor do perfil for maior que a média, esta coluna atrairá o ponto linha em sua direção. Multiplicando ambos valores por $p_{i.}$, nota-se que quando $p_{ij} > e_{ij}$ (o resíduo for positivo), a linha i será atraída pela coluna j e vice-versa, como é mostrado pela equação (2.9). Em geral, quanto maior a diferença $p_{ij} - e_{ij}$, mais próximos i e j estarão.

Substituído uma das equações (2.6) ou (2.7) em (2.5), obtêm-se

$$P = K + D_l L \Lambda C^t D_c = D_l (II^t + L\Lambda C^t) D_c \qquad (2.10)$$

que é conhecida como fórmula de reconstituição.

A equação (2.10) mostra que a análise de correspondência decompõe o afastamento da independência na matriz P. Assim sendo, a análise de correspondência só tem sentido quando os resíduos não são meramente resultantes da variação aleatória da independência. Na prática, freqüentemente, isso não é verificado. Para tal verificação, a equação (2.2) pode ser utilizada.

A relação entre o qui-quadrado e os valores singulares ao quadrado, em Λ^2, segue de (2.5) e (2.2), como sendo:

traço $(\Lambda^2) = \chi^2/n$.

Essa equação mostra que a análise de correspondência decompõe o valor χ^2 para testar a independência na matriz.

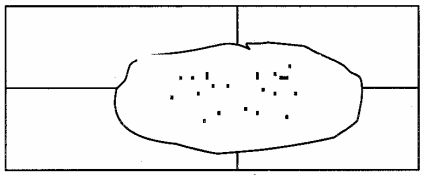
Da equação (2.8) pode-se observar que a coordenada do ponto linha i sobre o k-ésimo eixo coordenado é dada por $\lambda_k u_{ik} / \sqrt{p_{i.}}$, e de (2.9) que a coordenada do ponto coluna j sobre o k-ésimo eixo é dada por $\lambda_k u_{jk} / \sqrt{p_{.j}}$. Uma representação completa requer pelo menos min(I-1, J-1) dimensões. Assim, esta configuração é relativa aos seus eixos principais e a melhor aproximação *m*-dimensional é dada tomando as coordenadas correspondentes aos *m* maiores λ_k. Uma medida do *goodness of fit* desta aproximação é dada pela soma do traço de $\Lambda = \sum_{k=1}^{m} \lambda_k$, chamada inércia total do sistema. Portanto, procura-se um subespaço L de baixa dimensão que passa através do centro de gravidade da nuvem de pontos, isto é, o seu perfil médio, e que maximiza a inércia de N_{IxJ} paralela a L. A determinação do número de eixos envolve um complexo critério matemático, mas Benzecri (1992) (*apud* Micheloud (1997)) sugere que este número deve ser fixado pela sua capacidade de dar uma interpretação significativa para cada um dos eixos tomados.

Desta forma, a análise de correspondência pode ser entendida como um simples modelo de independência como sugere Van der Heijden *et al.* (1989) e como uma forma de mapear *a dispersão dos resultados* entre as categorias de linhas e categorias de coluna.

Representação gráfica e interpretação

Para uma linha i qualquer, passando através do centro de gravidade da nuvem de pontos, é possível decompor a inércia total da nuvem na soma da inércia paralela, projetada sobre a linha i, e da inércia perpendicular a esta linha. O primeiro eixo é a linha i para a qual a inércia paralela é máxima. O segundo eixo é, entre todas as linhas ortogonais a i, a linha para a qual a dispersão (inércia) projetada da nuvem ortogonalmente complementar a i é a maior. Assim, indo passo a passo, é possível extrair todos os eixos

que formam o novo conjunto de eixos ortogonais que descrevem totalmente a nuvem de pontos. Estes eixos são chamados eixos principais da inércia e são obtidos pela equação (2.8), considerando a ordem decrescente dos autovalores.



Observe na figura acima a nuvem de pontos que foi projetada sobre um subespaço L bidimensional passando pelo seu centro de gravidade. A projeção da nuvem fornece uma representação aproximada da mesma.

O conceito de distância qui-quadrado é utilizado na interpretação da configuração dos pontos. Quando duas linhas estão próximas, seus perfis devem ser similares e estas linhas estão relacionadas aproximadamente e da mesma forma para as colunas. Quando duas linhas estão distantes, elas estão relacionadas de modo diferente e o mesmo ocorre com as colunas. Quando um ponto linha está próximo do centro, seu perfil é similar ao perfil da coluna. Quando dois pontos linhas estão em direções opostas do centro, eles desviam em oposição aos perfis das colunas.

Aspecto teórico da análise *procrustes*

Como já foi discutido, após a análise, surge uma questão de grande interesse: qual (ou quais) das categorias, ou colunas da matriz, do conjunto contribuem mais para as diferenças entre amostras (linhas da tabela)?

Uma questão similar a esta está no contexto de componentes principais e foi respondida por Jolliffe (1972 e 1973), utilizando análise de *cluster* e por Krzanowski (1987), seguindo a utilização de análise *procrustes*.

No contexto da análise de correspondência, devido à grande afinidade matemática entre análise de correspondência e componentes principais, Krzanowski (1993) sugeriu a utilização de análise *procrustes*, e propôs uma medida *procrusteana* da importância de cada variável. Para detalhes sobre análise *procrustes,* ver Krzanowski (1987 e 1993) e Guedes e Ivanqui (1998).

Da análise de correspondência inicial, a configuração (gráfico) das linhas, obtida de todos os dados, é considerada como a configuração de referência na análise *procrustes*. As coordenadas dessa configuração são as obtidas como apresentado na Seção 2.

Para proceder à análise *procrustes*, cada coluna da matriz de dados N é omitida, uma de cada vez, e uma nova configuração de linhas é obtida para o novo conjunto de dados. A retirada de colunas de N implica que as proporções p_{ij} e, portanto, os valores e_{ij} devem ser redefinidos, mas o procedimento é exatamente igual ao anterior. Cada uma das novas configurações linhas é comparada com a configuração de referência por translação, rotação e reflexão e a soma de quadrado residual *procrustes* M^2 é calculada.

O conjunto resultante de valores M^2 apresenta um *ranking* da importância de cada coluna.

Krzanowski (1993) aponta dois pontos que devem ser considerados na aplicação dessa análise no contexto de análise de correspondência.

Primeiro, em geral, uma mudança substancial no padrão de entrada da matriz original é esperada quando as colunas são retiradas. Desta forma, n decrescerá uma quantidade considerável no decorrer da análise e os $p_{i.}$ estarão sujeitos a grandes mudanças.

Segundo, na análise de correspondência massas ou pesos são atribuídos para cada ponto da configuração. O i-ésimo ponto, isto é, a linha, tem massa $p_{i.}$, e isto leva ao argumento de que no cálculo de M^2 é mais apropriado minimizar uma soma de quadrado ponderada, ou seja, é mais fácil tolerar um erro na posição de um ponto com pouca massa que de um ponto com muita massa. Porém, não é muito claro decidir que peso será apropriado, pois as massas das linhas mudarão toda vez que uma coluna da tabela for retirada. Krzanowski (1993) sugere que sejam utilizadas como pesos as massas $p_{i.}$ calculadas da tabela completa.

Assim, cada linha de N e da matriz cuja linha foi retirada Y, será ponderada pela massa $p_{i.}$, calculada de N e M^2 e será calculado como o mínimo de:

$M^2 = \min \{ \text{traço } [D(N-Y)(N-Y)^T]\}$,

onde $D=\text{diag}\ (p_{1.}, p_{2.}, ..., p_{I.})$.

Estes dois pontos colocados por Krzanowski (1993) não estão totalmente claros, assim sendo, várias permutações de pesos e escalonamento deverão ser testados em qualquer situação.

Na utilização de M^2 como um critério para determinar a importância das colunas, a seleção das q colunas que fazem a distinção das linhas da tabela, não é simplesmente escolher as q colunas que apresentam as maiores M^2. Isto ocorre porque cada um desses valores mede a importância de uma coluna particular na presença de todas as outras, e se

uma ou mais destas são omitidas, então M^2 pode sofrer grandes mudanças.

A solução ideal para a seleção das melhores q colunas é calcular M^2 entre a configuração nova e a configuração de referência para cada escolha possível de q colunas e, então, selecionar as q colunas que correspondem aos menores valores de M^2. Isto, porém, requer um trabalho computacional muito grande. Para facilitar, o algoritmo de eliminação *backward* pode ser adotado da seguinte forma: retire a coluna que produz o menor valor M^2 e recalcule o valor de M^2 para as colunas restantes, repetindo este procedimento até que somente q colunas permaneçam.

Não é necessário fixar previamente o valor de q. Pode-se continuar retirando colunas até que um valor razoável para M^2 seja encontrado.

Aspectos computacionais da análise de correspondência utilizando a análise *procrustes*

A análise deve ser realizada utilizando o procedimento PROC CORRESP do programa estatístico SAS para a análise de correspondência e o módulo IML, também do SAS, para o cálculo de M^2.

Na implementação do procedimento de eliminação de colunas ou atributos de variáveis categóricas com base em M^2, Krzanowski (1993) sugere que se utilize tantas colunas para representar a configuração inicial ou de referência quantos forem os eixos principais na análise de correspondência necessários para acumular pelo menos 80% da variabilidade total. Nesse trabalho, notou-se que, utilizando o *ranking* formado pelos valores de qui-quadrado, calculados na análise de correspondência para o conjunto com todas as colunas, e a sugestão de Krzanowski (1993), chega-se ao melhor subconjunto representativo rapidamente, seguindo os passos descritos a seguir:

1º Passo - Realizar uma análise de componentes principais e observar quantos eixos principais são necessários para acumular pelo menos 80% da variabilidade total.

2º Passo - Realizar a análise de correspondência para o conjunto completo.

- Construir o gráfico da dim1xdim2, obtendo assim a configuração de referência.
- Classificar as colunas por ordem crescente, segundo a contribuição do qui-quadrado.

3º Passo - Realizar as análises de correspondência retirando as colunas com menor contribuição de qui-quadrado na presença das outras e calcular o valor de M^2 (comparando a configuração de referência com as novas configurações). Dessa análise, formar o *ranking* de importância de cada coluna na presença das outras.

4º Passo - Realizar a análise de correspondência retirando as colunas uma a uma conforme o *ranking* de importância e ir calculando os valores de M^2. Parar quando houver uma mudança drástica nesse valor ou até atingir o número de colunas sugeridos no passo 1, obtendo desta forma o menor conjunto de colunas representativo. As colunas que estavam na análise anterior permanecem no processo. Fazer uma análise sensitiva nos valores de M^2 para as colunas já excluídas neste passo mas com valores de qui-quadrado próximos aos das colunas selecionadas. Decidir pelo conjunto de colunas de menor M^2.

5º Passo - Construir o gráfico dim1xdim2 e comparar com o gráfico da configuração de referência.

Aplicação - principais vantagens percebidas pelos trabalhadores em suas diferentes ocupações

Este problema encontra-se em Lebart *et al.* (1984) (*apud* Krzanowski (1993)) e citado e analisado por Krzanowski (1993), no qual foram analisadas as freqüências de 17 características percebidas como sendo vantagens de seus trabalhos por respondentes agrupados em 26 categorias de trabalho. Nesse trabalho, as colunas (vantagens percebidas) serão denotadas pelas letras maiúsculas de A a Q e os tipos de trabalhos representarão os casos, isto é, as respostas de cada respondente, conforme Tabela 1 e 2.

Tabela 1. Denotação das colunas

A	Variety
B	Freedom
C	Human contact
D	Schedules
E	Salaries
F	Security
G	Family Life
H	Interesting
I	Near home
J	Good atmosphere
K	Social advantages
L	Own boss
M	I like it
N	Other
O	None
P	Out-doors
Q	No answer

Aplicando o procedimento proposto para o problema, obtém-se os seguintes resultados. No passo 1, após a análise de componentes principais, conclui-se que são necessários cinco eixos para acumular 81,5% da variabilidade total, ou seja, conforme Krzanowski (1993), serão necessários pelo menos cinco variáveis para representar a

configuração original ou de referência. No passo 2, realizou-se a análise de correspondência para o conjunto completo. Com os valores obtidos para dim1 e dim2 obteve-se a Figura 1a que representa a configuração de referência. A seguir, as variáveis foram ordenadas em ordem crescente, segundo a contribuição para o qui-quadrado, da seguinte forma: B, F, C, K, O, E, H, D, J, I, P, M, A, L, Q e N.

Tabela 2. Denotação das linhas

1	Farming-fishing
2	Farm-food industry
3	Energy-mines
4	Steel
5	Chemical-glass-oil
6	Wood-paper
7	Auto-Aviation-shipping
8	Textile-leather-shoes
9	Phamaceutical-industries
10	Manufaturing
11	Construtions
12	Food-grossery
13	Small busines
14	Miscellaneous busines
15	Administrative services
16	Telecomunications
17	Social services
18	Health services
19	Teaching-research
20	Transportations
21	Insurance-banking
22	Domestic workers
23	Other services
24	Printing-publishing
25	Private services
26	No answer

Tabela 3. Contribuições para a estatística qui-quadrado total

Variável	Qui-quadrado	Variável	Qui-quadrado
B	22,65	J	3,36
F	16,06	I	3,14
C	8,81	P	2,60
K	8,18	M	2,27
O	6,76	A	2,20
E	6,21	L	1,82
H	4,83	Q	1,69
G	4,26	N	1,64
D	3,43		

No passo 3, foram realizadas as análises de correspondência retirando as colunas (com menor contribuição de qui-quadrado – ver Tabela 3), cada uma na presença das outras e calculou-se o valor de M^2 (comparando a configuração de referência com as novas configurações). Dessa análise, foi formado o *ranking* de importância, em ordem crescente, de cada coluna na presença das outras, conforme tabela abaixo.

No passo 4, foi realizada a análise de correspondência retirando as colunas conforme o *ranking* de importância, dado na Tabela 4, e calculou-se os valores de M^2, conforme Tabela 5.

Tabela 4. *Ranking* das colunas que saem na presença das demais, segundo os M^2 em ordem crescente (decrescente em importância)

Variáveis	M^2	*Ranking*
N	0,00232	7º
Q	0,00035	3º
L	0,00025	2º
A	0,00087	5º
M	0,00323	10º
P	0,00007	1º
I	0,00319	9º
J	0,00041	4º
D	0,00434	11º
G	0,00485	12º
H	0,00311	8º
E	0,00091	6º
O	0,01649	13º

Tabela 5. Valores de M^2 para os subconjuntos de colunas

Colunas	M^2
A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, Q	0,00007
A, B, C, D, E, F, G, H, I, J, K, M, N, O, Q	0,00038
A, B, C, D, E, F, G, H, I, J, K, M, N, O	0,00140
A, B, C, D, E, F, G, H, I, K, M, N, O	0,0027
B, C, D, E, F, G, H, I, K, M, N, O	0,0074
B, C, D, F, G, H, I, K, M, N, O	0,0076
B, C, D, F, G, H, I, K, M, O	0,0164
B, C, D, F, G, I, K, M, O	0,0324
B, C, D, F, G, K, M, O	0,0586
B, C, D, F, G, K, O	0,0946
B, C, F, G, K, O[a]	0,2223
B, C, F, K, O[b]	0,3314
B, C, F, K, E[c]	0,2684
B, C, F, K	0,5468

[a] Conjunto de colunas onde se dá a mudança drástica no valor de M^2; [b] Conjunto das 5 colunas com maiores contribuições no qui-quadrado; [c] Conjunto das 4 colunas com maiores contribuições no qui-quadrado mais a coluna E, cuja contribuição para o qui-quadrado é equivalente à contribuição da coluna O.

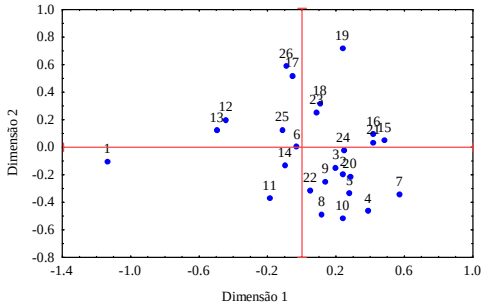


Figura 1a. Configuração de referência, (conjunto completo)

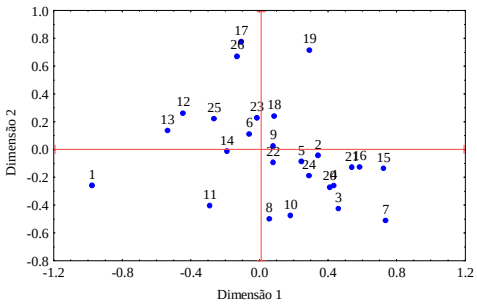


Figura 1b. Configuração com 7 colunas, selecionadas (B, C, D, F, G, K, O)

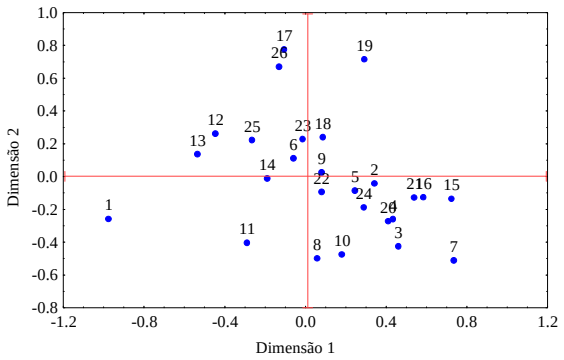


Figura 1c. Configuração com 5 colunas selecionadas (B, C, F, K, O)

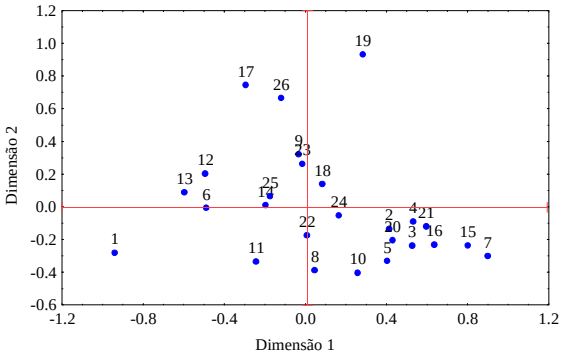


Figura 1d. Configuração com 5 colunas selecionadas (B, C, E, F, K)

Através das figuras acima, pode-se observar que as configurações são semelhantes, embora o subconjunto selecionado para a Figura 1b não seja parcimonioso. O conjunto escolhido deve ser o que apresenta a Figura 1d por apresentar um valor de M^2 inferior aos demais.

As duas dimensões utilizadas, em cada uma das figuras acima, acumulam 57,67%, 68,63%, 71,59% e 72,79%, respectivamente, da inércia total. Portanto, nota-se uma melhora na contribuição da inércia total.

Nessa análise, o conjunto com o menor valor de M^2 (B, C, E, F, K) coincide com o conjunto selecionado por Krzanowski (1993).

Discussão

Nesse trabalho, está sendo apresentado apenas uma aplicação dos resultados teóricos, embora outras análises tenham sido conduzidas para dados simulados.

Para os dados aqui analisados, pode-se observar que o subconjunto escolhido, através do critério M^2 em conjunto com a distância qui-quadrado, fornece uma configuração semelhante à configuração inicial conforme apresentado nas Figuras 1a e 1d.

A coluna E apresenta menos contribuição no valor do qui-quadrado que a coluna O, e também no *ranking* de importância foi pior classificada, porém quando substitui-se O por E no subconjunto, o valor de M^2 torna-se menor e este foi o motivo da opção por este conjunto. Vale ressaltar que, embora no *ranking* a coluna E esteja pior classificada, sua contribuição em conjunto com as colunas B, C, F e K é maior que a contribuição da coluna O.

Desta forma, observa-se que o procedimento proposto é efetivo para a redução do número de colunas para um conjunto relativamente pequeno, para o qual o pesquisador poderá avaliar pequenas mudanças nos valores de qui-quadrado que afetam grandemente o valor do resíduo M^2 e, conseqüentemente, a configuração.

Referências bibliográficas

Guedes, T.A.; Ivanqui, I.L. Análise *procrustes* aplicada à seleção de variáveis. *Acta Scient.*, *20*(4):505-509, 1998.

Jolliffe, I.T. Discarding variables in a principal components analysis, I: Artificial data. *Appl. Statist.*, *21*:160-173, 1972.

Jolliffe, I.T. Discarding variables in a principal components analysis, I: Real data. *Appl. Statist.*, *22*:21-31, 1973.

Jolliffe, I.T. Rotation of III-defined principal components. *Appl. Statist.*, *38*:139-147, 1989.

Krzanowski, W.J. Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.*, *36*:22-33, 1987.

Krzanowski, W.J. Attribute selection in correspondence analysis of incidence matrices. *Appl. Statist.*, *42*:529-541, 1993.

Micheloud, F. Correspondence analysis, www.micheloud.com, (1997).

Pack, P.; Jolliffe, I.T. Influence in correspondence analysis. *Appl. Statist.*, *41*:365-380, 1992.

Van der Heizden, P.G.M.; Falguerolles, A.; Lewm, J. A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Appl. Statist.*, *38*:249-292, 1989.

Received on September 16, 1999.
Accepted on November 29, 1999.