MEMÓRIA
CNPTIA
Com.Téc.05/99

Pesquisa Agropecuária
Agropecuária
Ifura e do Abastecimento

Comunicado Técnico da Embrapa Informática Agropecuária

Nº. 5, setembro/99 ISSN 1516-5620

APLICATIVO PARA ANÁLISE DE CORRESPONDÊNCIA

Fabiana Vittoria Patrícia Palumbo[1], José Ruy Porto de Carvalho[2], Maria Fernanda Moura[3]

Termos para indexação: Análise de correspondência; Algoritmos; SW-NTIA; Análise multivariada.
Index terms: Correspondence analysis; Algoritms; SW-NTIA; Multivariate analysis.

1. Introdução

Este trabalho está inserido em uma das etapas do subprojeto de Desenvolvimento de Algoritmos Relacionados à Metodologia de Análise Multivariada (Moran et al., 1997), que estão sendo implementados como aplicativos SW NTIA (Embrapa, 1997). A necessidade de cálculos matriciais, neste aplicativo, foi satisfeita desenvolvendo-o na linguagem de programação matricial do SW NTIA, módulo CM. Ainda, como o módulo CM não permite um tratamento de dados mais apurado, também foi utilizado o módulo GENESE, para criação de arquivos de dados e seleção de variáveis. Desta forma, os dados a serem analisados são transformados em arquivos *ntia* (Embrapa, 1997) e interpretados como uma matriz pelo módulo CM. Os resultados obtidos são apresentados em tela, podendo ser redirecionados para um arquivo ASCII, que por sua vez pode ser exportado para algum editor de dados a critério do usuário.

2. Análise de Correspondência

A análise exploratória de dados categorizados, particularmente para tabelas de contigência, data de 1933 com os trabalhos de Richardson & Kuder (1933). A partir dessa data, muitos autores tem trabalhado sobre o assunto. Entretanto, somente na década de 60, através de Benzécri (1992), com a definição de um método mostrando suas propriedades algébricas e matemáticas, na qual foi denominado de *Analyse Factorialle des Correspondences* é que esta análise começou a ter uma evolução constante. A Análise de Correspondência é uma técnica multivariada para análise exploratória de dados categorizados. Esta técnica permite a redução da matriz de dados originais para uma dimensão menor, de maneira que as associações entre linhas, entre colunas e entre linhas e colunas possam ser interpretadas. Os dados categorizados geralmente são apresentados em tabelas de contingências. Em algumas tabelas de contingências, as categorias podem ser agrupadas em classes, o que denomina-se estrutura de grupo. Estas tabelas, com estrutura em grupos, podem ter suas categorias agrupadas de acordo com os interesses do estudo. Assim, existem associações entre: linhas e agrupamento de colunas; colunas e agrupamento de linhas; e, agrupamento de linhas e colunas. Para maiores esclarecimentos sobre esta técnica consultar Benzécri (1992) e Greenacre (1984).

3. Descrição do Aplicativo

3.1. Entrada de Dados

Deve-se criar um arquivo contendo a matriz de dados iniciais; utilizando-se o módulo GENESE do SW NTIA e o seguinte roteiro:

[1] Estagiária da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
[2] Ph.D em Estatística, Pesquisador da Embrapa Informática Agropecuária. (jruy@cnptia.embrapa.br)
[3] Mestre em Engenharia Elétrica, Pesquisadora da Embrapa Informática Agropecuária.(fernanda@cnptia.embrapa.b

Tiragem:100 exemplares

Formar um arquivo ASCII não formatado contendo a matriz de dados originais.

a) Criar um arquivo programa com extensão ".gen" da seguinte forma:

```
Genese [nome da matriz]-
        Num [ nomes das colunas referentes às variáveis];
        Arquivo s = abref([ nome do arquivo que contém a matriz dos dados originais]) [nomes das colunas referentes às
        variáveis, na ordem em que se encontram no arquivo ASCII];
        {
        Leiaf(s);
        }
```

b) atualizar o nome da matriz no aplicativo CM, na segunda linha, como mostrado a seguir:

```
Cm
Matriz = leia " [ nome da matriz]";
```

3.2. Seleção de Variáveis

O programa trabalha com todas as variáveis existentes na matriz de dados. Uma sugestão para poder trabalhar com algumas variáveis é montar a matriz inicial de dados (matriz de entrada) reduzida, contendo apenas as variáveis de interesse, da seguinte forma:

```
Genese  nulo
        Num [ nomes das colunas referentes as variáveis];
        Arquivo e = (nome do arquivo ASCII não formatado de entrada) nome das variáveis na ordem em que se encontram
    no arquivo;
        Arquivo s = abref(nome da matriz,s) nome das variáveis  de interesse;
        {
            Leiaf (e);
            Grave(s);
        }
```

Os comandos anteriormente executados sob controle do GENESE permitirão gravar um arquivo apenas com as variáveis de interesse, que poderá ser lido no programa CM. Para maiores detalhes consulte o manual do GENESE (Embrapa, 1997).

3.3. Opções do Aplicativo

- Relatório de resultados: o usuário terá saída apenas em tela;
- seleção de Arquivo de Saída: se o usuário desejar gravar um Arquivo de Saída, deverá utilizar os recursos do DOS; com uso do comando "grave" do CM;
- saídas para a Ferramenta: como visto no item anterior, a ferramenta poderá gerar relatório de resultados em tela.

3.4. Algoritmo

O algoritmo implementado utiliza as formas descritas anteriormente de apresentação para os cálculos, contendo os resultados necessários para satisfazer os interesses do usuário.

Assim, o algoritmo possui as seguintes fases:

- procedimento para determinação dos eixos,
- procedimento para incluir pontos - linhas suplementares.

Os resultados são apresentados em tela, no formato de tabelas. Tanto o resultado em tela pode ser redirecionado para um arquivo quanto cada tabela pode ser transformada em um arquivo de dados. Para gravar tabelas como arquivo de dados, utiliza-se o comando "grave" do módulo CM.

3.4.1. Procedimento para Determinação dos Eixos

3.4.1.1. Parâmetros de Entrada

N_{ixj} = matriz de dados originais, onde n é qualquer matriz de dados não negativos de dimensão **i x j**, sendo que **i** é o número de linhas e **j** é o número de colunas, ou seja, $N = \{n_{ij}\}$.

3.4.1.2. Parâmetros de Saída

NT: número de elementos da matriz de dados
CM: coluna marginal da matriz de dados
LM: linha marginal da matriz de dados
P: matriz de correspondência
R: vetor de massas das linhas
C: vetor de massas das colunas
E: matriz dos valores esperados
OE: n matriz das diferenças entre os valores observados e os valores esperados
QOE: matriz onde cada elemento é o quadrado do correspondente elemento em OE
Q: matriz das contribuições totais do Qui-quadrado
RP: matriz de perfis linhas
CCC: matriz de perfis colunas
X: matriz de centralização, ponderada e transformada
Decomposição em valores singulares da matriz:
VV: matriz associada aos autovetores das linhas associados aos k valores diferentes de zero
UU: matriz associada aos autovetores das colunas associados aos k valores diferentes de zero
AA: vetor dos valores singulares
F: coordenadas principais dos perfis linhas
L: vetores com elementos ao quadrado
CIR: componentes da inércia de cada perfil em cada direção
IR: vetor de inércia de todos os perfis linhas na nuvem de perfis
CRR: contribuição relativa dos perfis linhas colunas
CAR: contribuição absoluta dos perfis linhas em cada eixo
FP: coordenadas padronizadas dos perfis linhas
G: coordenadas principais dos perfis colunas
QG: vetores com elementos ao quadrado
CIC: componentes da inércia de cada perfil coluna em cada direção
DIC: inércia dos perfis colunas
CRC: contribuição relativa dos perfis colunas em cada eixo
CAC: contribuição absoluta dos perfis colunas em cada eixo
GP: coordenadas padronizadas dos perfis colunas

3.4.1.3. Descrição

- Definição de Vetores Auxiliares
- Elementos Marginais e Total da Matriz de Dados
- Matriz de Correspondência
- Vetor de Massas das Linhas
- Vetor de Massas das Colunas
- Matriz dos Valores Esperados
- Matriz de Diferenças entre Valores Observados e Esperados
- Matriz das Contribuições Totais do Qui-Quadrado
- Definição das Matrizes Diagonais
- Matriz dos Perfis Linhas
- Matriz dos Perfis Colunas
- Matriz de Centralização, Ponderada e Transformada
- Decomposição em Valores Singulares da Matriz X

- Coordenadas Principais dos Perfis Linhas
- Vetores com Elementos ao Quadrado
- Componente da Inércia de Cada Perfil Linha em Cada Direção
- Inércia dos Perfis Linhas
- Contribuição Relativa dos Perfis Linhas em Cada Eixo
- Contribuição Absoluta dos Perfis em Cada Eixo
- Coordenadas Padronizadas dos Perfis Linhas
- Coordenadas Principais dos Perfis Colunas
- Vetores com Elementos ao Quadrado
- Componente da Inércia de cada Perfil Coluna em cada Direção
- Inércia dos Perfis Colunas
- Contribuição Relativa dos Perfis Coluna em Cada Eixo
- Contribuição Absoluta dos Perfis Colunas em cada Eixo
- Coordenadas Padronizadas dos Perfis Colunas

3.4.2. Procedimento para Incluir Pontos - Colunas Suplementares

3.4.2.1. Parâmetros de Entrada

NSC$_{ix}$ = é qualquer matriz de dados não negativos de dimensão **i x z**, sendo que **i** é o número de linhas e **z** é o número de colunas, NSC = [n_{ij}].

3.4.2.2. Parâmetros deSaída

NTCS: número de elementos da matriz de dados (NZ)

CMCS: coluna marginal da matriz de dados

LMCS: linha marginal da matriz de dados

PCS: matriz de correspondência

CS: vetor de massas das colunas

CCS: matriz dos perfis colunas

GS : coordenadas principais dos perfis colunas suplementares

QGS: vetores com elementos ao quadrado

CRCS: contribuição relativa dos perfis colunas suplementares em cada eixo

GPS: coordenadas padronizadas dos perfis linhas suplementares

3.4.2.3. Descrição

- Definição de Vetores Auxiliares
- Elementos Marginais e Total da Matriz de Dados Suplementares
- Matriz de Correspondência
- Vetor de Massas das Colunas
- Definição de Matrizes Diagonais
- Matriz dos Perfis Colunas
- Coordenadas Principais dos Perfis Colunas Suplementares
- Vetores com Elementos ao Quadrado
- Contribuição Relativa dos Perfis Colunas Suplementares em Cada Eixo
- Coordenadas Padronizadas dos Perfis Colunas Suplementares

4. Referências Bibliográficas

BENZÉCRI, J.P. *Correspondence analysis handbook.* New York: Marcel Dekker, 1992, 665p. *handbook*. New York:

EMBRAPA. Centro Nacional de Pesquisa Tecnológica em Informática para a Agricultura. Software NTIA: manuais do usuário. Campinas, 1997. 4v. não paginado.

GREENACRE, M.J. Theory and applications of correspondence analysis. New York: Academic Press, 1984. 364p.

MORAN, R.C.C.P.; MARTINEZ, E.; CARVALHO, J.R.P. de; SUGHAWARA, M.; KOMI, R. *Desenvolvimento de algoritmos relacionados a metodologia de análise multivariada.* Campinas: Embrapa Informática Agropecuária, 1997. 10p. (EMBRAPA. Programa 14. Intercâmbio e Produção de Informação em Apoio às Ações de Pesquisa e Desenvolvimento. Subprojeto 14.0.97.362-04).

RICHARDSON, M.; KUDER, G.F. Marking a rating scale that measures. *Personnel Journal,* Costa Mesa, CA, v.2, p.36-40, 1933.

IMPRESSO

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura e do Abastecimento
Rua Dr. André Tosello, s/nº Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone (19) 289-9800 Fax (19) 289-9594
E-mail: sac@cnptia.embrapa.br
http://www.cnptia.embrapa.br

GOVERNO FEDERAL