DISCARDING VARIABLES IN PRINCIPAL COMPONENT ANALYSIS: ALGORITHMS FOR ALL-SUBSETS COMPARISONS BASED ON THE RV COEFFICIENT

ANTÓNIO PEDRO DUARTE SILVA

FACULDADE DE CIÊNCIAS ECONÓMICAS E EMPRESARIAIS
UNIVERSIDADE CATÓLICA PORTUGUESA
CENTRO REGIONAL DO PORTO

Discarding and/or Selecting Variables in PCA

WHY DISCARD VARIABLES ?

- **PARSIMONITY**
- **REDUCING COSTS OF DATA COLLECTION**
- **HELPING THE INTREPETATION OF A CLASSICAL PCA**

  - THE TRADITIONAL APPROACH OF IGNORING VARIABLES WITH SMALL LOADINGS, IMPLICITLY DISCARDS THEM

  - ASSESSING VARIABLE IMPORTANTE ONLY BASED ON LOADINGS CAN BE MISLEADING (CADIMA AND JOLLIFFE 1995)

- **AS A DIRECT APPROACH TO THE PROBLEM OF DIMENSIONALITY REDUCTION (McCabe 1984)**

CRITERIA FOR COMPARING VARIABLE SUBSETS

McCABE APROACH (PRINCIPAL VARIABLES 1984):

$Y = A\ X\ ;\quad Z = B\ Y$

$(n*k)\quad (n*p)\quad (n*p)\quad (n*k)$

$(A = [I_1 \mid 0\ ]\ ;\ X = [X_1 \mid X_2\ ] \Rightarrow Y = X_1)$

(1) $\text{MIN}\ |Z - X| \Leftrightarrow \text{MIN}\ |S_{22|1}| = \prod \theta_i$

(2) $\text{MIN}\ \text{Tr}\,(Z - X) \Leftrightarrow \text{MIN}\ \text{Tr}\ S_{22|1} = \sum \theta_i$

(3) $\text{MIN}\ ||Z - X||^2 \Leftrightarrow \text{MIN}\ ||S_{22|1}||^2 = \sum \theta_i^2$

(4) $\text{MIN}\ (Z - X)^T S_X^{-1} (Z - X) \Leftrightarrow \text{MIN Tr}\,(S_{22}^{-1} S_{22|1}) = (p\text{-}K) - \sum \rho_i^2$

$(S_{22|1} = S_{22} - S_{21} S_{11}^{-1} S_{12})$

CRITERIA BASED ON CORRELATIONAL MEASURES :

$r_1(A,B) = Tr(A^T B) / [Tr (A^T A) Tr (B^T B)]^{1/2}$

$A = P_A \Lambda_A Q_A^T \qquad B = P_B \Lambda_B Q_B^T$

YANAI' S GENERALIZED CORRELATION COEFFICIENT

$GCD(A,B) = r_1(P_A P_A^T, P_B P_B^T)$

MAX $GCD(X_1, PC_q(X))$ (Cadima and Jolliffe 1996)

ESCOUFIER'S RV COEFFICIENT

$RV(A,B) = r_1(P_A \Lambda_A^2 P_A^T, P_B \Lambda_B^2 P_B^T)$

MAX $RV(X, M X_1)$ (Robert and Escoufier 1976)

ALL-SUBSETS COMPARISONS IN LINEAR REGRESSION

$$y = X' \hat{\beta} + e$$

FURNIVAL'S ALGORITHM (1971):

$$\begin{bmatrix} X'X & X'y \\ y'X & y'y \end{bmatrix} \Rightarrow \begin{bmatrix} -(X'X)^{-1} & \hat{\beta} \\ \hat{\beta}' & e'e \end{bmatrix}$$

ELEMENTARY MATRIX OPERATIONS (SYMMETRIC SWEEPS)

COMMENTS:

- MATRIX SIMMETRY IS ALWAYS PRESERVED
- ONLY THE RIGHT-LOWER CORNER NEEDS TO BE UPDATED AT EACH STEP
- BY SEQUECING THE SUBSETS EVALUATIONS PROPERLY:

  1/2 OF THE SWEEPS UPDATE (1*1) SUBMATRICES

  1/4 OF THE SWEEPS UPDATE (2*2) SUBMATRICES

  …

  ONLY ONE SWEEP UPDATES THE FULL (P*P) MATRIX

TOTAL NUMBER OF FLOATING POINT OPERATIONS:

$$\approx 6 * 2^{P}$$

$$(6 (2^{p}) - p(p+7)/2 - 6)$$

FURNIVAL AND WILSON'S "LEAPS AND BOUNDS" (1974):

$$S_A \subset S_B \quad \Rightarrow \quad (e^T e)_A \geq (e^T e)_B$$

DOUBLE TREE SEARCH

ϕ	$X_1, X_2, \ldots, X_P$
↓ ↓ ↓ ↓ ↓ ↓	↓ ↓ ↓ ↓ ↓ ↓
ADDING VARIABLES	**REMOVING VARIABLES**
CREATE BOUNDS	$e^T e$ > **BOUND** ⇒ **PRUNE BRANCH**

ALL-SUBSETS COMPARISONS BASED ON THE RV COEFFICIENT

$$\text{MAX RV}(X, M X_1) \Leftrightarrow \text{MAX Tr}\{ [(S_{11})^{-1} (S^2)_{11}]^2 \} =$$

$$= \sum_{i,j} \{ [(S_{11})^{-1} (S^2)_{11}](i,j) * [(S_{11})^{-1} (S^2)_{11}](j,i) \}$$

SYMMETRIC SWEEPING OF COVARIANCE MATRICES:

$$A = \begin{bmatrix} -(S_{11})^{-1} & -(S_{11})^{-1} S_{12} \\ -S_{21} (S_{11})^{-1} & S_{22|1} \end{bmatrix} \qquad B = \begin{bmatrix} -(S_{1'1'})^{-1} & -(S_{1'1'})^{-1} S_{1'2'} \\ -S_{2'1'} (S_{1'1'})^{-1} & S_{2'2'|1'} \end{bmatrix}$$

$$X_{1'} = X_1 \cup \{X_u\} \; ; \quad X_{2'} = X_2 \setminus \{X_u\}$$

$$B(u,u) = - 1 / A(u,u)$$

$$B(i,u) = A(i,u) * B(u,u) \qquad (i \neq u)$$

$$B(i,j) = A(i,,j) + A(i,u) * B(j,u) \qquad (i \neq u, i \neq u)$$

UPDATING THE ELEMENTS OF $[(S_{11})^{-1}\,(S^2)_{11}]$

$$\left((S_{1'1'})^{-1}\,(S^2)_{1'\bullet}\right)(u,j) = \frac{\left(S^2\right)(u,j) - \sum_{a:X_a \in X_1}\left[\left((S_{11})^{-1}\,S_{12}\right)(a,u) * \left(S^2\right)(a,j)\right]}{S_{22|1}(u,u)}$$

$$\left((S_{1'1'})^{-1}\,(S^2)_{1'\bullet}\right)(i,j) = \left((S_{11})^{-1}\,(S^2)_{1\bullet}\right)(i,j) - \frac{\left((S_{11})^{-1}\,S_{12}\right)(i,u)}{S_{22|1}(u,u)} *$$

$$* \left\{ S^2(u,j) - \sum_{a:X_a \in X_1}\left[\left((S_{11})^{-1}\,S_{12}\right)(a,u) * \left(S^2\right)(a,j)\right] \right\} \qquad (i \neq u)$$

NUMBER OF FLOATING POINT OPERATIONS

PER SWEEP:

$$(5/2)\ K_1^2 + (11/2)\ K_1 + 2$$

TOTAL:

$$(5/2)\ K_1^2 + (11/2)\ K_1 + 2$$

OPERATORS FOR REMOVING VARIABLES

$$A = \begin{bmatrix} (S_{1'1'})^{-1} & (S_{1'1'})^{-1} S_{1'2'} \\ S_{2'1'}(S_{1'1'})^{-1} & -S_{2'2'|1'} \end{bmatrix} \qquad B = \begin{bmatrix} (S_{11})^{-1} & (S_{11})^{-1} S_{12} \\ S_{21}(S_{11})^{-1} & -S_{22|1} \end{bmatrix}$$

$$X_{1'} = X_1 \cup \{X_u\} \;\; ; \;\; X_{2'} = X_2 \setminus \{X_u\}$$

$B(u,u) = - 1 / A(u,u)$

$B(i,u) = A(i,u) * B(u,u) \qquad (i \neq u)$

$B(i,j) = A(i,,j) + A(i,u) * B(j,u) \qquad (i \neq u, i \neq u)$

$$\left((S_{11})^{-1}(S^2)_{1\bullet}\right)(i,j) = \left((S_{1'1'})^{-1}(S^2)_{1'\bullet}\right)(i,j) - (S_{1'1'})^{-1}(i,u)*(S^2)(j,u) -$$

$$- \frac{(S_{1'1'})^{-1}(i,u)}{(S_{1'1'})^{-1}(u,u)} * \sum_{a:X_a \in X_1} \left[(S_{1'1'})^{-1}(a,u)*\left(S^2\right)(a,j) \right]$$

CONCLUSIONS

ALL-SUBSETS COMPARISONS ARE POSSIBLE:

(a) IF THE NUMBER OF ORIGINAL VARIABLES IS NOT TOO LARGE (SAY $P \leq 25$)

OR

(b) IF A "FEW" SUBSETS ARE CLEARLY PREFERABLE TO THE OTHERS

FURTHER RESEARCH

WHEN ALL-SUBSETS COMPARISONS ARE NOT POSSIBLE
CAN HEURISTIC METHODS BE ALMOST OPTIMAL ???

SOFTWARE IMPLEMENTATION

LOOK AT:

http://www.porto.ucp.pt/psilva