Subselect 0.9-99: Selecting variable subsets in multivariate linear models

A. PEDRO DUARTE SILVA[1][*]

JORGE CADIMA[2]

MANUEL MINHOTO[3]

JORGE ORESTES CERDEIRA[2]

<u>THE PROBLEM:</u> **Finding a k-variable subset that is a good surrogate for a full p-variable data set**

CONTEXT:

Exploratory data analysis – Subselect 0.1-- 0.9 **Multivariate Linear Models – Subselect 0.9-99**

(Cadima, Cerdeira, Duarte Silva and Minhoto -- useR! 2004)

A LINEAR HYPOTHESIS FRAMEWORK

$X = A\,\Psi + U$ $H0: C\,\Psi = 0$

- **SELECT COLUMNS OF X IN ORDER TO EXPLAIN H1**

<u>PARTICULAR CASES:</u>

CANONICAL CORRELATION ANALYSIS /

LINEAR REGRESSION ANALYSIS

$A = [1 \mid Y]$ $C = [0 \mid I]$

LINEAR DISCRIMINANT ANALYSIS

$A = [1_g]$ $\Psi = [\mu_g]$ $C = \begin{bmatrix} 1 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & -1 \end{bmatrix}$

MULTI-WAY MANOVA/MANCOVA EFFECTS

<u>COMPARISON CRITERIA:</u>

ccr_1^2

max ccr_1^2 ⇔

max **Roy first root**

$\varsigma^2 = 1 - \frac{r}{\sum_{i=1}^{r}\left(1 - ccr_i^2\right)^{-1}}$

max ς^2 ⇔

max **Lawley-Hotelling trace)**

MULTIVARIATE INDICES

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

max τ^2 ⇔

min **Wilks** Λ

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

max ξ^2 ⇔

max **Bartlei-Pillai trace)**

$T = X'(I - P_\omega)X$

$ccr_i^2 = Eigval_i(T^{-1}H)$ $\Omega = \mathcal{R}(A)$ $\omega = \mathcal{R}(A) \cap \mathcal{M}(C)$ $r = \dim(\Omega) - \dim(\omega)$

$H = X'(P_\Omega - P_\omega)X$

<u>THE SUBSELECT PACKAGE</u>

Search routines for (combinatorial) criteria optimization

Exact Algorithm:

leaps - **based on Furnival and Wilson´s leaps and bounds algorithm**

- **viable with up to 30 - 35 original variables**

Heuristics:

anneal - **simulated annealing**

genetic - **genetic algorithm**

improve – **restricted local improvement**