AN UNIFIED APPROACH TO DIMENSIONALITY IDENTIFICATION AND FEATURE SELECTION IN CANONICAL VARIATE ANALYSIS

ANTÓNIO PEDRO DUARTE SILVA

FACULDADE DE ECONOMIA E GESTÃO
UNIVERSIDADE CATÓLICA PORTUGUESA
CENTRO REGIONAL DO PORTO

Canonical Variate Analysis

- n ENTITIES (OBSERVATIONS)
- PARTIONED BY k GROUPS
- DESCRIBED BY p ATTRIBUTES (FEATURES)

$\Rightarrow$

- **$r = \min(k-1, p)$ CANONICAL VARIATES:**

$$Z_i = \sum_j c_{ij} X_j$$

$$[c_{i1}, c_{i2}, \ldots, c_{ip}] = \text{Eigvct}_i(B\, W^{-1}) = \text{Eigvct}_i(B\, T^{-1})$$

$$W = \sum_{g=1}^{k} \sum_{i=1}^{n_g} (x_{gi} - \bar{x}_g)(x_{gi} - \bar{x}_g)^T \qquad B = \sum_{g=1}^{k} n_g (\bar{x}_g - \bar{\bar{x}})(\bar{x}_g - \bar{\bar{x}})^T$$

$$T = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{\bar{x}}_j)(x_{ij} - \bar{\bar{x}}_j)^T = B + W$$

Example I: Talent Data (Cooley and Lohnes 1968)

- **442 STUDENTS ATTENDING HIGH SCHOOL IN 1960**
- **DESCRIBED BY 15 ATTRIBUTES(1)**
- **ENROLLED IN FOUR DIFFERENT TYPES OF INSTITUTIONS IN 1962(2)**

(1) ATTRIBUTE VARIABLES:

Cognitive:

Literature information	--	LINFO
Social Science information	--	SINFO
English proficiency	--	EPROF
Mathematics reasoning	--	MRSNG
Visualization in three dimensions	--	VTDIM
Mathematics information	--	MINFO
Clerical-perceptual speed	--	CPSPD

Interest:

Physical science	--	PSINT
Literary-linguistic	--	LLINT
Business management	--	BMINT
Computation	--	CMINT
Skilled trade	--	TRINT

Temperament:

Sociability	--	SOCBL
Impulsiveness	--	IMPLS
Mature personality	--	MATRP

(2) STUDENT GROUPS:

G1 --- 89 students in a teacher college

G2 --- 75 students in a vocational school

G3 --- 78 students in a business or technical school

G4 --- 200 students in a university

RESULTS

Table 1: Eigenvalues

Function	Eigenvalue (λ_i)	% of Variance	Cumulative %	Canonical Correlation (r_i)
1	.587	84.9	84.9	.608
2	.082	11.9	96.8	.275
3	.022	3.2	100.0	.148

Table 2: Standardized Canonical Function Coefficients

	Z_1 Coef.	Z_2 Coef.	Z_3 Coef.		Z_1 Coef.	Z_2 Coef.	Z_3 Coef.
LINFO	-.123	.417	-.316	**PSINT**	**.343**	-.091	.086
SINFO	.255	-.246	-.456	**LLINT**	**.351**	-.033	-.014
EPROF	-.283	**-.683**	-.161	**BMINT**	**.401**	-.061	.297
MRSNG	.014	**.759**	-.438	CMINT	-.233	-.143	.074
VTDIM	-.131	.316	.360	**TRINT**	**-.630**	.335	.076
MINFO	**.700**	-.062	.412	SCOLB	.042	-.026	.072
CPSPD	.093	.293	.167	IMPLS	-.020	-.034	.461
				MATRP	.197	.154	.029

2º Exemplo: Sector Financeiro Português

- **33 INSTITUÍÇÕES FINANCEIRAS A OPERAR EM PORTUGAL EM 1993**
- **DESCRITAS POR 17 ATRIBUTOS(1)**
- **DIVIDIDAS POR TRÊS GRUPOS(2)**

(1) VARIÁVEIS:

Liquidez Reduzida	LR
Capacidade Creditícia Geral	CCG
Transformação dos Recursos de Clientes em Crédito (logt.)	ln TRCC
Grau de Endividamento	GE
Solvabilidade Bruta	SB
Taxa Média das Aplicações	TMA
Taxa Média dos Recursos	TMR
Margem Financeira	MF
Margem de Negócio	MN
Relevância dos Custos Pessoal	RCPE
Relevância Custos no Produto (logt.)	ln RCPD
Nº. Empregados por Balcão (logt.)	ln EB
Activo Líquido por Empregado	ALE
Rendibilidade Bruta do Activo	RBA
Rendibilidade Bruta Capitais Próprios	RBCP
Rendibilidade do Activo	RA
Rendibilidade dos Capitais Próprios	RCP

(2) GRUPOS DE INSTITUIÇÕES:

G1 --- 14 Instituições nacionais creadas antes de 1984

G2 --- 7 Instituições nacionais creadas depois de 1984

G3 --- 12 Instituições estrangeiras

RESULTADOS

Tabela 1: Valores Próprios

Função	Valor Próprio (λ_i)	% de Variância	% Variância Acumulada	Correlação Canónica (r_i)
1	9.214	76.7	76.7	.950
2	2.799	23.3	100.0	.858

Tabela 2: Coeficientes Canónicos Padronizados

	Z_1 Coef.	Z_2 Coef.		Z_1 Coef.	Z_2 Coef.
LR	1.690	.081	**MN**	**-2.748**	-1.773
CCG	.511	.523	RCPE	-.523	-1.508
ln TRCC	**2.791**	.572	**ln RCPD**	**2.263**	-.139
GE	.777	-.796	ln EB	-.701	.648
SB	**-3.165**	**-3.568**	ALE	.669	-1.605
TMA	-8.278	-6.947	RBA	.773	-.971
TMR	7.455	6.328	**RBCP**	.774	**1.916**
MF	8.582	7.887	RA	.735	.122
			RCP	-.713	-.397

Dimensionality Identification in Canonical Variate Analysis

- Asymptotic Statistics:

$$\lambda_i = Eigval_i(\mathbf{B}\mathbf{W}^{-1}) \; ; \; r_i^2 = Eigval_i(\mathbf{B}\mathbf{T}^{-1})$$

Bartlett (1947)

$$U_t = \left(n - 1 - \frac{p+k}{2}\right) \sum_{i=t+1}^{r} \ln(1+\lambda_i) = \left(n - 1 - \frac{p+k}{2}\right) \sum_{i=t+1}^{r} -\ln(1-r_i^2) \sim \chi^2_{(p-t)(k-1-t)}$$

Rao (1973)

$$T^2_{0,t} = (n-k) \sum_{i=t+1}^{r} \lambda_i = (n-k) \sum_{i=t+1}^{r} \frac{r_i^2}{1-r_i^2} \sim \chi^2_{(p-t)(k-1-t)}$$

- Closed Test Procedure

(Calinski and Lejeune 1998)

$$T^2_{0,t} \sim T^2_0(p-t, k-1-t, n-k)$$

- Model Selection through the Akaike Information Criterion

(Fujikoshi 1979)

$$A_t = n \sum_{i=t+1}^{r} \ln(1+\lambda_i) - 2(p-t)(k-1-t) = n \sum_{i=t+1}^{r} -\ln(1-r_i^2) - 2(p-t)(k-1-t)$$

Example I: Talent Data

t	U_t	(p-value)	$T^2_{0,t}$	p-value (Rao Ap.)	p-value (Lejeune UB)	A_t
0	242.6	(0.000)	302.4	(0.000)	(0.000)	158.5
1	43.6	(0.031)	45.8	(0.018)	(0.030)	-11.3
2	9.6	(0.726)	9.9	(0.702)	(0.724)	-16.2
3						0.0

2º Exemplo: Sector Financeiro Português

t	U_t	(p-value)	$T^2_{0,t}$	p-value (Rao Ap.)	p-value (Lejeune UB)	A_t
0	80.5	(0.000)	360.4	(0.000)	(0.000)	49.3
1	29.4	(0.022)	84.0	(0.000)	(0.035)	12.4
2						0.0

Variable Selection in Canonical Variate Analysis

- Additional Information Testes and Stepwise Strategies

- Model Selection through the Akaike Information Criterion (Fushikoshi 1985)

- Comparison of Multivariate Indices (Duarte Silva 2001)

Additional Information Testes and Stepwise Strategies

$$X_g = \underset{(1\times q)}{X_{g1}} \mid \underset{(1\times(p-q))}{X_{g2}} \sim N_p(\mu_g, \Sigma) \qquad \mu_g = \mu_{g1} \mid \mu_{g2} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Hypothesis of No Aditional Information

$$H_0(X_2 \mid X_1): (\mu_{g2} - \mu_{g'2}) = (\mu_{g1} - \mu_{g'1})\Sigma_{11}^{-1}\Sigma_{12} \qquad \forall\, g, g' = 1,2,\ldots,k$$

Statistical Tests

$$\Lambda = \Lambda_1 \Lambda_{2|1} \qquad \Lambda_1 = \frac{|W_{11}|}{|T_{11}|} \qquad \Lambda_{2|1} = \frac{|W_{22} - W_{21}W_{11}^{-1}W_{12}|}{|T_{22} - T_{21}T_{11}^{-1}T_{12}|}$$

$$H_0(X_2 \mid X_1) \Rightarrow \Lambda_{2|1} \sim \Lambda(p-q, r, n-p-q)$$

Stepwise Procedures

Forward: $H_0(X_i), H_0(X_j \mid X_i), H_0(X_k \mid X_i, X_j), \ldots$

Backward: $H_0(X_1,,\ldots,X_p), H_0(X_i \mid X_1,\ldots,X_{i-1}, X_{i+1},\ldots,X_p), \ldots$

Mixed

Akaike Information Criterion

$$A_{H_0(X_2|X_1)} = -n \ln \Lambda_{2|1} - 2r(p-q) = -n \Delta\left(\sum_{i=1}^{r} \ln(1-r_i^2)\right) - 2r(p-q)$$

Multivariate Indices

$$r_1^2$$

$$\tau^2 = 1 - \left(\prod_{i=1}^{r}(1-r_i^2)\right)^{1/r}$$

$$\xi^2 = \frac{\sum_{i=1}^{r} r_i^2}{r}$$

$$\varsigma^2 = 1 - \frac{r}{\sum_{i=1}^{r} \frac{1}{1-r_i^2}}$$

Example I: Talent Data

Stepwise Selection ($\alpha = 5\%$)

S_i = {SINFO, EPROF, MRNSG, MINFO, PSINT, LLINT, BMINT, TRINT}

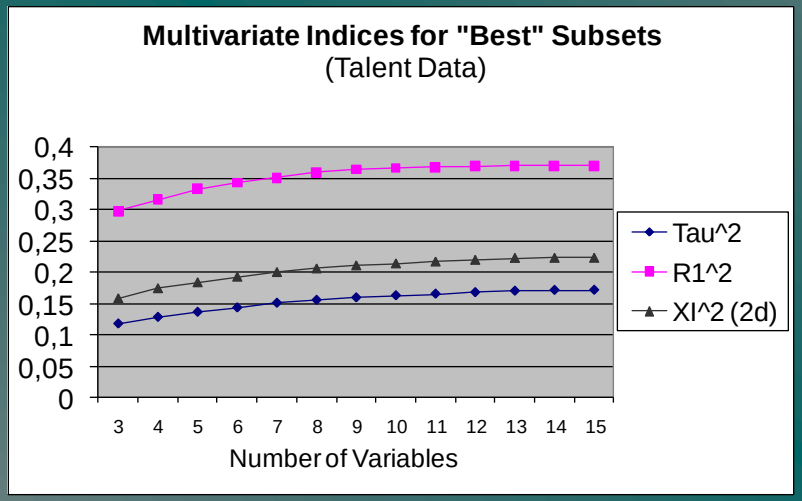
Akaike Information Criterion

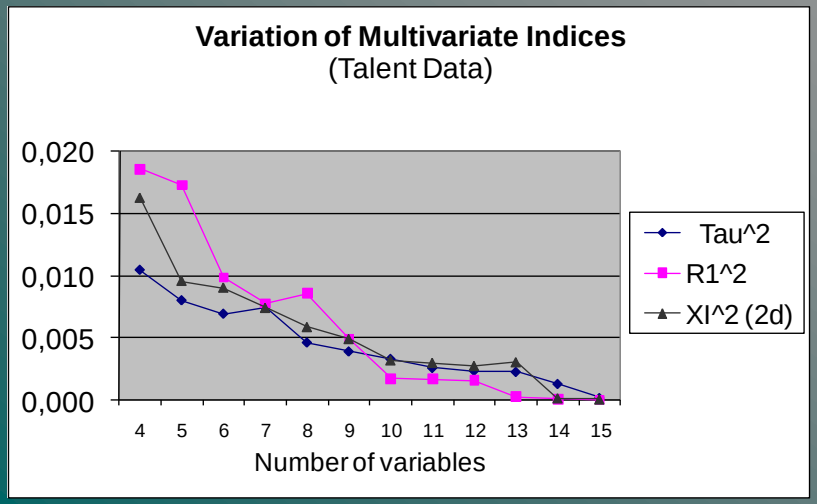
$S_j = S_i \cup \{MATRP\}$ $\quad (A_{H_0(X_2|X_1)} = -16.9)$

S_i $\quad (A_{H_0(X_2|X_1)} = -16.8)$

$S_k = S_j \setminus \{PSINT\}$ $\quad (A_{H_0(X_2|X_1)} = -16.4)$

Multivariate Indices





Best 8- and 9- Variable subsets according to r_1^2

S_1 = {**CMINT**, EPROF, **MATRP**, MINFO, PSINT, LLINT, BMINT, TRINT}

S_m = {SINFO, **CMINT**, EPROF, MATRP, MINFO, PSINT, LLINT, BMINT, TRINT}

Best 13- Variable subset according to $\xi^2(2d)$

$S_n = S_1 \setminus$ {SOCBL, IMPLS}

2º Exemplo: Sector Financeiro Português

Selecção passo a passo (α = 10%)

Ascendente:

S_i = {CCG, ln TRCC,MN}

Descendente:

S_j = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP}

Critério de Akaike

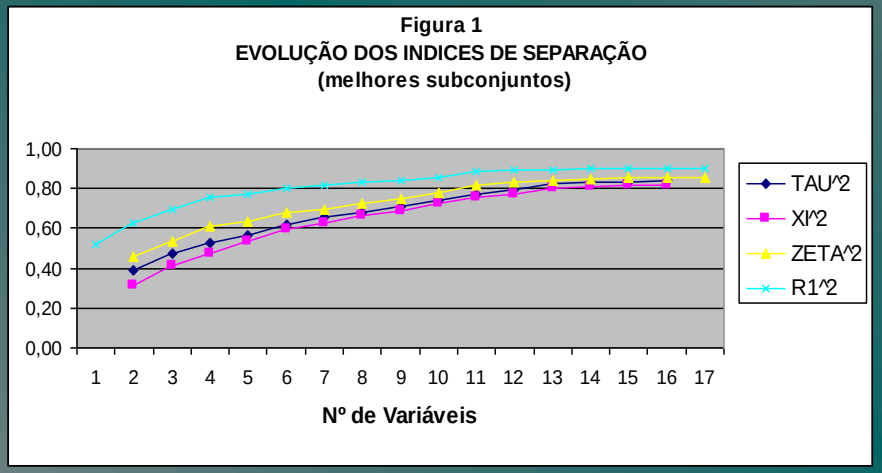
S_k = {LR, **CCG**, SB, GE, TMA, TMR, MF, **RA**, RCPE, ALE} $(A_{H_0(X_2|X_1)} = -7.73)$

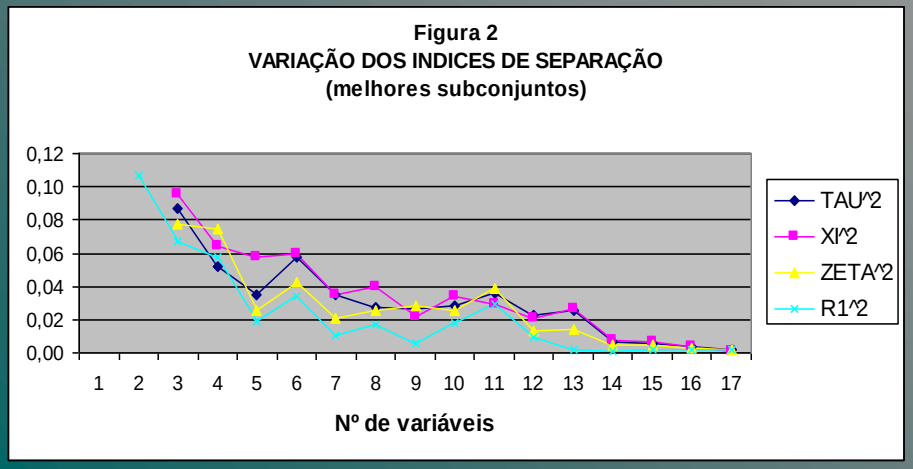
S_l = {**CCG**, SB,TMA,TMR,MF,RBCP,ln RCPD,ALE,**TCC**} $(A_{H_0(X_2|X_1)} = -6.41)$

S_m = {RBCP, **CCG**, SB, **TCC**, TMA, TMR, MF, **RA**, RCPE, ALE}

$(A_{H_0(X_2|X_1)} = -6.36)$

Indices Multivariados





Melhor subconjunto de 11 variáveis segundo r_1^2

S_n = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, ln RCPD, ln EB, ALE}

Melhor subconjunto de 13 variáveis segundo ξ^2

S_j = {LR, ln TRCC, SB, GE, TMA, TMR, MF, MN, RCPE, ln RCPD, ln EB, ALE, RBCP}

An Unified Model for Dimensionality Identification and Feature Selection

$$X_g = \underset{(1\times q)}{X_{g1}} \,|\, \underset{(1\times(p-q))}{X_{g2}} \sim N_p(\mu_g, \Sigma) \qquad \mu_g = \mu_{g1} \,|\, \mu_{g2} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$H_{0t}(X_2 \mid X_1):$$

$$\Omega_{11} = \sum_{g=1}^{k} \frac{n_g}{n} (\mu_{g1} - \bar{\mu}_1)(\mu_{g1} - \bar{\mu}_1)^T \qquad rank(\Omega_{11}) = t \qquad (t \le r)$$

$$(\mu_{g2} - \mu_{g'2}) = (\mu_{g1} - \mu_{g'1}) \Sigma_{11}^{-1} \Sigma_{12} \qquad \forall\, g, g' = 1,2,\ldots,k$$

Akaike Information Criterion

$$A_{H_0 t(X_2|X_1)} = -n \left[\sum_{i=t+1}^{t} \ln(1 - r_i^2) + \Delta\left(\sum_{i=1}^{t} \ln(1 - r_i^2) \right) \right] - 2(q-t)(k-1-t) + r(p-q)$$

Inference from t-Dimensional Indices: A Bootstrap Approach

1 - Generate a Model, M(j), Consistent with the Data and $H_{0t}(X_2 \mid X_1)$

Adjust means in order to respect $H_{0t}(X_2 \mid X_1)$

Keep all higher-order data moments unchanged

2 - Resample with Replication from M(j)

3 - Generate Distribution of t-Dimensional Multivariate Indices for Best Subsets

4 - Compare Original Indices with the Percentiles of the Distribution Generated in 3

References

Bartlett, M.S. (1947), "Multivariate Analysis", *Journal of the Royal Statistical Society Suppl.*, **9**, 176-190

Calinski, C. and Lejeune, M. (1998), "Dimensionality in MANOVA Tested by a Closed Testing Procedure", *Journal of Multivariate Analysis*, **65**, 181-194.

Cooley, W.W. and Lohnes, P.R. (1968), *Multivariate Procedures for the Behavioral Sciences*, John Wiley.

Duarte Silva, A.P. (1998), "Análise Discriminante com Selecção de Variáveis. 1ª Parte: Descrição", *Revista de Estatística*, 5-42.

Duarte Silva, A.P. (2001), "Efficient Variable Screening for Multivariate Analysis" , *Journal of Multivariate Analysis*, **76**, 35-62.

Fujikoshi, Y. (1979), "Estimation of Dimensionality in Canonical Correlation Analysis", *Biometrika*, **66** (2), 345-351.

Fujikoshi, Y. (1985), "Selection of Variables in Discriminant Analysis and Canonical Correlation Analysis", *Multivariate Analysis VI* (P.R. Krishnaiah Ed.), Elsevier Science Publishers, 219-236.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, 2nd ed., John Wiley.