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]

(1) FEG/CEGE – UNIV. CATÓLICA PORTUGUESA – C.R. PORTO

(2) I.S. AGRONOMIA – UNIV. TÉCNICA DE LISBOA

(3) DEP. MATEMÁTICA – UNIVERSIDADE DE ÉVORA

(*) Supported by: FEDER / POCI 2010

Ciência.Inovação 2010 Programa Operacional Ciência e Inovação 2010 MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

FCT Fundação para a Ciência e a Tecnologia MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA

Subselect 0.9-99

<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**

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

- **Multivariate Linear Models – Subselect 0.9-99**

Subselect 0.9-99

A LINEAR HYPOTHESIS FRAMEWORK

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

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

PARTICULAR CASES:

CANONICAL CORRELATION ANALYSIS $A = [1 \mid Y] \quad C = [0 \mid I]$

LINEAR DISCRIMINANT ANALYSIS

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

MULTI-WAY MANOVA/MANCOVA EFFECTS

Subselect 0.9-99

$\Omega = \mathcal{R}(\mathbf{A})$ $\omega = \mathcal{R}(\mathbf{A}) \cap \mathcal{N}(\mathbf{C})$ $\mathbf{r = dim(\Omega) - dim(\omega)}$

$\mathbf{ccr_i^2 = Eigval_i(T^{-1}H)}$ $\mathbf{T = X'(I - P_\omega)X}$ $\mathbf{H = X'(P_\Omega - P_\omega)X}$

Comparison Criteria:

$$\mathbf{ccr_1^2}$$

(*max* ccr_1^2 $\Leftrightarrow$ *max* Roy λ_1)

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

(*max* τ^2 $\Leftrightarrow$ *min* Wilks Λ)

Multivariate Indices

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

(*max* ς^2 $\Leftrightarrow$ *max* Lawley-Hotelling trace)

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

(*max* ξ^2 $\Leftrightarrow$ *max* Bartlei-Pillai trace)

Subselect 0.9-99

The Subselect Package

Search routines for (combinatorial) criteria optimization

Exact Algorithm:

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

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

Heuristics:

anneal - **simulated annealing**

genetic - **genetic algorithm**

improve - **restricted local improvement**

Subselect 0.9-99

Subselect in Multivariate Linear Models

Principal arguments of search routines :

mat	- **Total SSCP data matrix (T)**
H	- **Effect SSCP data matrix**
r	- **Expected rank of the H matrix**
criterion	- **“ccr12”, “tau2”, “xi2” or “zeta2”**
kmin, kmax	- **minimum and maximum subset dimensionalities sought**

Subselect 0.9-99

Subselect in Multivariate Linear Models

<u>Other arguments :</u>

- Tuning parameters for heuristics

- Maximum time allowed for exact search

- Variables forcibly included or excluded in the selected subsets

- Number of solutions by subset dimensionality

- Numerical tolerance for detecting singular or non-symmetrical matrices

Subselect 0.9-99

Subselect in Multivariate Linear Models

Auxiliary functions:

lmHmat - creates H and mat matrices for linear regression/canonical correlation analysis

ldaHmat - creates H and mat matrices for linear discriminant analysis

glhHmat - creates H and mat matrices for an analysis based on a linear hypothesis specified by the user

Subselect 0.9-99

Subselect in Multivariate Linear Models

Auxiliary functions :

ccr12.coef, tau2.coef zeta2.coef, xi2.coef	**- computes a comparison criterion for a subset supplied by the user**
trim.matrix	**- deletes rows and columns of singular or ill-conditioned matrices**
	- until all linear dependencies (perfect or almost perfect) are removed

Subselect 0.9-99

Example: Hubbard Brook Forest soil data

Source: Morrison (1990)

Description:

58 pits were analyzed before (1983) and after (1986) harvesting (83-84) trees larger than a minimum diameter

Continuous variables: **gr/m^2 of exchangeable cations**

Al - Aluminum

Ca - Calcium

Mg - Magnesium

K - Potassium

Na - Sodium

Subselect 0.9-99

Example: Hubbard Brook Forest soil data

Source: Morrison (1990)

<u>Factors:</u>	<u>Factor levels:</u>
F - Forest Type	1 - Spruce- fir
	2 - High elevation hardwood
	3 - Low elevation hardwood
D - Logging Disturbance	0 - Uncut forest
	1 - Cut, undisturbed by machinery
	2 - Cut, disturbed by machinery
Year	1983 or 1986

Subselect 0.9-99

Example: Hubbard Brook Forest soil data

Source: Morrison (1990)

<u>Reading and preparing the data:</u>

```
> library(subselect)

> HubForest <- read.table("Hubbard Brook.txt“ ,header=T,
  col.names=c("Pit","F","D","Al","Ca","Mg","K","Na","Year"),
  colClasses=c("factor","factor","factor","numeric",
  "numeric","numeric","numeric","numeric","factor") )
```

<u>Analysis #1:</u> Explaining the levels of calcium

```
> Hmat <- lmHmat(Ca ~ F*D + Al + Mg+ K + Na ,HubForest)
> colnames(Hmat$mat)
> leaps(Hmat$mat,H=Hmat$H,r=1,nsol=3)
```

Subselect 0.9-99

Example: Hubbard Brook Forest soil data

Source: Morrison (1990)

<u>Analysis #2:</u> Looking for combinations of Forest type and Disturbance that best explain the nutrient levels

```
> Hmat <- lmHmat(cbind(Al,Ca,Mg,K,Na) ~ F*D,HubForest)
> colnames(Hmat$mat)
> leaps(Hmat$mat,H=Hmat$H,r=5,criterion="tau2",nsol=3)
```

<u>Analysis #3:</u> Finding which subsets of nutrients were most affected by the harvesting in 1983-84

```
> Hmat <- ldaHmat(Year ~  Al + Ca + Mg + K + Na , HubForest)
> leaps(Hmat$mat,H=Hmat$H,r=1,nsol=3)
```

Subselect 0.9-99

Example: Hubbard Brook Forest soil data

Source: Morrison (1990)

<u>Analysis #4:</u> Finding which subsets of nutrients are most affected by interactions between harvesting and logging disturbances, after controlling for the effect of forest type

```
>  C <- matrix(0.,2,8)
>  C[1,7] = C[2,8] = 1.
>   Hmat <- glhHmat(cbind(Al,Ca,Mg,K,Na) ~ D*Year + F, C,
    HubForest)
>   leaps(Hmat$mat,H=Hmat$H,r=2, criterion="tau2",
    nsol=3,tolsym=1E-10)
```

References

Cadima J, Cerdeira JO and Minhoto M (2004). Computational Aspects of Algorithms for Variable Selection in the Context of Principal Components. *Computational Statistics and Data Analysis* **47**: 225-236.

Cadima J, Cerdeira JO, Duarte Silva AP and Minhoto M (2004). The Subselect Package; Selecting Variable Subsets in an Exploratory Data Analysis. *useR! 2004. 1rst Internatinal R User Conference. Vienna, Austria.*

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

Furnival, G.M. & Wilson, R.W. (1974). Regressions by Leaps and Bounds. *Technometrics* **16**: 499-511.

Morrison D.F. (1990). *Multivariate Statistical Methods*, 3rd ed., McGraw-Hill. New York, NY.