ANÁLISE DE VARIÁVEIS CANÔNICAS

Carlos Alberto Alves Varella[1]

ÍNDICE

[1] Professor. Universidade Federal Rural do Rio de Janeiro, IT-Departamento de Engenharia, BR 465 km 7 - CEP 23890-000 – Seropédica – RJ. E-mail: varella@ufrrj.br.

INTRODUÇÃO

A análise de variáveis canônicas é uma técnica da estatística multivariada que permite a redução da dimensionalidade de dados, é semelhante a componentes principais e correlações canônicas. Essa técnica é especialmente empregada em análises discriminantes realizadas a partir de amostras com observações repetidas. A análise também pode ser utilizada para representar várias populações em um subespaço de menor dimensão. A análise procura, com base em um grande número de características originais correlacionadas, obter combinações lineares dessas características denominadas variáveis canônicas de tal forma que a correlação entre essas variáveis seja nula (KHATTREE & NAIK, 2000). A utilização dessa técnica permite capturar o efeito simultâneo de características originais e com isso pode capturar variações não percebidas quando do uso de características originais isoladamente. É importante observar que a primeira variável canônica é a função discriminante linear de Fisher. Variáveis canônicas são funções discriminantes ótimas, ou seja, maximizam a variação entre tratamentos em relação à variação residual. A variação de tratamentos, nesta análise, é expressa por uma matriz denominada **H**, composta pela soma de quadrados e produtos de tratamentos; a variação residual é expressa pela matriz **E**, composta pela soma de quadrados e produtos do resíduo. As matrizes **H** e **E** são obtidas de uma análise de variância multivariada: MANOVA.

DIMENSIONALIDADE DAS VARIÁVEIS CANÔNICAS

A dimensionalidade é o número de variáveis canônicas obtidas na análise. Pode ser entendida como o número de raízes não nulas da Equação1.

$$|H - \lambda \cdot n_e \cdot \Sigma| = 0 \tag{1}$$

A dimensionalidade, portanto, é a ordem do hiperplano gerado pelas diferentes médias de tratamentos. A dimensionalidade, em termos das médias populacionais, é o número de autovalores não nulos da matriz Λ da Equação 2.

$$\Lambda = E^{-1} \cdot H = \left|ne \cdot \Sigma\right|^{-1} \cdot H \tag{2}$$

em que,

Λ = matriz determinante;
E = matriz de soma de quadradros e produtos de resíduo;
H = matriz de soma de quadrados e produtos de tratamentos;

n_e = número de graus de liberdade do resísuo;
Σ = matriz de covariância.

Teste de dimensionalidade

Quando a dimensionalidade é igual a zero (d=0) as médias são coincidentes, se d=1 as médias são colineares e se d=2 as médias são perpendiculares, isto é independentes. Numa análise de variância variânica multivariada com k tratamentos, usualmente testamos a hipótese:

$$H_0: \mu_1 = \cdots = \mu_k$$

A hipótese que testamos é se os vetores de médias são iguais. Esta hipótese é equivalente ao teste de que não há diferença entre os vetores de médias de tratamentos, isto é:

$$H_0: t_1 = \cdots = t_k$$

Se H_0 é verdadeira, concluímos que os vetores $\mu_1, \ldots, \mu_k$ são idênticos. Então H_0 verdadeira implica em d=0.

Se H_0 é rejeitada, é de importância se determinar a real dimensionalidade d, onde d=0, ... , t. Se d=t não há nenhuma restrição sobre os vetores de médias, e d<t ocorre se e somente se houver exatamente s=t-d relações linearmente dependentes entre os k vetores de médias.

Em qualquer caso tem-se que:

$$d \leq \min(p, q) = t, \text{com } q = k - 1$$

em que,

d = dimensionalidade das variáveis canônicas;
p = número de variáveis originais;
q = número de graus de liberdade de tratamentos;
t = número de vetores de médias linearmente independentes.

Considerando-se que em uma análise de variância multivariada o número de variáveis estudas normalmente é maior que número de tratamentos, a regra acima significa que: o número de variáveis canônicas será no máximo igual ao número de graus de liberdade de tratamentos.

Quando trabalhamos com dados observados, um autovalor pode ser muito pequeno sem propriamente ser nulo. Assim um teste de verificação da dimensionalidade torna-se necessário. A aproximação mais adequada, nesse caso, segundo REGAZZI (2000), é aquela proposta por BARTLETT (1947). O teste é feito sequencialmente para d=0, d=1, etc, até que um resultado não significativo apareça. Se até d-1 se obtiver resultados significativos, mas em

d não, infere-se que a dimensionalidade é d. A estatística proposta por BARTLETT (1947) é obtida através da Equação 3.

$$D_d^2 = \left(n_e - \frac{p-q+1}{2}\right) \cdot \sum_{j=d+1}^{p} \ln(1+\lambda_j) \tag{3}$$

Na Equação 3, λ_j com j=1, 2, ... , p, são autovalores da matriz Λ. A estatístca D_d^2, assintoticamente tem distribuição qui-quadrada χ_f^2 com $f = (p-q)\cdot(q-d)$.

VETORES CANÔNICOS

Vetores canônicos são os autovetores v_j associados aos autovalores λ_j não nulos da matriz determinante Λ. Seja dessa maneira, L_j o autovetor associado ao autovalor λ_j, onde L_j é normalizado de modo que:

$$L_j' \cdot \frac{E}{n_e} \cdot L_j = 1$$

Então L é o j-ésimo vetor canônico obtido na análise.

A projeção de um ponto X (observações) sobre o hiperplano estimado pode ser representada em termos de coordenadas canônicas d-dimensional

$$L_1'X, \cdots, L_d'X$$

As médias canônicas dos k tratamentos são:

$$\widehat{mc}_i = \left[L_1'\widehat{m}_i, \cdots, L_d'\widehat{m}_i\right]', i = 1, 2, \cdots, k$$

As médias canônicas representam a projeção do grupo de médias sobre o hiperplano estimado e podem ser usadas para estudar as diferenças entre grupos (tratamentos). O vetor L_j é o vetor canônico para a j-ésima variável canônica.

$$CAN_j = L_j' \cdot X$$

em que,

CAN_j = j-ésima variável canônica;
L_j' = j-ésimo vetor canônico;
X = vetor de características originais.

PORCENTAGEM DE VARIAÇÃO

A porcentagem de variação entre tratamentos explicada pelas primeiras d variáveis canônicas é o resultado da divisão da soma dos autovalores λ_d pela soma dos autovalores λ_p, isto é:

$$\mathbf{PV} = \frac{(\lambda_1 + \lambda_2 + \cdots + \lambda_d)}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

em que,

PV	=	porcentagem de variação explicada pelas primeiras d variáveis canônicas;
d	=	número de variáveis canônicas;
p	=	número de variáveis originais.

EXEMPLO DE APLICAÇÃO

Neste exemplo as análises serão realizadas com o procedimento CANDISC do programa computacional SAS (SAS, 2007).

Vamos estudar o caso em que temos k tratamentos com p variáveis e r repetições em um delineamento estatístico inteiramente casualizado. Neste caso a variância total é decomposta como segue:

$$\mathbf{A} = \mathbf{H} + \mathbf{E}$$

em que,

A	=	matriz de totais;
H	=	matriz de tratamentos;
E	=	matriz de resíduos.

A, H, e E são matrizes de dimensões *p* x *p* de somas de quadrados e produtos.

No Quadro 1 estão os valores observados das variáveis X_1 e X_2 provenientes de um delineamento estatístco inteiramente casualizado com três tratamentos e cinco repetições.

Quadro 1. Valores observados das variáveis X_1 e X_2 com cinco repetições

Tratamentos	Repetição	X_1	X_2
1	1	4,63	0,95
1	2	4,38	0,89
1	3	4,94	1,01
1	4	4,96	1,23
1	5	4,48	0,94

2	1	6,03	1,08
2	2	5,96	1,05
2	3	6,16	1,08
2	4	6,33	1,19
2	5	6,08	1,08
3	1	4,71	0,96
3	2	4,81	0,93
3	3	4,49	0,87
3	4	4,43	0,82
3	5	4,56	0,91

Procedimento CANDISC para análise de variáveis canônicas

O exercício abaixo exemplifica o uso do procedimento CANDISC do programa computacional SAS para fazer análise de variáveis canônicas dos dados apresentados no Quadro 1.

```
data exemplo;
title 'Exemplo de Análise de Variáveis Canônicas DIC';
input  trat rep X1 X2;
cards;
1  1  4.63  0.95
1  2  4.38  0.89
1  3  4.94  1.01
1  4  4.96  1.23
1  5  4.48  0.94
2  1  6.03  1.08
2  2  5.96  1.19
2  3  6.16  1.08
2  4  6.33  1.19
2  5  6.08  1.08
3  1  4.71  0.96
3  2  4.81  0.93
3  3  4.49  0.87
3  4  4.43  0.82
3  5  4.56  0.91
;
proc candisc data=exemplo out=can all;
class trat;
var X1 X2;
run;
proc plot;
plot can2*can1 = trat / vpos=20;
run;
```

Descrição dos comandos utilizados

data nome do arquivo que será utilizado na análise;

title título do cabeçalho da análise;

input define as variáveis em ordem de apresentação no arquivo;

cards é o arquivo de dados;

proc candisc é o procedimento do SAS que realiza a análise de variáveis canônicas;

out nome do arquivo para armazenar resultados da análise;

all ativa todas as funções de impressão;

class define a fonte de variação, no caso tratamentos;

var são as variáveis independentes, neste caso X_1 e X_2;

run processa os comandos anteriores;

proc plot ajusta diversos parâmetros para plotagem de gráficos;

plot define variáveis para plotagem;

=trat plota a dispersão em função de tratamentos;

vpos=20 localiza o gráfico na posição central.

Interpretação dos resultados do SAS

```
Exemplo de Análise de Variáveis Canônicas DIC                          16
21:59 Thursday, March 28, 2007

                        The CANDISC Procedure O Procedimento CANDISC

Observations      15          DF Total                 14 GL total
Variables          2          DF Within Classes        12 GL de resíduo
Classes (trat)     3          DF Between Classes        2 GL de tratamentos


                        Class Level Information Probabilidades a priori

            Variable
trat    Name        Frequency       Weight    Proportion

1    _1                   5       5.0000      0.333333
2    _2                   5       5.0000      0.333333
3    _3                   5       5.0000      0.333333


Exemplo de Análise de Variáveis Canônicas DIC                          17
                                          21:59 Thursday, March 28, 2007

                          The CANDISC Procedure
                     Within-Class SSCP Matrices

                                  trat = 1

             Variable           X1                X2

             X1            0.2784800000      0.1145400000
             X2            0.1145400000      0.0711200000

--------------------------------------------------------------------------------
```

trat = 2

Variable	X1	X2
X1	0.0806800000	0.0072600000
X2	0.0072600000	0.0145200000

---

trat = 3

Variable	X1	X2
X1	0.0988000000	0.0294000000
X2	0.0294000000	0.0118800000

The CANDISC Procedure

Pooled Within-Class SSCP Matrix Matriz **E** resíduo

Variable	X1	X2
X1	0.4579600000	0.1512000000
X2	0.1512000000	0.0975200000

Between-Class SSCP Matrix Matriz **H** tratamentos

Variable	X1	X2
X1	7.247640000	0.870100000
X2	0.870100000	0.127853333

Total-Sample SSCP Matrix Matriz **A** total

Variable	X1	X2
X1	7.705600000	1.021300000
X2	1.021300000	0.225373333

Neste caso como o delineamento estatístico é inteiramente casualiuzado (DIC) temos que:

$$E = A - E$$

The CANDISC Procedure
Within-Class Covariance Matrices Matrizes Cov dentro de trat

trat = 1, DF = 4

Variable	X1	X2
X1	0.0696200000	0.0286350000
X2	0.0286350000	0.0177800000

---

trat = 2, DF = 4

Variable	X1	X2
X1	0.0201700000	0.0018150000
X2	0.0018150000	0.0036300000

---

trat = 3, DF = 4

Variable	X1	X2
X1	0.0247000000	0.0073500000
X2	0.0073500000	0.0029700000

The CANDISC Procedure

Pooled Within-Class Covariance Matrix, DF = 12 Resíduo

Variable	X1	X2
X1	0.0381633333	0.0126000000
X2	0.0126000000	0.0081266667

Between-Class Covariance Matrix, DF = 2 Tratamentos

Variable	X1	X2
X1	0.7247640000	0.0870100000
X2	0.0870100000	0.0127853333

Total-Sample Covariance Matrix, DF = 14 Total

Variable	X1	X2
X1	0.5504000000	0.0729500000
X2	0.0729500000	0.0160980952

The CANDISC Procedure

Within-Class Correlation Coefficients / Pr > |r|

trat = 1

Variable	X1	X2
X1	1.00000	0.81389 Correlação 0.0936 Significância
X2	0.81389	1.00000 0.0936

---

trat = 2

Variable	X1	X2
X1	1.00000	0.21211 Correlação 0.7320 Significância
X2	0.21211	1.00000 0.7320

trat = 3

Variable	X1	X2
X1	1.00000	0.85814 Correlação 0.0628 Significância
X2	0.85814	1.00000 0.0628

The CANDISC Procedure

Pooled Within-Class Correlation Coefficients / Pr > |r|

Variable	X1	X2
X1	1.00000	0.71547 0.0060
X2	0.71547 0.0060	1.00000

Between-Class Correlation Coefficients / Pr > |r|

Variable	X1	X2
X1	1.00000	0.90389 0.2814
X2	0.90389 0.2814	1.00000

Total-Sample Correlation Coefficients / Pr > |r|

Variable	X1	X2
X1	1.00000	0.77499 0.0007
X2	0.77499 0.0007	1.00000

The CANDISC Procedure
Simple Statistics

Total-Sample

Variable	N	Sum	Mean	Variance	Standard Deviation
X1	15	76.95000	5.13000	0.55040	0.7419
X2	15	15.13000	1.00867	0.01610	0.1269

trat = 1

Variable	N	Sum	Mean	Variance	Standard Deviation
X1	5	23.39000	4.67800	0.06962	0.2639
X2	5	5.02000	1.00400	0.01778	0.1333

trat = 2

Variable	N	Sum	Mean	Variance	Standard Deviation
X1	5	30.56000	6.11200	0.02017	0.1420
X2	5	5.62000	1.12400	0.00363	0.0602

trat = 3

Variable	N	Sum	Mean	Variance	Standard Deviation
X1	5	23.00000	4.60000	0.02470	0.1572
X2	5	4.49000	0.89800	0.00297	0.0545

The CANDISC Procedure

Pairwise Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Squared Distance to trat

From trat	1	2	3
1	0	85.37718	1.78287
2	85.37718	0	78.72086
3	1.78287	78.72086	0

F Statistics, NDF=2, DDF=11 for Squared Distance to trat

From trat	1	2	3
1	0	97.82801	2.04287
2	97.82801	0	90.20099
3	2.04287	90.20099	0

Prob > Mahalanobis Distance for Squared Distance to trat

From trat	1	2	3
1	1.0000	<.0001	0.1760
2	<.0001	1.0000	<.0001
3	0.1760	<.0001	1.0000

The CANDISC Procedure

Univariate Test Statistics

F Statistics, Num DF=2, Den DF=12

Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
X1	0.7419	0.1954	0.8513	0.9406	15.8259	94.96	<.0001
X2	0.1269	0.0901	0.1131	0.5673	1.3110	7.87	0.0066

Average R-Square

Unweighted	0.7539318
Weighted by Variance	0.9299607

Multivariate Statistics and F Approximations MANOVA

S=2 M=-0.5 N=4.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.03142928	25.52	4	22	<.0001
Pillai's Trace	1.21304168	9.25	4	24	0.0001
Hotelling-Lawley Trace	23.03901513	61.97	4	12.235	<.0001
Roy's Greatest Root	22.69629642	136.18	2	12	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.978672	0.977020	0.011279	0.957799
2	0.505215	.	0.199045	0.255242

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

Test of H0: The canonical correlations in the current row and all that follow are zero

Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	22.6963	22.3536	0.9851	0.0314	25.52	4	22	<.0001
2	0.3427	0.0149	1.0000	0.7447	4.11	1	12	0.0654

The CANDISC Procedure

Total Canonical Structure

Variable	Can1	Can2
X1	0.987661	0.156610
X2	0.666459	0.745541

Between Canonical Structure

Variable	Can1	Can2
X1	0.996667	0.081583
X2	0.865977	0.500084

Pooled Within Canonical Structure

Variable	Can1	Can2
X1	0.832256	0.554392
X2	0.208132	0.978101

```
Exemplo de Análise de Variáveis Canônicas DIC                          29
21:59 Thursday, March 28, 2007

                              The CANDISC Procedure

                Total-Sample Standardized Canonical Coefficients

                  Variable              Can1              Can2

                  X1             5.316720535      -1.131352800
                  X2            -1.116842127       1.676610061

               Pooled Within-Class Standardized Canonical Coefficients

                  Variable              Can1              Can2

                  X1             1.399999146      -0.297907882
                  X2            -0.793525275       1.191244875


            Raw Canonical Coefficients Vetores canônicos

                  Variable              Can1              Can2

                  X1              7.16645900       -1.52496137
                  X2             -8.80246974       13.21432007


            Class Means on Canonical Variables Médias canônicas

                     trat              Can1              Can2

                        1      -3.198161274       0.627615714
                        2       6.022244556       0.026539512
                        3      -2.824083283      -0.654155226
```

Escores das variáveis canônicas

O Quadro 2 é o arquivo ‘can’ definido em ‘out=can’. Este arquivo fica armazenado na biblioteca (library) denominada WORK e deve ser exportado para o formato Excel ‘xls’ antes de se fechar o programa. Os arquivos gravados na biblioteca WORK são temporários e são apagados pelo SAS quando o programa é fechado.

Quadro 2. Arquivo ‘can’ temporário gravado na biblioteca WORK do SAS

trat	rep	X1	X2	Can1	Can2
1	1	4.63	0.95	-3.06682	-0.01276
1	2	4.38	0.89	-4.33028	-0.42438
1	3	4.94	1.01	-1.37336	0.307362
1	4	4.96	1.23	-3.16658	3.184013
1	5	4.48	0.94	-4.05376	0.083842
2	1	6.03	1.08	5.821904	-0.42984
2	2	5.96	1.19	4.35198	1.130479
2	3	6.16	1.08	6.753543	-0.62809
2	4	6.33	1.19	7.00357	0.566243
2	5	6.08	1.08	6.180227	-0.50609
3	1	4.71	0.96	-2.58153	-0.00261
3	2	4.81	0.93	-1.60081	-0.55154
3	3	4.49	0.87	-3.36592	-0.85641
3	4	4.43	0.82	-3.35579	-1.42563
3	5	4.56	0.91	-3.21637	-0.43458

Gráficos de dispersão

Os gráficos para d=1 ou d=2 envolverndo as médias canônicas podem representar uma ajuda importante na discriminação de tratamentos. A Figura 1 ilustra o gráfico de dispersão entre tratamentos representado pelos escores das duas primeiras variáveis canônicas. Observa-se que o efeito conjunto das variáveis X_1 e X_2 pode capturar a variação entre os tratamentos 2 e os demais (1 e 3). Contudo a análise não foi capaz de capturar a variância entre 1 e 3. Dessa forma podemos concluir que apenas essas características (X_1, X_2) não são suficientes para discriminar os indivíduos dessa população em três grupos diferentes. A interpretação da análise depende do fenômeno analisado, e a experiência do pesquisador é fator importante. Podemos também observar no Quadro 3 que não houve diferença siginificativa entre as distância de Mahalanobis entre os tratamentos 1 e 3, indicando que a separação desses indivíduos não é possível.

Quadro 3. Matriz de significância das distâncias de Mahalanobis entre tratamentos

Prob > Mahalanobis Distance for Squared Distance to trat

From trat	1	2	3
1	1.0000	<.0001	0.1760
2	<.0001	1.0000	<.0001
3	0.1760	<.0001	1.0000

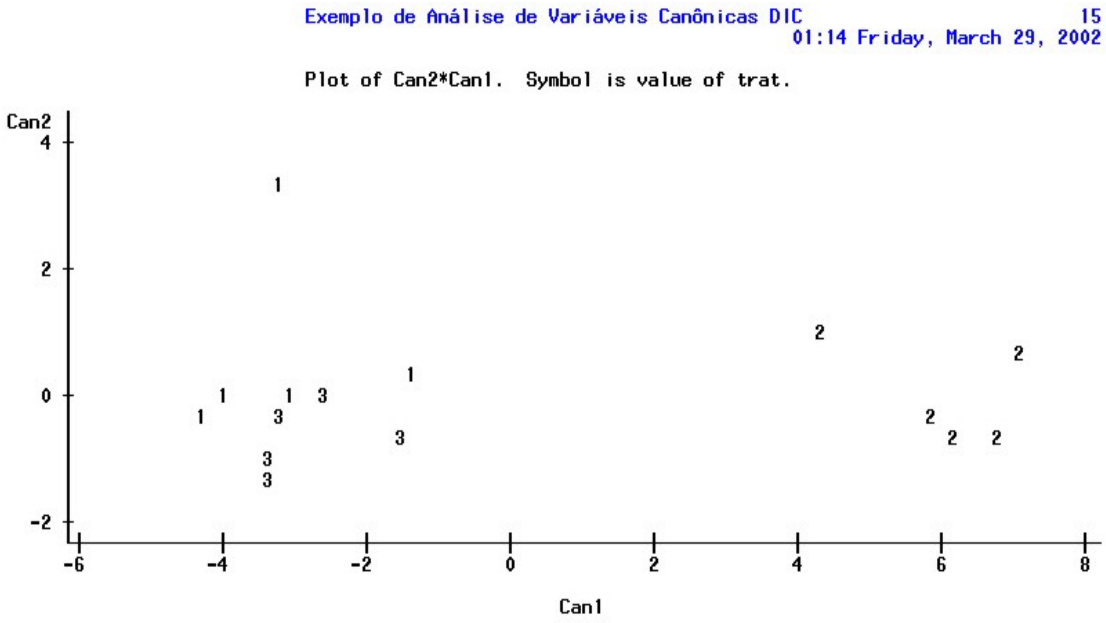

Figura 1. Dispersão dos escores das duas primeiras variáveis canônicas.

BIBLIOGRAFIA

FISHER, R.A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v.**7**, p.179-188, 1936.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999, 815 p.

KHATTREE, R. & NAIK, D.N**. Multivariate data reduction and discrimination with SAS software**. Cary, NC, USA: SAS Institute Inc., 2000. 558 p.

KHOURY JR, J.K. **Desenvolvimento e avaliação de um sistema de visão artificial para classificação de madeira serrada de eucalipto**. 2004. 101 f. Tese (Doutorado em Engenharia Agrícola) – Universidade Federal de Viçosa, Viçosa, 2004.

REGAZZI, A.J. Análise multivariada, notas de aula INF 766, Departamento de Informática da Universidade Federal de Viçosa, v.2, 2000.

VARELLA, C.A.A. **Estimativa da produtividade e do estresse nutricional da cultura do milho usando imagens digitais**. 2004. 92 f. Tese (Doutorado em Engenharia Agrícola) – Universidade Federal de Viçosa, Viçosa, 2004.

SAS. Online doc version 8. Disponível em: http://v8doc.sas.com/sashtml/. Acesso em 14 mar. 2007.

BARTLETT, M.S. Multivariate Analysis. J.R. Statist. Soc., Serie B, v.9, p.176-197, London, 1947.