EXTENDED LEAPS

A MODERN IMPLEMENTATION OF A TIME TESTED CLASSIC

ANTÓNIO PEDRO DUARTE SILVA

FACULDADE DE ECOMIA E GESTÃO
CENTRO DE ESTUDOS EM GESTÃO E ECONOMIA
UNIVERSIDADE CATÓLICA PORTUGUESA – C.R. PORTO

OVERVIEW

- ✓ FURNIVAL AND WILSON'S LEAPS AND BOUNDS
- ✓ LEAPS AND BOUND STRATEGIES:
  WHY DOES IT WORK?
- ✓ EXTENDED LEAPS: BEYOND LINEAR REGRESSION
- ✓ A C++ LIBRARY OF LEAPS SEARCH ROUTINES
- ✓ USING THE LIBRARY:
  A CONSOLE APPLICATION AND A R PACKAGE
- ✓ FUTURE DEVELOPMENTS

ALL-SUBSETS COMPARISONS IN LINEAR REGRESSION

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

FURNIVAL'S EXHAUSTIVE ALGORITHM (1971):

$$\begin{bmatrix} R_{XX} & R^{T}_{Xy} \\ R_{Xy} & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} -R^{-1}_{XX} & -\hat{\beta}^{*T} \\ -\hat{\beta}^{*} & -R^{2}_{\hat{y}y} \end{bmatrix}$$

$$\hat{\beta}^{*}(j) = \frac{s(X_j)}{s(y)}\hat{\beta}(j)$$

SPECIALIZED GAUSSIAN ELIMINATION OPERATIONS

(SYMMETRIC SWEEPS)

ALL-SUBSETS COMPARISONS IN LINEAR REGRESSION

SYMMETRIC SWEEPING OF CORRELATION MATRICES:

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

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

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

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

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

ALL-SUBSETS COMPARISONS IN LINEAR REGRESSION

COMMENTS:

- ✓ **MATRIX SIMMETRY IS ALWAYS PRESERVED**

  BY SEQUENCING THE SUBSETS EVALUATIONS PROPERLY

  (Ex: $I(X1) = \sum_{a=1}^{p} \delta_a(X_1) * 2^{a-1}$ $\delta_a(X_1) = 1$ **if** $X_a \in X_1$; $\delta_a(X_1) = 0$ **if** $X_a \notin X_1$**)**

- ✓ **ONLY** 2^{p-t-1} **SWEEPS NEED TO UPDATE THE DATA ASSOCIATED WITH t GIVEN VARIABLES**

  $\Rightarrow$

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

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

  …

  ONLY ONE SWEEP NEEDS TO UPDATE THE FULL ((P+1)*(P+1)) MATRIX

ALL-SUBSETS COMPARISONS IN LINEAR REGRESSION

NUMBER OF FLOATING POINT OPERATIONS:

PER SWEEP:

$(1/2)\ (t^2 + 5t + 4)$	multiplications/divisions
$(1/2)\ (t^2 + t + 2)$	additions/subtractions
$t^2 + 3t + 3$	flops

TOTAL:

$6(2^p) - (1/2)p^2 - (7/2)p - 6$	multiplications/divisions
$3(2^p) - (1/2)p^2 - (3/2)p - 3$	additions/subtractions
$9(2^p) - p^2 - 5\,p - 9$	flops

LEAPS AND BOUNDS (1974)

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

DOUBLE TREE SEARCH

ϕ $\qquad\qquad$ $X_1, X_2, \ldots, X_P$

↓ ↓ ↓ ↓ ↓ ↓ $\qquad$ ↓ ↓ ↓ ↓ ↓ ↓

ADDING VARIABLES $\qquad$ **REMOVING VARIABLES**

CREATE BOUNDS $\qquad$ $e^T e$ > **BOUND**

$\Rightarrow$ **PRUNE BRANCH**

TOTAL NUMBER OF FLOATING POINT OPERATIONS (NO PRUNING):

$6(2^p) + O(p^2)$	multiplications/divisions
$3(2^p) + O(p^2)$	additions/subtractions
$9(2^p) + O(p^2)$	flops

LEAPS EXTENSIONS

- ✓ **Linear Discriminant Analysis**
  (McCabe, *Technometrics*, 1975)

- ✓ **Generalized linear models**
  (Lawless and Singhal, *Biometrics*, 1978)

- ✓ **Linear models with multivariate responses**
  (Duarte Silva, *JMVA*, 2001)

- ✓ **Exploratory Multivariate Analysis**
  (Duarte Silva, *Comp. St.*, 2002)

EXTENDED LEAPS

A general implementation of the leaps and bounds algorithm

The basic algorithm is separated from comparison criteria specific details

Diferent problems and criteria can be easily incorporated:

- **Defining functions for updating a criterion when adding and removing single variables (“pivot variables”)**
- **Specifying the information required by each criterion update**
- **Defining functions for updating the information that will be required by *t* future “pivot variables”**

EXTENDED LEAPS

NUMBER OF FLOATING POINT OPERATIONS:

IF

IN EACH UPDATE : $a\,t^2 + b\,t + c$

THEN

TOTAL: $(3a + b + c)\ (2^p)\ +\ O(p^2)$

EXTENDED LEAPS

AVAILABLE CRITERIA:

REGRESSION PROBLEMS

QUADRATIC FORMS: $v_1^T \; Q_{11}^{-1} \; v_1$

LINEAR MODELS WITH MULTIVARIATE RESPONSES

WILKS RATIO: $\det(E) / \det(E+H)$

TRACE STATISTICS: $tr\left(H\,E^{-1}\right)$; $tr\left(H\,(H+E)^{-1}\right)$

SQUARED CANONICAL CORRELATIONS

EXPLORATORY MULTIVARIATE ANALYSIS

McCABE *PRINCIPAL VARIABLES* CRITERIA

ESCOUFIER'S RV COEFICIENT

CADIMA AND JOLIFFE GCD COEFICIENT

SSCMA – A CONSOLE APPLICATION FOR MULTIVARIATE PROBLEMS

- ✓ **A console application to search variable subsets in multivariate linear models and exploratory problems**
- ✓ **Uses extended leaps rotines either as exact algorithms or as heuristics (early stoping)**

User has control over:

- ✓ **Choice of comparison criterion**
- ✓ **Number of best subsets to look for**
- ✓ **Groups of variables to forcibly include or exclude from selected subsets**

- ✓ **Input and outpu through ASCII text files**
- ✓ **Portable (extensively tested under Windows/DOS and Linux)**

SUBSELECT – AN R PACKAGE FOR EXPLORATORY DATA ANALYSIS

- ✓ **A package to assess and search variable subsets in exploratory data analysis**
- ✓ **Implements meta-heuristic (sim. annealing; genetic algorithms., etc.) as well as exact (Extended Leaps) search functions**

Options:

- ✓ **Comparison criteria: RV, GCD, RM (McCabe 2nd criterion)**
- ✓ **Number of best subsets (for each cardinality) to keep**
- ✓ **Forcibly include or exclude variables**
- ✓ **Set time limit for exact searchs**
- ✓ **Set starting points and parameters for heuristic searchs**

Future Developments:

- ✓ **Linear models with univariate and multivariate responses**

WORK IN PROGRESS

Reliable routines for updating spectral and single value matrix decompositions

Interfaces to standard linear algebra systems (Lapack, Blas,...)

Cross-validation routines to measure and correct the effects of selection bias

Implementation of comparison criteria based on Procrustes Rotations

DOCUMENTATION AND DOWNLOADS

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

(LOOK AFTER DECEMBER 2005)