Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura e do Abastecimento

Comunicado Técnico da Embrapa Informática Agropecuária

Nº. 6, outubro/99 ISSN 1516-5620

Ministério da Agricultura e do Abastecimento

APLICATIVO PARA ANÁLISE DE COMPONENTES PRINCIPAIS

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

Termos para indexação: Componentes principais; Algoritmos; SW-NTIA; Análise multivariada.
Index terms: Principals components; Algoritms; SW-NTIA; Multivariate analysis.

1. Introdução

Este trabalho tem por objetivo descrever o aplicativo sobre Componentes Principais, o qual está inserido em uma das etapas do subprojeto 14.0.97.362-02 - Desenvolvimento de Algoritmos Relacionados à Metodologia de Análise Multivariada (Moran et al., 1997). Estes algoritmos estão sendo implementados como aplicativos do Ambiente de Software NTIA-SW NTIA (Embrapa, 1997), através da linguagem de programação matricial CM.

Como o módulo CM não permite um tratamento de dados mais apurado, também foi utilizado o módulo GENESE do mesmo software, 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 por Componentes Principais

A análise de componentes principais preocupa-se em explicar a estrutura de variância - covariância através de poucas combinações lineares das variáveis originais. Seus objetivos são a redução de dimensão da matriz de dados, facilitando assim a interpretação do fenômeno estudado (Ferreira, 1988). Embora p componentes são necessários para reproduzir a variabilidade total dos dados, freqüentemente muita dessa variabilidade pode ser explicada por um pequeno número de k ($k < p$) componentes principais. Esta técnica é valiosa como um passo intermediário em grandes investigações. Para obter mais detalhes e explicações consultar os trabalhos de Mardia et al. (1982) e Johnson & Wichern (1982).

3. Entrada de Dados para o Aplicativo

3.1. Procedimento

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

1. Formar um arquivo contendo a matriz de dados originais e criar um arquivo .gen da forma
   Genese [nome da matriz]-
   Num [nomes das colunas referentes as variáveis]
   Arquivo [nome do arquivo que contém a matriz dos dados originais]
   Leiaf [matriz];

[1] Estagiária da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campina:
[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.emb

Tiragem: 100 exemplares

2. O nome da matriz deve ser atualizado na segunda linha do arquivo de programa conforme a seguir:
   Cm
   Matriz = leia "[nome da matriz]";

3.2. Opções do Aplicativo

- **Seleção de variáveis:** para esta ferramenta, pode-se selecionar uma lista de variáveis de interesse, isto é, as variáveis para as quais se deseja usar o método de cálculo para a análise de componentes principais parciais. Nos outros casos, são utilizadas todas as variáveis;
- **seleção de métodos utilizados:** o usuário deve optar pelos métodos a serem utilizados;
- **opções de 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 MS-DOS;
- **saídas para o aplicativo:** como visto no item anterior, a ferramenta poderá gerar relatório de resultados impresso na tela.

4. Especificação de Algoritmos

4.1. Descrição de Método

Este documento cobre todos os algoritmos envolvidos na análise de componentes principais. Os componentes principais de um conjunto de dados podem ser calculados através das seguintes técnicas:

- análise componentes principais utilizando a matriz de variâncias e covariâncias;
- análise componentes principais utilizando a matriz de correlação;
- análise componentes principais utilizando a matriz parcial (neste caso o usuário escolhe as variáveis que deseja utilizar).
- análise componentes principais parciais utilizando a matriz de variâncias e covariâncias;
- análise componentes principais parciais utilizando a matriz correlação.

Estas técnicas diferem basicamente pela matriz de dados utilizada e pela forma de cálculo das componentes principais. Nos dois primeiros casos, utiliza-se a matriz completa e, no terceiro caso, apenas algumas variáveis escolhidas pelo usuário para formar uma matriz parcial. Além disso, em todos os casos podemos efetuar os cálculos utilizando a matriz de variâncias e covariâncias ou a matriz de correlação. A diferença se dá pela matriz utilizada no cálculo dos autovetores e autovalores.

4.2. Cálculo usando a Matriz de Variâncias e Covariâncias (S)

O algoritmo a seguir utiliza a matriz de dados completa e encontra os componentes principais através da matriz de variâncias e covariâncias a qual denominamos S.

4.2.1. Parâmetros de Entrada

X_{nxp} = matriz de dados originais, onde

N = número de indivíduos (linhas);
P = número de variáveis (colunas).

4.2.2. Parâmetros de Saída

S: matriz de variâncias e covariâncias
M: vetor de médias
DP: vetor de desvios
R: matriz de correlação
b,c: autovalor e autovetor da matriz S
Matriz de resultados contendo: final, autovalor, diferença, proporção e proporção acumulada

COMP: componentes principais

COV: matriz de variâncias e covariâncias entre as componentes e as variáveis originais

CORR: matriz de correlação entre a variáveis originais e os componentes

ME: matriz de proporções acumuladas explicadas pelas componentes retiradas de cada variável original

4.2.3. Algoritmo

- Guardar n e p;
- Definir um vetor unitário, n x 1, com: $D' = (1, 1, \ldots, 1)_{1 \times p}$;
- Calcular a Matriz de variâncias e covariâncias $S = (1/(n-1))\ X'(I\ \ 1/nDD')X$, onde I é uma matriz identidade n x n;
- Calcular o vetor das Médias $M = 1/n\,X'\,D$;
- Calcular os Desvios Padrões $Dp = (\mathrm{diag}Si)^{1/2}$, $i = 1, 2, 3, \ldots, p$.
- Calcular os autovetores e autovalores utilizando algoritmos que use o fato de S ser simétrica: S simétrica e positiva definida; S simétrica e positiva semi definida.
- Ordenar os autovalores em ordem decrescente e os correspondentes autovetores ortonormalizados, nomeá-los por CP_1, $CP_2, \ldots, CP_p$.
- Montar a tabela das proporções de variações das componentes:

Componentes	Autovalor	Proporção	Acumulada
CP_1	λ_1	$\lambda_1/\Sigma^p_{i=1}\lambda_i$	$\lambda_1/\Sigma^p_{i=1}\lambda_i$
CP_2	λ_2	$\lambda_2/\Sigma^p_{i=1}\lambda_i$	$\lambda_1+\lambda_2/\Sigma^p_{i=1}\lambda_i$
.	.	.	.
.	.	.	.
.	.	.	.
CP_p	λ_π	$\lambda_\pi/\Sigma^p_{i=1}\lambda_i$	1

- Imprimir as componentes principais CP1, CP2,............,CPp por coluna, nomeando em cada linha as p variáveis de entrada, ou seja:

Variáveis	CP_1	....	CP_p
Var. 1	γ_{11}	....	γ_{1p}
.	.	.	.
.	.	.	.
.	.	.	.
Var. p	γ_{p1}	....	γ_{pp}

Onde: Γ

$$= \begin{bmatrix} \gamma_{11} \cdots\cdots\cdots \gamma_{1p} \\ \gamma_{21} \cdots\cdots\cdots \gamma_{2p} \\ \cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots \\ \gamma_{p1} \cdots\cdots\cdots \gamma_{pp} \end{bmatrix}$$

- Calcular a matriz de variâncias e covariâncias entre as componentes e as variáveis originais:
  COV = $\Gamma\lambda$ onde Γ é a matriz dos autovetores e λ é a matriz diagonal dos autovalores.
  Logo, COV = (cov_{ij}), cov_{ij} = cov(var_i, Cp_j).
- Matriz de correlação entre as variáveis originais e as componentes principais:
  CORR = $(DP)^{-1}$ COV$(diag\lambda)^{-1/2}$
- Calcular a matriz das proporções acumuladas explicadas pelas componentes retiradas, de cada uma variável original:

$$ME = \begin{bmatrix} corr_{11}^2 & corr_{11}^2 + corr_{12}^2 .... corr_{11}^2 + ... + corr_{1p}^2 \\ corr_{21}^2 & corr_{21}^2 + corr_{22}^2 .... corr_{21}^2 + ... + corr_{2p}^2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ corr_{p1}^2 & corr_{p1}^2 + corr_{p2}^2 .... corr_{p1}^2 + ... + corr_{pp}^2 \end{bmatrix}$$

- Calcular os escores dos indivíduos nas componentes:
  ESCORES = (X-DM')Γ onde Γ é a matriz de autovetores
  $ESCORES_{n \times p}$, indivíduos por linha e valores da CP_i no indivíduo i por coluna, i = 1,...,p.

4.3. Cálculo usando a Matriz de Correlação (R)

O algoritmo é semelhante ao anterior, entretanto utiliza-se a matriz de correlação em lugar da matriz de variâncias e covariâncias no cálculo de autovalores e autovetores. São os mesmos parâmetros de entrada e de saída. O algoritmo calcula a matriz de correlação:

R = (DP)-1*S*(DP)-1 , onde S = 1/(N-1)*X (I-(1/N)DD')*X

Segue o algoritmo anterior com R (matriz de correlação) no lugar de S (matriz de variâncias e covariâncias).

4.4. Cálculo usando a Matriz Parcial

Nos algoritmos a seguir, as variáveis utilizadas são escolhidas previamente, ou seja a matriz de dados será uma nova matriz denominada de matriz parcial.

Do mesmo modo que os algoritmos anteriores temos dois métodos, um utilizando a matriz de variâncias e covariâncias e outro a matriz de correlação.

4.4.1. Utilizando a Matriz de Variâncias e Covariâncias (S)

4.4.1.1. Parâmetros de Entrada

$X_{n \times p}$ = matriz de dados originais, onde: N = número de indivíduos (linhas); P = número de variáveis (colunas)

$X1_{q \times n}$ = matriz formada pelas variáveis principais

$X2_{(p-q) \times n}$ = matriz formada pelas variáveis restantes

4.4.1.2. Parâmetros de Saída

S: matriz de variâncias e covariâncias
S particionada em blocos, onde

S11 = matriz de variâncias e covariâncias das q variáveis restantes
S12 = matriz de variâncias e covariâncias entre as variáveis restantes e principais
S22 = matriz de variâncias e covariâncias entre variáveis consideradas parciais

SP: matriz de variâncias e covariâncias removendo o efeito das parciais

DP: vetor de desvios

DPC: vetor de desvios parciais

M: vetor de médias

MP: vetor de médias parciais

R: matriz de correlação

b,c: autovalor e autovetor da matriz **SP**

Matriz de resultados contendo: final, autovalor, diferença e proporção

COMP: componentes principais

COV: matriz de covariância entre as componentes e as variáveis originais

CORR: matriz de correlação entre a variáveis originais e os componentes

ME: matriz de proporções acumuladas explicadas pelas componentes retiradas de cada variável original

Escores onde **escores1**: escores corrigidos e **escores2**: escores não corrigidos.

4.4.1.3. Algoritmo

Ao entrar com as p variáveis var1,...,varp e selecionar as que serão consideradas parciais:

$$\begin{bmatrix} \text{var}1 \\ . \\ . \\ . \\ \text{var}\,q \\ \text{var}\,q+1 \\ . \\ . \\ . \\ \text{var}\,p \end{bmatrix}$$

onde q + 1...............varp são consideradas parciais

calcular S = X'*(I - 1/n*D*D')*X

particionar S em blocos correspondentes a q e p.

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

Onde: S11q x q matriz de variâncias e covariâncias das variáveis

S12q x (p-q) matriz das covariâncias entre as variáveis e as variáveis consideradas parciais

S22(p-q) x (p-q) matriz de variâncias e covariâncias das variáveis consideradas parciais

- Cálculo da matriz de variâncias e covariâncias removendo o efeito das parciais:
  SP = S11 - S12 - S12S21

- Substituir Sp no lugar de S e proceder aos cálculos.

- Para calcular o vetor das médias obter:
  Onde M1 é q x 1 e M2 é (p-q) x 1.

$$M = (\tfrac{1}{n}) * X' * D = \frac{M_1}{M_2}$$

- Calcular o vetor de médias parciais:
  MP = M1 S12S22-1M2, onde Mp é q x 1

- Para cálculo dos escores corrigidos, fazer:
  ESCORES = (X1 D(MP'))

- Para o cálculo dos escores não corrigidos, fazer:
  ESCORES = (X1 M1)

4.4.2. Utilizando a Matriz de Correlação (R)

Temos os mesmos parâmetros de entrada e saída conforme a situação anterior e pequena alteração no algoritmo, conforme especificado a seguir:

- Calcular DPC = (diagSP)1/2

- CORRP = (DPC)-1SP(DPC)-1

- Formar R no lugar de S e proceder a partir do cálculo para parciais.

5. Referências Bibliográficas

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

FERREIRA; A. M. *Apostila de técnicas multivariadas*. São Carlos: Universidade Federal de São Carlos, SP. Departamento de Estatística, 1988.

JOHNSON, N.R.; WICHERN, D.W. *Aplied multivariate statistical analysis*. 2.ed. Prentice-Hall, 1982.

MARDIA, K.V.; KENT, J.T,; BIBBY, J.M. *Multivariate analysis*. London: Academic Press/Harcourt Brace Jovanovich, 1982.

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 de Informação em Apoio às Ações de Pesquisa e Desenvolvimento. Subprojeto 14.0.97.362-04).

IMPRESSO

Embrapa

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