Package ‘subselect’

October 14, 2022

Version 0.15.3
Date 2022-08-30
Title Selecting Variable Subsets
Author Jorge Orestes Cerdeira [aut],
Pedro Duarte Silva [aut],
Jorge Cadima [aut, cre],
Manuel Minhoto [aut]
Maintainer Jorge Cadima <jcadima@isa.ulisboa.pt>
Description A collection of functions which (i) assess the quality of variable subsets as surro-
gates for a full data set, in either an exploratory data analysis or in the context of a multivari-
ate linear model, and (ii) search for subsets which are optimal under various criteria. Theoreti-
cal support for the heuristic search methods and exploratory data analysis crite-
ria is in Cadima, Cerdeira, Minhoto (2003, <doi:10.1016/j.csda.2003.11.001>). Theoretical sup-
port for the leap and bounds algorithm and the criteria for the general multivariate lin-
ear model is in Duarte Silva (2001, <doi:10.1006/jmva.2000.1920>). There is a package vi-
gnette ``subselect'', which includes additional references.
Depends R (>= 3.5.0)
Imports MASS, ISwR, corpcor
License GPL (>= 2)
Repository CRAN
LazyData TRUE
NeedsCompilation yes
Date/Publication 2022-08-30 16:50:02 UTC

R topics documented:

anneal .................................................. 2
ccrl2.coef ........................................... 13
eleaps ................................................. 15
farm .................................................. 25
gcd.coef ............................................. 27
 genetic ............................................. 29
Description

Given a set of variables, a Simulated Annealing algorithm seeks a k-variable subset which is optimal, as a surrogate for the whole set, with respect to a given criterion.

Usage

```r
anneal(mat, kmin, kmax = kmin, nsol = 1, niter = 1000, exclude = NULL, include = NULL, improvement = TRUE, setseed = FALSE, cooling = 0.05, temp = 1, coolfreq = 1, criterion = "default", pcindices = "first_k", initalsol=NULL, force=FALSE, H=NULL, r=0, tolval=1000*.Machine$double.eps,tolsym=1000*.Machine$double.eps)
```

Arguments

- **mat**: a covariance/correlation, information or sums of squares and products matrix of the variables from which the k-subset is to be selected. See the Details section below.
- **kmin**: the cardinality of the smallest subset that is wanted.
- **kmax**: the cardinality of the largest subset that is wanted.
- **nsol**: the number of initial/final subsets (runs of the algorithm).
- **niter**: the number of iterations of the algorithm for each initial subset.
- **exclude**: a vector of variables (referenced by their row/column numbers in matrix `mat`) that are to be forcibly excluded from the subsets.
- **include**: a vector of variables (referenced by their row/column numbers in matrix `mat`) that are to be forcibly included in the subsets.
improvement is a logical variable indicating whether or not the best final subset (for each cardinality) is to be passed as input to a local improvement algorithm (see function improve).

setseed is a logical variable indicating whether to fix an initial seed for the random number generator, which will be re-used in future calls to this function whenever setseed is again set to TRUE.

cooling is a variable in the ]0,1[ interval indicating the rate of geometric cooling for the Simulated Annealing algorithm.

temp is a positive variable indicating the initial temperature for the Simulated Annealing algorithm.

coolfreq is a positive integer indicating the number of iterations of the algorithm between coolings of the temperature. By default, the temperature is cooled at every iteration.

criterion is a character variable, which indicates which criterion is to be used in judging the quality of the subsets. Currently, the "RM", "RV", "GCD", "Tau2", "Xi2", "Zeta2", "ccr12" and "Wald" criteria are supported (see the Details section, the References and the links rm.coef, rv.coef, gcd.coef, tau2.coef, xi2.coef, zeta2.coef and ccr12.coef for further details). The default criterion is "Rm" if parameter r is zero (exploratory and PCA problems), "Wald" if r is equal to one and mat has a "FisherI" attribute set to TRUE (generalized linear models), and "Tau2" otherwise (multivariate linear model framework).

pcindices can be either a vector of ranks of Principal Components that are to be used for comparison with the k-variable subsets (for the GCD criterion only, see gcd.coef) or the default text first_k. The latter will associate PCs 1 to k with each cardinality k that has been requested by the user.

initialsol is a vector, matrix or 3-d array of initial solutions for the simulated annealing search. If a single cardinality is required, initialsol may be a vector of length k, in which case it is used as the initial solution for all nsol final solutions that are requested; a 1 x k matrix (as produced by the $bestsets output value of the algorithm functions anneal, genetic, or improve), or a 1 x k x 1 array (as produced by the $subsets output value), in which case it will be treated as the above k-vector; or an nsol x k matrix, or nsol x k x 1 3-d array, in which case each row (dimension 1) will be used as the initial solution for each of the nsol final solutions requested. If more than one cardinality is requested, initialsol can be a length(kmin:kmax) x kmax matrix (as produced by the $bestsets option of the algorithm functions), in which case each row will be replicated to produced the initial solution for all nsol final solutions requested in each cardinality, or a nsol x kmax x length(kmin:kmax) 3-d array (as produced by the $subsets output option), in which case each row (dimension 1) is interpreted as a different initial solution.

If the exclude and/or include options are used, initialsol must also respect those requirements.

force is a logical variable indicating whether, for large data sets (currently p > 400) the algorithm should proceed anyways, regardless of possible memory problems which may crash the R session.
H Effect description matrix. Not used with the RM, RV or GCD criteria, hence the NULL default value. See the Details section below.

r Expected rank of the effects (H) matrix. Not used with the RM, RV or GCD criteria. See the Details section below.

tolval the tolerance level for the reciprocal of the 2-norm condition number of the correlation/covariance matrix, i.e., for the ratio of the smallest to the largest eigenvalue of the input matrix. Matrices with a reciprocal of the condition number smaller than tolval will activate a restricted-search for well conditioned subsets.

tolsym the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes (A+t(A))/2.

Details

An initial k-variable subset (for k ranging from kmin to kmax) of a full set of p variables is randomly selected and passed on to a Simulated Annealing algorithm. The algorithm then selects a random subset in the neighbourhood of the current subset (neighbourhood of a subset S being defined as the family of all k-variable subsets which differ from S by a single variable), and decides whether to replace the current subset according to the Simulated Annealing rule, i.e., either (i) always, if the alternative subset’s value of the criterion is higher; or (ii) with probability \( \exp \frac{ac-cc}{t} \) if the alternative subset’s value of the criterion (ac) is lower than that of the current solution (cc), where the parameter t (temperature) decreases throughout the iterations of the algorithm. For each cardinality k, the stopping criterion for the algorithm is the number of iterations (niter) which is controlled by the user. Also controlled by the user are the initial temperature (temp) the rate of geometric cooling of the temperature (cooling) and the frequency with which the temperature is cooled, as measured by coolfreq, the number of iterations after which the temperature is multiplied by 1-cooling.

Optionally, the best k-variable subset produced by Simulated Annealing may be passed as input to a restricted local search algorithm, for possible further improvement.

The user may force variables to be included and/or excluded from the k-subsets, and may specify initial solutions.

For each cardinality k, the total number of calls to the procedure which computes the criterion values is nsol x (niter + 1). These calls are the dominant computational effort in each iteration of the algorithm.

In order to improve computation times, the bulk of computations is carried out by a Fortran routine. Further details about the Simulated Annealing algorithm can be found in Reference 1 and in the comments to the Fortran code (in the src subdirectory for this package). For datasets with a very large number of variables (currently p > 400), it is necessary to set the force argument to TRUE for the function to run, but this may cause a session crash if there is not enough memory available.

The function checks for ill-conditioning of the input matrix (specifically, it checks whether the ratio of the input matrix’s smallest and largest eigenvalues is less than tolval). For an ill-conditioned input matrix, the search is restricted to its well-conditioned subsets. The function trim.matrix may be used to obtain a well-conditioned input matrix.
In a general descriptive (Principal Components Analysis) setting, the three criteria $R_m$, $R_v$ and $G_{cd}$ can be used to select good $k$-variable subsets. Arguments $H$ and $r$ are not used in this context. See references [1] and [2] and the Examples for a more detailed discussion.

In the setting of a multivariate linear model, $X = A\Psi + U$, criteria $C_{c12}$, $\tau_2$, $\xi_2$ and $\zeta_2$ can be used to select subsets according to their contribution to an effect characterized by the violation of a reference hypothesis, $C\Psi = 0$ (see reference [3] for further details). In this setting, arguments $\text{mat}$ and $H$ should be set respectively to the usual Total (Hypothesis + Error) and Hypothesis, Sum of Squares and Cross-Products (SSCP) matrices. Argument $r$ should be set to the expected rank of $H$. Currently, for reasons of computational efficiency, criterion $C_{c12}$ is available only when $r \leq 3$. Particular cases in this setting include Linear Discriminant Analysis (LDA), Linear Regression Analysis (LRA), Canonical Correlation Analysis (CCA) with one set of variables fixed and several extensions of these and other classical multivariate methodologies.

In the setting of a generalized linear model, criterion Wald can be used to select subsets according to the (lack of) significance of the discarded variables, as measured by the respective Wald’s statistic (see reference [4] for further details). In this setting arguments $\text{mat}$ and $H$ should be set respectively to $FI$ and $FI \times b \times t(b) \times FI$, where $b$ is a column vector of variable coefficient estimates and $FI$ is an estimate of the corresponding Fisher information matrix.

The auxiliary functions $\text{lmHmat}$, $\text{ldaHmat}$, $\text{glhHmat}$ and $\text{glmHmat}$ are provided to automatically create the matrices $\text{mat}$ and $H$ in all the cases considered.

Value

A list with five items:

- **subsets**: An $n_{sol} \times k_{max} \times \text{length}(k_{min}:k_{max})$ 3-dimensional array, giving for each cardinality (dimension 3) and each solution (dimension 1) the list of variables (referenced by their row/column numbers in matrix $\text{mat}$) in the subset (dimension 2). (For cardinalities smaller than $k_{max}$, the extra final positions are set to zero).

- **values**: An $n_{sol} \times \text{length}(k_{min}:k_{max})$ matrix, giving for each cardinality (columns), the criterion values of the $n_{sol}$ (rows) subsets obtained.

- **bestvalues**: A length($k_{min}:k_{max}$) vector giving the best values of the criterion obtained for each cardinality. If $\text{improvement}$ is TRUE, these values result from the final restricted local search algorithm (and may therefore exceed the largest value for that cardinality in $\text{values}$).

- **bestsets**: A length($k_{min}:k_{max}$) x $k_{max}$ matrix, giving, for each cardinality (rows), the variables (referenced by their row/column numbers in matrix $\text{mat}$) in the best $k$-subset that was found.

- **call**: The function call which generated the output.

References


See Also

rm.coef, rv.coef, gcd.coef, tau2.coef, xi2.coef, zeta2.coef, ccr12.coef, genetic, anneal, eleaps, trim.matrix, lmHmat, ldaHmat, glhHmat, glmHmat.

Examples

```r
## --------------------------------------------------------------------
##
## (1) For illustration of use, a small data set with very few iterations
## of the algorithm, using the RM criterion.
##
## data(swiss)
anneal(cor(swiss),2,3,nsol=4,niter=10,criterion="RM")

##$subsets
## , , Card.2
##     Var.1 Var.2 Var.3
##Solution 1 3 6 0
##Solution 2 4 5 0
##Solution 3 1 2 0
##Solution 4 3 6 0
##
## , , Card.3
##     Var.1 Var.2 Var.3
##Solution 1 4 5 6
##Solution 2 3 5 6
##Solution 3 3 4 6
##Solution 4 4 5 6
##
##$values
##  card.2  card.3
##Solution 1 0.8016409 0.9043760
##Solution 2 0.7982296 0.8769672
##Solution 3 0.7945390 0.8777509
##Solution 4 0.8016409 0.9043760

##$bestvalues
##   Card.2  Card.3
##0.8016409 0.9043760
```
```r
##$bestsets
## [[1]]
##       Var.1 Var.2 Var.3
##Card.2    3    6    0
##Card.3    4    5    6
##
##$call
## anneal(cor(swiss), 2, 3, nsol = 4, niter = 10, criterion = "RM")
##
## # (2) An example excluding variable number 6 from the subsets.
##
## data(swiss)
## anneal(cor(swiss), 2, 3, nsol = 4, niter = 10, criterion = "RM", exclude = c(6))

##$subsets
## , , Card.2
##
## #
## #   Var.1 Var.2 Var.3
## #Solution 1  4    5    0
## #Solution 2  4    5    0
## #Solution 3  4    5    0
## #Solution 4  4    5    0
##
## , , Card.3
##
## #
## #   Var.1 Var.2 Var.3
## #Solution 1  1    2    5
## #Solution 2  1    2    5
## #Solution 3  1    2    5
## #Solution 4  1    4    5
##
##
####$values
##
## #   card.2   card.3
## #Solution 1 0.7982296 0.8791856
## #Solution 2 0.7982296 0.8791856
## #Solution 3 0.7982296 0.8791856
## #Solution 4 0.7982296 0.8686515

##$bestvalues
##       Card.2 Card.3
##0.7982296 0.8791856

##$bestsets
## [[1]]
##       Var.1 Var.2 Var.3
##Card.2    4    5    0
##Card.3    1    2    5

##$call
## anneal(cor(swiss), 2, 3, nsol = 4, niter = 10, criterion = "RM")
```
## exclude=c(6))
## --------------------------------------------------------------------

## (3) An example specifying initial solutions: using the subsets produced
## by simulated annealing for one criterion (RM, by default) as initial
## solutions for the simulated annealing search with a different criterion.

data(swiss)
rmresults<-anneal(cor(swiss), 2, 3, nsol=4, niter=10, setseed=TRUE)
anneal(cor(swiss), 2, 3, nsol=4, niter=10, criterion="gcd",
initialsol=rmresults$subsets)

###$subsets
### , , Card.2
###
### Var.1 Var.2 Var.3
###Solution 1 3 6 0
###Solution 2 3 6 0
###Solution 3 3 6 0
###Solution 4 3 6 0
###
### , , Card.3
###
### Var.1 Var.2 Var.3
###Solution 1 4 5 6
###Solution 2 4 5 6
###Solution 3 3 4 6
###Solution 4 4 5 6
###
###$values
### card.2 card.3
###Solution 1 0.8487026 0.9253720
###Solution 2 0.8487026 0.9253720
###Solution 3 0.8487026 0.7988640
###Solution 4 0.8487026 0.9253720
###
###$bestvalues
### Card.2 Card.3
###0.8487026 0.9253720
###
###$bestsets
### Var.1 Var.2 Var.3
###Card.2 3 6 0
###Card.3 4 5 6
###
###$call
###anneal(cor(swiss), 2, 3, nsol = 4, niter = 10, criterion = "gcd",
### initialsol = rmresults$subsets)
(4) An example of subset selection in the context of Multiple Linear Regression. Variable 5 (average car price) in the Cars93 MASS library data set is regressed on 13 other variables. A best subset of linear predictors is sought, using the "TAU_2" criterion which, in the case of a Linear Regression, is merely the standard Coefficient of Determination, R^2 (like the other three criteria for the multivariate linear hypothesis, "XI_2", "CCR1_2" and "ZETA_2").

```
library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,5])

names(Cars93[,5,drop=FALSE])
## [1] "Price"

colnames(CarsHmat$mat)
## [1] "MPG.city" "MPG.highway" "EngineSize"
## [4] "Horsepower" "RPM" "Rev.per.mile"
## [7] "Fuel.tank.capacity" "Passengers" "Length"
## [10] "Wheelbase" "Width" "Turn.circle"
## [13] "Weight"

anneal(CarsHmat$mat, kmin=4, kmax=6, H=CarsHmat$H, r=1, crit="tau2")
```

````
## $subsets
## , , Card.4
## Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Solution 1 4 5 10 11 0 0
## , , Card.5
## Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Solution 1 4 5 10 11 12 0
## , , Card.6
## Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Solution 1 4 5 9 10 11 12

## $values
## card.4 card.5 card.6
## Solution 1 0.7143794 0.7241457 0.731015

## $bestvalues
## Card.4 Card.5 Card.6
## 0.7143794 0.7241457 0.731015

## $bestsets
## Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Card.4 4 5 10 11 0 0
## Card.5  4  5  10  11  12  0
## Card.6  4  5  9  10  11  12

## $call
## anneal(mat = CarsHmat$mat, kmin = 4, kmax = 6, criterion = "xi2",
## H = CarsHmat$H, r = 1)
##
## --------------------------------------------------------------------
## (5) A Linear Discriminant Analysis example with a very small data set.
## We consider the Iris data and three groups, defined by species (setosa,
## versicolor and virginica). The goal is to select the 2- and 3-variable
## subsets that are optimal for the linear discrimination (as measured
## by the "CCR1_2" criterion).

data(iris)
irisHmat <- ldaHmat(iris[1:4], iris$Species)
anneal(irisHmat$mat, kmin=2, kmax=3, H=irisHmat$H, r=2, crit="ccr12")

## $subsets
## , , Card.2
## # Solution 1 1 3 0
## #
## , , Card.3
## #
## # Solution 1 2 3 4
## #
## # $values
## #
## # Solution 1 0.9589055 0.967897
## #
## # $bestvalues
## #
## # 0.9589055 0.9678971
## #
## # $bestsets
## #
## # Card.2 1 3 0
## # Card.3 2 3 4
## #
## # $call
## # anneal(irisHmat$mat, kmin=2, kmax=3, H=irisHmat$H, r=2, crit="ccr12")
## #
## # --------------------------------------------------------------------
## (6) An example of subset selection in the context of a Canonical
## Correlation Analysis. Two groups of variables within the Cars93
## MASS library data set are compared. The goal is to select 4- to
## 6-variable subsets of the 13-variable 'X' group that are optimal in
## terms of preserving the canonical correlations, according to the
## "XI_2" criterion (Warning: the 3-variable 'Y' group is kept
## intact; subset selection is carried out in the 'X'
## group only). The 'tolsym' parameter is used to relax the symmetry
## requirements on the effect matrix H which, for numerical reasons,
## is slightly asymmetric. Since corresponding off-diagonal entries of
## matrix H are different, but by less than tolsym, H is replaced
## by its symmetric part: (H+t(H))/2.

library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,4:6])
names(Cars93[,4:6])
colnames(CarsHmat$mat)
## [1] "MPG.city" "MPG.highway" "EngineSize"
## [4] "Horsepower" "RPM" "Rev.per.mile"
## [7] "Fuel.tank.capacity" "Passengers" "Length"
## [10] "Wheelbase" "Width" "Turn.circle"
## [13] "Weight"

anneal(CarsHmat$mat, kmin=4, kmax=6, H=CarsHmat$H, r=CarsHmat$r,
crit="tau2", tolsym=1e-9)

## $subsets
## , , Card.4
##   Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Solution 1  4  9 10 11  0  0
## , , Card.5
##   Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Solution 1  3  4  9 11 11  0
## , , Card.6
##   Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Solution 1  3  4  5  9 10 11

## $values
## card.4 card.5 card.6
## Solution 1 0.2818772 0.2943742 0.3057831

## $bestvalues
## Card.4 Card.5 Card.6
## 0.2818772 0.2943742 0.3057831
##
## $bestsets
## Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## Card.4  4  9  10  11  0  0
## Card.5  3  4  9  10  11  0
## Card.6  3  4  5  9  10  11
##
## $call
## anneal(mat = CarsHmat$mat, kmin = 4, kmax = 6, criterion = "xi2",
## H = CarsHmat$H, r = CarsHmat$r, tolsym = 1e-09)
##
## Warning message:
##
## The effect description matrix (H) supplied was slightly asymmetric:
## symmetric entries differed by up to 3.63797880709171e-12.
## (less than the 'tolsym' parameter).
## The H matrix has been replaced by its symmetric part.
## in: validnovcrit(mat, criterion, H, r, p, tolval, tolsym)

# (?) An example of variable selection in the context of a logistic regression model. We consider the last 100 observations of the iris data set (versicolor and virginica species) and try to find the best variable subsets for the model that takes species as response variable.

data(iris)
iris2sp <- iris[iris$Species != "setosa",]
logrfit <- glm(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, iris2sp,family=binomial)
Hmat <- glmHmat(logrfit)
anneal(Hmat$mat,1,3,H=Hmat$H,r=1,criterion="Wald")

## $subsets
## , , Card.1
##
## Var.1 Var.2 Var.3
## Solution 1  4  0  0
##
## , , Card.2
##
## Var.1 Var.2 Var.3
## Solution 1  1  3  0
##
## , , Card.3
##
## Var.1 Var.2 Var.3
## Solution 1  2  3  4

## $values
## Solution

1. 4.894554 3.522885 1.060121

### bestvalues

<table>
<thead>
<tr>
<th>Card.1</th>
<th>Card.2</th>
<th>Card.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.894554</td>
<td>3.522885</td>
<td>1.060121</td>
</tr>
</tbody>
</table>

### bestsets

<table>
<thead>
<tr>
<th>Var.1</th>
<th>Var.2</th>
<th>Var.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card.1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Card.2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Card.3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

### call

```
anneal(mat = Hmat$mat, kmin = 1, kmax = 3, criterion = "Wald",
H = Hmat$H, r = 1)
```

It should be stressed that, unlike other criteria in the subselect package, the Wald criterion is not bounded above by 1 and is a decreasing function of subset quality, so that the 3-variable subsets do, in fact, perform better than their smaller-sized counterparts.

### Description

**First Squared Canonical Correlation for a multivariate linear hypothesis**

Computes the first squared canonical correlation. The maximization of this criterion is equivalent to the maximization of the Roy first root.

### Usage

```
crr12.coef(mat, H, r, indices,
tolval=10*.Machine$double.eps, tolsym=1000*.Machine$double.eps)
```

### Arguments

- **mat**: the Variance or Total sums of squares and products matrix for the full data set.
- **H**: the Effect description sums of squares and products matrix (defined in the same way as the mat matrix).
- **r**: the Expected rank of the H matrix. See the Details section below.
- **indices**: a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different k-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.
tolval the tolerance level to be used in checks for ill-conditioning and positive-definiteness of the 'total' and 'effects' (H) matrices. Values smaller than tolval are considered equivalent to zero.

tolsym the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes \((A + t(A))/2\).

Details

Different kinds of statistical methodologies are considered within the framework, of a multivariate linear model:

\[ X = A\Psi + U \]

where \(X\) is the (nxp) data matrix of original variables, \(A\) is a known (nxp) design matrix, \(\Psi\) an (qxp) matrix of unknown parameters and \(U\) an (nxp) matrix of residual vectors. The \(ccr^2\) index is related to the traditional test statistic (the Roy first root) and measures the contribution of each subset to an Effect characterized by the violation of a linear hypothesis of the form \(C\Psi = 0\), where \(C\) is a known coefficient matrix of rank r. The Roy first root is the first eigenvalue of \(HE^{-1}\), where \(H\) is the Effect matrix and \(E\) is the Error matrix. The index \(ccr^2\) is related to the Roy first root \(\lambda_1\) by:

\[ ccr^2 = \frac{\lambda_1}{1 + \lambda_1} \]

The fact that indices can be a matrix or 3-d array allows for the computation of the \(ccr^2\) values of subsets produced by the search functions \texttt{anneal}, \texttt{genetic}, \texttt{improve} and \texttt{anneal} (whose output option $subsets$ are matrices or 3-d arrays), using a different criterion (see the example below).

Value

The value of the \(ccr^2\) coefficient.

Examples

### 1) A Linear Discriminant Analysis example with a very small data set.

We considered the Iris data and three groups, defined by species (setosa, versicolor and virginica).

```r
data(iris)
irisHmat <- ldaHmat(iris[,1:4],iris$Species)
ccr12.coef(irisHmat$mat,H=irisHmat$H,r=2,c(1,3))
## [1] 0.9589055
```

### 2) An example computing the value of the ccr1_2 criteria for two subsets produced when the anneal function attempted to optimize the zeta_2 criterion (using an absurdly small number of iterations).
eleaps

A Leaps and Bounds Algorithm for finding the best variable subsets

Description

An exact Algorithm for optimizing criteria that measure the quality of k-dimensional variable subsets as approximations to a given set of variables, or to a set of its Principal Components.

Usage

eleaps(mat,kmin=length(include)+1,kmax=ncol(mat)-length(exclude)-1,nsol=1,
exclude=NULL,include=NULL,criterion="default",pcindices="first_k",timelimit=15,
H=NULL,r=0,tolval=1000*.Machine$double.eps,
tolsym=1000*.Machine$double.eps,maxaperr=1E-4)

Arguments

mat a covariance/correlation, information or sums of squares and products matrix of the variables from which the k-subset is to be selected. See the Details section below.

kmin the cardinality of the smallest subset that is wanted.

kmax the cardinality of the largest subset that is wanted.

nsol the number of different subsets of each cardinality that are requested.

exclude a vector of variables (referenced by their row/column numbers in matrix mat) that are to be forcibly excluded from the subsets.

include a vector of variables (referenced by their row/column numbers in matrix mat) that are to be forcibly included in the subsets.

criterion Character variable, which indicates which criterion is to be used in judging the quality of the subsets. Currently, the "Rm", "Rv", "Gcd", "Tau2", "Xi2", "Zeta2", "Ccr12" and "Wald" criteria are supported (see the Details section, the References and the links rm.coef, rv.coef, gcd.coef, tau2.coef, xi2.coef, zeta2.coef, ccr12.coef and wald.coef for further details). The default criterion is "Rm" if parameter r is zero (exploratory and PCA problems), "Wald" if r is equal to one and mat has a "FisherI" attribute set to TRUE (generalized linear models), and "Tau2" otherwise (multivariate linear model framework).
either a vector of ranks of Principal Components that are to be used for comparison with the k-variable subsets (for the Gcd criterion only, see `gcd.coef`) or the default text `first_k`. The latter will associate PCs 1 to k with each cardinality k that has been requested by the user.

a user specified limit (in seconds) for the maximum time allowed to conduct the search. After this limit is exceeded, `eleaps` exits with a warning message stating that it was not possible to find the optimal subsets within the allocated time.

Effect description matrix. Not used with the Rm, Rv or Gcd criteria, hence the NULL default value. See the Details section below.

Expected rank of the effects (H) matrix. Not used with the Rm, Rv or Gcd criteria. See the Details section below.

the tolerance level for the reciprocal of the 2-norm condition number of the correlation/covariance or sums of squares matrix, i.e., for the ratio of the smallest to the largest eigenvalue of the input matrix. Matrices with a reciprocal of the condition number smaller than `tolval` will activate a restricted-search (for well conditioned sets as defined by the value of the `maxaperr` argument) algorithm.

the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes $(A+t(A))/2$.

the tolerance level for the relative rounding error of the criterion. When a restricted search in employed subsets where a first order estimate of this error is higher than `maxaperr` will be excluded from the analysis.

Details

For each cardinality k (with k ranging from `kmin` to `kmax`), `eleaps` performs a branch and bound search for the best `nsol` variable subsets according to a specified criterion. Leaps implements Duarte Silva’s adaptation (references [2] and [3]) of Furnival and Wilson’s Leaps and Bounds Algorithm (reference [4]) for variable selection in Regression Analysis. If the search is not completed within a user defined time limit, `eleaps` exits with a warning message.

The user may force variables to be included and/or excluded from the k-subsets.

In order to improve computation times, the bulk of computations are carried out by C++ routines. Further details about the Algorithm can be found in references [2] and [3] and in the comments to the C++ code. A discussion of the criteria considered can be found in References [1] and [3].

The function checks for ill-conditioning of the input matrix (specifically, it checks whether the ratio of the input matrix’s smallest and largest eigenvalues is less than `tolval`). For an ill-conditioned input matrix, the search is restricted to its well-conditioned subsets. The function `trim.matrix` may be used to obtain a well-conditioned input matrix.

In a general descriptive (Principal Components Analysis) setting, the three criteria Rm, Rv and Gcd can be used to select good k-variable subsets. Arguments H and r are not used in this context. See reference [1] and the Examples for a more detailed discussion.

In the setting of a multivariate linear model, $X = A\Psi + U$, criteria Ccr12, Tau2, Xi2 and Zeta2 can be used to select subsets according to their contribution to an effect characterized by the violation
of a reference hypothesis, \( C \Psi = 0 \) (see reference [3] for further details). In this setting, arguments \( \text{mat} \) and \( H \) should be set respectively to the usual Total (Hypothesis + Error) and Hypothesis, Sum of Squares and Cross-Products (SSCP) matrices. Argument \( r \) should be set to the expected rank of \( H \). Currently, for reasons of computational efficiency, criterion Ccr12 is available only when \( r \leq 3 \). Particular cases in this setting include Linear Discriminant Analysis (LDA), Linear Regression Analysis (LRA), Canonical Correlation Analysis (CCA) with one set of variables fixed, and several extensions of these and other classical multivariate methodologies.

In the setting of a generalized linear model, criterion Wald can be used to select subsets according to the (lack of) significance of the discarded variables, as measured by the respective Wald’s statistic (see reference [5] for further details). In this setting arguments \( \text{mat} \) and \( H \) should be set respectively to \( \text{FI} \) and \( \text{FI} \% \% b \% \% t(b) \% \% \text{FI} \), where \( b \) is a column vector of variable coefficient estimates and \( \text{FI} \) is an estimate of the corresponding Fisher information matrix.

The auxiliary functions \( \text{lmHmat}, \text{ldaHmat}, \text{glhHmat} \) and \( \text{glmHmat} \) are provided to automatically create the matrices \( \text{mat} \) and \( H \) in all the cases considered.

### Value

A list with five items:

- **subsets**
  An \( \text{n.sol} \times \text{k.max} \times \text{length(k.min:k.max)} \) 3-dimensional array, giving for each cardinality (dimension 3) and each solution (dimension 1) the list of variables (referenced by their row/column numbers in matrix \( \text{mat} \)) in the subset (dimension 2). (For cardinalities smaller than \( \text{k.max} \), the extra final positions are set to zero).

- **values**
  An \( \text{n.sol} \times \text{length(k.min:k.max)} \) matrix, giving for each cardinality (columns), the criterion values of the best \( \text{n.sol} \) (rows) subsets according to the chosen criterion.

- **bestvalues**
  A \( \text{length(k.min:k.max)} \) vector giving the overall best values of the criterion for each cardinality.

- **bestsets**
  A \( \text{length(k.min:k.max)} \times \text{k.max} \) matrix, giving, for each cardinality (rows), the variables (referenced by their row/column numbers in matrix \( \text{mat} \)) in the best \( k \)-subset.

- **call**
  The function call which generated the output.

### References


See Also

rm.coef, rv.coef, gcd.coef, tau2.coef, wald.coef, xi2.coef, zeta2.coef, ccr12.coef, anneal, genetic, anneal, trim.matrix, lmHmat, ldaHmat, glhHmat, glmHmat.

Examples

```r
## 1) For illustration of use, a small data set.
## Subsets of variables of all cardinalities are sought using the
## RM criterion.
##
data(swiss)
eleaps(cor(swiss), nsol=3, criterion="RM")
```

```
# $subsets
# , , Card.1
#
#   Var.1 Var.2 Var.3 Var.4 Var.5
# Solution 1 3 0 0 0 0
# Solution 2 1 0 0 0 0
# Solution 3 4 0 0 0 0
#
# , , Card.2
#
#   Var.1 Var.2 Var.3 Var.4 Var.5
# Solution 1 3 6 0 0 0
# Solution 2 4 5 0 0 0
# Solution 3 1 2 0 0 0
#
# , , Card.3
#
#   Var.1 Var.2 Var.3 Var.4 Var.5
# Solution 1 4 5 6 0 0
# Solution 2 1 2 5 0 0
# Solution 3 3 4 6 0 0
#
# , , Card.4
#
#   Var.1 Var.2 Var.3 Var.4 Var.5
# Solution 1 2 4 5 6 0
# Solution 2 1 2 5 6 0
# Solution 3 1 4 5 6 0
#
# , , Card.5
#
#   Var.1 Var.2 Var.3 Var.4 Var.5
# Solution 1 1 2 3 5 6
# Solution 2 1 2 4 5 6
```
## Solution 3
2 3 4 5 6

### $values

card.1 card.2 card.3 card.4 card.5

### Solution 1
0.6729689 0.8016409 0.9043760 0.9510757 0.9804629

### Solution 2
0.6286185 0.7982296 0.8791856 0.9506434 0.9776338

### Solution 3
0.6286130 0.7945390 0.8777509 0.9395708 0.9752551

### $bestvalues

<table>
<thead>
<tr>
<th>Card.1</th>
<th>Card.2</th>
<th>Card.3</th>
<th>Card.4</th>
<th>Card.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6729689</td>
<td>0.8016409</td>
<td>0.9043760</td>
<td>0.9510757</td>
<td>0.9804629</td>
</tr>
</tbody>
</table>

### $bestsets

<table>
<thead>
<tr>
<th>Var.1</th>
<th>Var.2</th>
<th>Var.3</th>
<th>Var.4</th>
<th>Var.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card.1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Card.2</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Card.3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Card.4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Card.5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

### $call

eleaps(cor(swiss), nsol = 3, criterion="RM")

data(swiss)
eleaps(cor(swiss), nsol = 3, criterion="RM")

### subsets

<table>
<thead>
<tr>
<th>, , Card.2</th>
</tr>
</thead>
</table>

### Solution 1
1 4 5 0

### Solution 2
1 2 0

### Solution 3
1 3 0

<table>
<thead>
<tr>
<th>, , Card.3</th>
</tr>
</thead>
</table>

### Solution 1
1 2 5

### Solution 2
1 4 5

### Solution 3
2 4 5

### $values

<table>
<thead>
<tr>
<th>card.2</th>
<th>card.3</th>
</tr>
</thead>
</table>
## Solution 1
0.7982296 0.8791856

## Solution 2
0.7945390 0.8686515

## Solution 3
0.7755232 0.8628693

## bestvalues
Card.2 Card.3
0.7982296 0.8791856

## bestsets
Var.1 Var.2 Var.3
Card.2 4 5 0
Card.3 1 2 5

## call
eleaps(cor(swiss), 2, 3, exclude = 6, nsol = 3, criterion = "gcd")

data(swiss)
eleaps(cor(swiss), 2, 3, criterion = "gcd", pcindices = 1:3, nsol = 3)

## subsets

### , , Card.2

### Var.1 Var.2 Var.3

#### Solution 1
4 5 0

#### Solution 2
5 6 0

#### Solution 3
4 6 0

### , , Card.3

### Var.1 Var.2 Var.3

#### Solution 1
4 5 6

#### Solution 2
3 5 6

#### Solution 3
2 5 6

## values
card.2 card.3

#### Solution 1
0.7831827 0.9253684

#### Solution 2
0.7475630 0.8459302

#### Solution 3
0.7383665 0.8243032

## bestvalues
Card.2 Card.3
0.7831827 0.9253684

## 3) Searching for 2- and 3-dimensional subsets that best approximate the spaces generated by the first three Principal Components

data(swiss)
eleaps(cor(swiss), 2, 3, exclude = 6, nsol = 3, criterion = "gcd", pcindices = 1:3, nsol = 3)
```r
###$bestsets
### Var.1 Var.2 Var.3
###Card.2 4 5 0
###Card.3 4 5 6
###
###$call
eleaps(cor(swiss), 2, 3, criterion = "gcd", pcindices = 1:3, nsol = 3)

###
### 4) An example of subset selection in the context of Multiple Linear Regression. Variable 5 (average car price) in the Cars93 MASS library data set is regressed on 13 other variables. A best subset of linear predictors is sought, using the default criterion ("TAU_2") which, in the case of a Linear Regression, is merely the standard Coefficient of Determination, R^2 (as are the other three criteria for the multivariate linear hypothesis, "XI_2", "CCR1_2" and "ZETA_2").
###
library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,5])
names(Cars93[,5,drop=FALSE])
## [1] "Price"
colnames(CarsHmat$mat)
## [1] "MPG.city" "MPG.highway" "EngineSize"
## [4] "Horsepower" "RPM" "Rev.per.mile"
## [7] "Fuel.tank.capacity" "Passengers" "Length"
## [10] "Wheelbase" "Width" "Turn.circle"
## [13] "Weight"
eleaps(CarsHmat$mat, kmin=4, kmax=6, H=CarsHmat$H, r=1)

### $subsets
### , , Card.4
###
###   Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
### Solution 1 4 5 10 11 0 0
###
### , , Card.5
###
###   Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
### Solution 1 4 5 10 11 12 0
###
### , , Card.6
###
###   Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
```

## Solution 1 4 5 9 10 11 12
##
## $values
card.4 card.5 card.6
## Solution 1 0.7143794 0.7241457 0.731015
##
## $bestvalues
Card.4 Card.5 Card.6
0.7143794 0.7241457 0.731015
## $bestsets
Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
Card.4 4 5 10 11 0 0
Card.5 4 5 10 11 12 0
Card.6 4 5 9 10 11 12

# 5) A Linear Discriminant Analysis example with a very small data set.
# We consider the Iris data and three groups, defined by species (setosa,
# versicolor and virginica). The goal is to select the 2- and 3-variable
# subsets that are optimal for the linear discrimination (as measured
# by the "CCR1_2" criterion).

data(iris)
irisHmat <- ldaHmat(iris[1:4], iris$Species)
eleaps(irisHmat$mat, kmin=2, kmax=3, H=irisHmat$H, r=2, crit="ccr12")

## $subsets
## , , Card.2
## Var.1 Var.2 Var.3
## Solution 1 1 3 0
##
## , , Card.3
## Var.1 Var.2 Var.3
## Solution 1 2 3 4

##
## $values
card.2 card.3
## Solution 1 0.9589055 0.967897
##
## $bestvalues
Card.2 Card.3
0.9589055 0.967897
## $bestsets
Var.1 Var.2 Var.3
Card.2 1 3 0
6) An example of subset selection in the context of a Canonical Correlation Analysis. Two groups of variables within the Cars93 MASS library data set are compared. The goal is to select 4- to 6-variable subsets of the 13-variable 'X' group that are optimal in terms of preserving the canonical correlations, according to the "ZETA_2" criterion (Warning: the 3-variable 'Y' group is kept intact; subset selection is carried out in the 'X' group only). The 'tolsym' parameter is used to relax the symmetry requirements on the effect matrix H which, for numerical reasons, is slightly asymmetric. Since corresponding off-diagonal entries of matrix H are different, but by less than tolsym, H is replaced by its symmetric part: (H+t(H))/2.

```
library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,4:6])
names(Cars93[,4:6])

colnames(CarsHmat$mat)
## [1] "MPG.city"    "MPG.highway"  "EngineSize"
## [4] "Horsepower"  "RPM"         "Rev.per.mile"
## [7] "Fuel.tank.capacity" "Passengers"  "Length"
## [10] "Wheelbase"   "Width"       "Turn.circle"
## [13] "Weight"

eleaps(CarsHmat$mat, kmin=4, kmax=6, H=CarsHmat$H, r=3, crit="zeta2", tolsym=1e-9)

## $subsets
## , , Card.4
## #
## # Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## # Solution 1  3  4 10 11  0  0
## 
## , , Card.5
## #
## # Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## # Solution 1  4  5  9 10 11  0
## 
## , , Card.6
## #
## # Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## # Solution 1  4  5  9 10 11 12
```
# $values
# card.4  card.5  card.6
# Solution 1 0.4827353 0.5018922 0.5168627
#
# $bestvalues
# Card.4  Card.5  Card.6
# 0.4827353 0.5018922 0.5168627
#
# $bestsets
# Var.1  Var.2  Var.3  Var.4  Var.5  Var.6
# Card.4 3 4 10 11 0 0
# Card.5 4 5 9 10 11 0
# Card.6 4 5 9 10 11 12
#
# Warning message:
# The effect description matrix (H) supplied was slightly asymmetric:
# symmetric entries differed by up to 3.63797880709171e-12.
# (less than the 'tolsym' parameter).
# The H matrix has been replaced by its symmetric part.
# in: validnovcrit(mat, criterion, H, r, p, tolval, tolsym)
#
# 7) An example of variable selection in the context of a logistic regression model. We consider the last 100 observations of the iris data set (versicolor an veriginica species) and try to find the best variable subsets for the model that takes species as response variable.

data(iris)
iris2sp <- iris[iris$Species != "setosa",]
logrfit <- glm(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, iris2sp, family=binomial)
Hmat <- glmHmat(logrfit)
eleaps(Hmat$mat, H=Hmat$H, r=1, criterion="Wald", nsol=3)

# $subsets
# , , Card.1
#
# Var.1  Var.2  Var.3
# Solution 1 4 0 0
# Solution 2 1 0 0
# Solution 3 3 0 0
#
# , , Card.2
#
# Var.1  Var.2  Var.3
# Solution 1 1 3 0
# Solution 2 3 4 0
Sixty-two economic indicators from 99 Portuguese farms.

Description

This data set is a very small subset of economic data regarding Portuguese farms in the mid-1990s, from Portugal’s Ministry of Agriculture.

Usage

farm
Format

A 99x62 matrix. The 62 columns are numeric economic indicators, referenced by their database code. Monetary units are in thousands of Escudos (Portugal’s pre-Euro currency).

<table>
<thead>
<tr>
<th>Column Number</th>
<th>Column Name</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[.1] R15</td>
<td>1000 Escudos</td>
<td>Total Standard Gross Margins (SGM)</td>
<td></td>
</tr>
<tr>
<td>[.2] R24</td>
<td>Hectares</td>
<td>Total land surface</td>
<td></td>
</tr>
<tr>
<td>[.3] R35</td>
<td>Hectares</td>
<td>Total cultivated surface</td>
<td></td>
</tr>
<tr>
<td>[.4] R36</td>
<td>Man Work Units</td>
<td>Total Man Work Units</td>
<td></td>
</tr>
<tr>
<td>[.5] R46</td>
<td>1000 Escudos</td>
<td>Land Capital</td>
<td></td>
</tr>
<tr>
<td>[.6] R59</td>
<td>1000 Escudos</td>
<td>Total Capital (without forests)</td>
<td></td>
</tr>
<tr>
<td>[.7] R65</td>
<td>1000 Escudos</td>
<td>Total Loans and Debts</td>
<td></td>
</tr>
<tr>
<td>[.8] R72</td>
<td>1000 Escudos</td>
<td>Total Investment</td>
<td></td>
</tr>
<tr>
<td>[.9] R79</td>
<td>1000 Escudos</td>
<td>Subsidies for Investment</td>
<td></td>
</tr>
<tr>
<td>[.10] R86</td>
<td>1000 Escudos</td>
<td>Gross Plant Product Formation</td>
<td></td>
</tr>
<tr>
<td>[.11] R91</td>
<td>1000 Escudos</td>
<td>Gross Animal Product Formation</td>
<td></td>
</tr>
<tr>
<td>[.12] R104</td>
<td>1000 Escudos</td>
<td>Current Subsidies</td>
<td></td>
</tr>
<tr>
<td>[.13] R110</td>
<td>1000 Escudos</td>
<td>Wheat Production</td>
<td></td>
</tr>
<tr>
<td>[.14] R111</td>
<td>1000 Escudos</td>
<td>Maize Production</td>
<td></td>
</tr>
<tr>
<td>[.15] R113</td>
<td>1000 Escudos</td>
<td>Other Cereals (except rice) Production</td>
<td></td>
</tr>
<tr>
<td>[.16] R114</td>
<td>1000 Escudos</td>
<td>Dried Legumes Production</td>
<td></td>
</tr>
<tr>
<td>[.17] R115</td>
<td>1000 Escudos</td>
<td>Potato Production</td>
<td></td>
</tr>
<tr>
<td>[.18] R116</td>
<td>1000 Escudos</td>
<td>Industrial horticulture and Melon Production</td>
<td></td>
</tr>
<tr>
<td>[.19] R117</td>
<td>1000 Escudos</td>
<td>Open-air horticultural Production</td>
<td></td>
</tr>
<tr>
<td>[.20] R118</td>
<td>1000 Escudos</td>
<td>Horticultural forcing Production</td>
<td></td>
</tr>
<tr>
<td>[.21] R119</td>
<td>1000 Escudos</td>
<td>Flower Production</td>
<td></td>
</tr>
<tr>
<td>[.22] R121</td>
<td>1000 Escudos</td>
<td>Sub-products Production</td>
<td></td>
</tr>
<tr>
<td>[.23] R122</td>
<td>1000 Escudos</td>
<td>Fruit Production</td>
<td></td>
</tr>
<tr>
<td>[.24] R123</td>
<td>1000 Escudos</td>
<td>Olive Production</td>
<td></td>
</tr>
<tr>
<td>[.25] R124</td>
<td>1000 Escudos</td>
<td>Wine Production</td>
<td></td>
</tr>
<tr>
<td>[.26] R125</td>
<td>1000 Escudos</td>
<td>Horses</td>
<td></td>
</tr>
<tr>
<td>[.27] R126</td>
<td>1000 Escudos</td>
<td>Bovines (excluding milk)</td>
<td></td>
</tr>
<tr>
<td>[.28] R127</td>
<td>1000 Escudos</td>
<td>Milk and dairy products</td>
<td></td>
</tr>
<tr>
<td>[.29] R129</td>
<td>1000 Escudos</td>
<td>Sheep</td>
<td></td>
</tr>
<tr>
<td>[.30] R132</td>
<td>1000 Escudos</td>
<td>Goats</td>
<td></td>
</tr>
<tr>
<td>[.31] R135</td>
<td>1000 Escudos</td>
<td>Pigs</td>
<td></td>
</tr>
<tr>
<td>[.32] R137</td>
<td>1000 Escudos</td>
<td>Birds</td>
<td></td>
</tr>
<tr>
<td>[.33] R140</td>
<td>1000 Escudos</td>
<td>Bees</td>
<td></td>
</tr>
<tr>
<td>[.34] R142</td>
<td>1000 Escudos</td>
<td>Other animals (except rabbits)</td>
<td></td>
</tr>
<tr>
<td>[.35] R144</td>
<td>1000 Escudos</td>
<td>Wood production</td>
<td></td>
</tr>
<tr>
<td>[.36] R145</td>
<td>1000 Escudos</td>
<td>Other forest products (except cork)</td>
<td></td>
</tr>
<tr>
<td>[.37] R146</td>
<td>Hectares</td>
<td>Land surface affected to cereals</td>
<td></td>
</tr>
<tr>
<td>[.38] R151</td>
<td>Hectares</td>
<td>Land surface affected to dry legumes</td>
<td></td>
</tr>
<tr>
<td>[.39] R152</td>
<td>Hectares</td>
<td>Land surface affected to potatoes</td>
<td></td>
</tr>
<tr>
<td>[.40] R158</td>
<td>Hectares</td>
<td>Land surface affected to fruits</td>
<td></td>
</tr>
<tr>
<td>[.41] R159</td>
<td>Hectares</td>
<td>Land surface affected to olive trees</td>
<td></td>
</tr>
<tr>
<td>[.42] R160</td>
<td>Hectares</td>
<td>Land surface affected to vineyards</td>
<td></td>
</tr>
<tr>
<td>[.43] R164</td>
<td>Hectares</td>
<td>Fallow land surface area</td>
<td></td>
</tr>
</tbody>
</table>
gcd.coef

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[.44]</td>
<td>R166</td>
<td>Hectares</td>
<td>Forest surface area</td>
</tr>
<tr>
<td>[.45]</td>
<td>R168</td>
<td>Head</td>
<td>Bovines</td>
</tr>
<tr>
<td>[.46]</td>
<td>R174</td>
<td>Head</td>
<td>Adult sheep</td>
</tr>
<tr>
<td>[.47]</td>
<td>R176</td>
<td>Head</td>
<td>Adult goats</td>
</tr>
<tr>
<td>[.48]</td>
<td>R178</td>
<td>Head</td>
<td>Adult pigs</td>
</tr>
<tr>
<td>[.49]</td>
<td>R209</td>
<td>Kg/hectare</td>
<td>Maize yield</td>
</tr>
<tr>
<td>[.50]</td>
<td>R211</td>
<td>Kg/hectare</td>
<td>Barley yield</td>
</tr>
<tr>
<td>[.51]</td>
<td>R214</td>
<td>Kg/hectare</td>
<td>Potato yield</td>
</tr>
<tr>
<td>[.52]</td>
<td>R215</td>
<td>L/cow/year</td>
<td>Cow milk productivity</td>
</tr>
<tr>
<td>[.53]</td>
<td>R233</td>
<td>1000 Escudos</td>
<td>Wages and social expenditure</td>
</tr>
<tr>
<td>[.54]</td>
<td>R237</td>
<td>1000 Escudos</td>
<td>Taxes and tariffs</td>
</tr>
<tr>
<td>[.55]</td>
<td>R245</td>
<td>1000 Escudos</td>
<td>Interest and financial costs</td>
</tr>
<tr>
<td>[.56]</td>
<td>R250</td>
<td>1000 Escudos</td>
<td>Total real costs</td>
</tr>
<tr>
<td>[.57]</td>
<td>R252</td>
<td>1000 Escudos</td>
<td>Gross Product</td>
</tr>
<tr>
<td>[.58]</td>
<td>R256</td>
<td>1000 Escudos</td>
<td>Gross Agricultural Product</td>
</tr>
<tr>
<td>[.59]</td>
<td>R258</td>
<td>1000 Escudos</td>
<td>Gross Value Added (GVA)</td>
</tr>
<tr>
<td>[.60]</td>
<td>R263</td>
<td>1000 Escudos</td>
<td>Final Results</td>
</tr>
<tr>
<td>[.61]</td>
<td>R270</td>
<td>1000 Escudos</td>
<td>Family labour income</td>
</tr>
<tr>
<td>[.62]</td>
<td>R271</td>
<td>1000 Escudos</td>
<td>Capital Income</td>
</tr>
</tbody>
</table>

**Source**

Obtained directly from the source.

**gcd.coef**

*Computes Yanai’s GCD in the context of the variable-subset selection problem*

**Description**

Computes Yanai’s Generalized Coefficient of Determination for the similarity of the subspaces spanned by a subset of variables and a subset of the full data set’s Principal Components.

**Usage**

```r
gcd.coef(mat, indices, pcindices = NULL)
```

**Arguments**

- `mat` the full data set’s covariance (or correlation) matrix.
- `indices` a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different `k`-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.
- `pcindices` a numerical vector of indices of Principal Components. By default, the first `k` PCs are chosen, where `k` is the cardinality of the subset of variables whose criterion value is being computed. If a vector of PCs is specified by the user, those PCs will be used for all cardinalities that were requested.
Details

Computes Yanai’s Generalized Coefficient of Determination for the similarity of the subspaces spanned by a subset of variables (specified by indices) and a subset of the full-data set’s Principal Components (specified by pcindices). Input data is expected in the form of a (co)variance or correlation matrix. If a non-square matrix is given, it is assumed to be a data matrix, and its correlation matrix is used as input. The number of variables (k) and of PCs (q) does not have to be the same.

Yanai’s GCD is defined as:

\[ GCD = \frac{\text{tr}(P_v \cdot P_c)}{\sqrt{k \cdot q}} \]

where \( P_v \) and \( P_c \) are the matrices of orthogonal projections on the subspaces spanned by the k-variable subset and by the q-Principal Component subset, respectively.

This definition is equivalent to:

\[ GCD = \frac{1}{\sqrt{kq}} \sum_i (r_{mi})^2 \]

where \((r_{mi})_i\) stands for the multiple correlation between the i-th Principal Component and the k-variable subset, and the sum is carried out over the q PCs (i=1,...,q) selected.

These definitions are also equivalent to the expression used in the code, which only requires the covariance (or correlation) matrix of the data under consideration.

The fact that indices can be a matrix or 3-d array allows for the computation of the GCD values of subsets produced by the search functions `anneal`, `genetic` and `improve` (whose output option $subsets$ are matrices or 3-d arrays), using a different criterion (see the example below).

Value

The value of the GCD coefficient.

References


Examples

```r
## An example with a very small data set.

data(iris3)
x<-iris3[,,1]
gcd.coef(cor(x),c(1,3))
## [1] 0.7666286
gcd.coef(cor(x),c(1,3),pcindices=c(1,3))
## [1] 0.584452

gcd.coef(cor(x),c(1,3),pcindices=1)
## [1] 0.6035127
```

```
```
An example computing the GCDs of three subsets produced when the anneal function attempted to optimize the RV criterion (using an absurdly small number of iterations).

```r
data(swiss)
rvresults<-anneal(cor(swiss),2,nsol=4,niter=5,criterion="Rv")
gcd.coef(cor(swiss),rvresults$subsets)
```

```r
# Card.2
# Solution 1 0.4962297
# Solution 2 0.7092591
# Solution 3 0.4748525
# Solution 4 0.4649259
```

---

**genetic**

*Genetic Algorithm searching for an optimal k-variable subset*

---

**Description**

Given a set of variables, a Genetic Algorithm algorithm seeks a k-variable subset which is optimal, as a surrogate for the whole set, with respect to a given criterion.

**Usage**

```r
genetic( mat, kmin, kmax = kmin, popsize = max(100,2*ncol(mat)), nger = 100,
mutate = FALSE, mutprob = 0.01, maxclone = 5, exclude = NULL,
include = NULL, improvement = TRUE, setseed= FALSE, criterion = "default",
pcindices = "first_k", initialpop = NULL, force = FALSE, H=NULL, r=0,
tolval=1000*.Machine$double.eps,tolsym=1000*.Machine$double.eps)
```

**Arguments**

- `mat` a covariance/correlation, information or sums of squares and products matrix of the variables from which the k-subset is to be selected. See the Details section below.
- `kmin` the cardinality of the smallest subset that is wanted.
- `kmax` the cardinality of the largest subset that is wanted.
- `popsize` integer variable indicating the size of the population.
- `nger` integer variable giving the number of generations for which the genetic algorithm will run.
- `mutate` logical variable indicating whether each child undergoes a mutation, with probability mutprob. By default, FALSE.
- `mutprob` variable giving the probability of each child undergoing a mutation, if mutate is TRUE. By default, 0.01. High values slow down the algorithm considerably and tend to replicate the same solution.
maxclone

integer variable specifying the maximum number of identical replicates (clones) of individuals that is acceptable in the population. Serves to ensure that the population has sufficient genetic diversity, which is necessary to enable the algorithm to complete the specified number of generations. However, even maxclone=0 does not guarantee that there are no repetitions: only the offspring of couples are tested for clones. If any such clones are rejected, they are replaced by a k-variable subset chosen at random, without any further clone tests.

exclude

a vector of variables (referenced by their row/column numbers in matrix mat) that are to be forcibly excluded from the subsets.

include

a vector of variables (referenced by their row/column numbers in matrix mat) that are to be forcibly included in the subsets.

improvement

a logical variable indicating whether or not the best final subset (for each cardinality) is to be passed as input to a local improvement algorithm (see function improve).

setseed

logical variable indicating whether to fix an initial seed for the random number generator, which will be re-used in future calls to this function whenever setseed is again set to TRUE.

criterion

Character variable, which indicates which criterion is to be used in judging the quality of the subsets. Currently, the "Rm", "Rv", "Gcd", "Tau2", "Xi2", "Zeta2", "ccr12" and "Wald" criteria are supported (see the Details section, the References and the links rm.coef, rv.coef, gcd.coef, tau2.coef, xi2.coef, zeta2.coef and ccr12.coef for further details). The default criterion is "Rm" if parameter r is zero (exploratory and PCA problems), "Wald" if r is equal to one and mat has a "FisherI" attribute set to TRUE (generalized linear models), and "Tau2" otherwise (multivariate linear model framework).

pcindices

either a vector of ranks of Principal Components that are to be used for comparison with the k-variable subsets (for the Gcd criterion only, see gcd.coef) or the default text first_k. The latter will associate PCs 1 to k with each cardinality k that has been requested by the user.

initialpop

vector, matrix or 3-d array of initial population for the genetic algorithm. If a single cardinality is required, initialpop may be a popsize x k matrix or a popsize x k x 1 array (as produced by the $subsets output value of any of the algorithm functions anneal, genetic, or improve). If more than one cardinality is requested, initialpop must be a popsize x kmax x length(kmin:kmax) 3-d array (as produced by the $subsets output value). If the exclude and/or include options are used, initialpop must also respect those requirements.

force

a logical variable indicating whether, for large data sets (currently p > 400) the algorithm should proceed anyways, regardless of possible memory problems which may crash the R session.

H

Effect description matrix. Not used with the Rm, Rv or Gcd criteria, hence the NULL default value. See the Details section below.

r

Expected rank of the effects (H) matrix. Not used with the Rm, Rv or Gcd criteria. See the Details section below.
tolval

the tolerance level for the reciprocal of the 2-norm condition number of the correlation/covariance matrix, i.e., for the ratio of the smallest to the largest eigenvalue of the input matrix. Matrices with a reciprocal of the condition number smaller than tolval will activate a restricted-search for well conditioned subsets.

tolsym

the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes (A+t(A))/2.

Details

For each cardinality k (with k ranging from kmin to kmax), an initial population of popsize k-variable subsets is randomly selected from a full set of p variables. In each iteration, popsize/2 couples are formed from among the population and each couple generates a child (a new k-variable subset) which inherits properties of its parents (specifically, it inherits all variables common to both parents and a random selection of variables in the symmetric difference of its parents’ genetic makeup). Each offspring may optionally undergo a mutation (in the form of a local improvement algorithm – see function improve), with a user-specified probability. The parents and offspring are ranked according to their criterion value, and the best popsize of these k-subsets will make up the next generation, which is used as the current population in the subsequent iteration.

The stopping rule for the algorithm is the number of generations (nger).

Optionally, the best k-variable subset produced by the Genetic Algorithm may be passed as input to a restricted local improvement algorithm, for possible further improvement (see function improve).

The user may force variables to be included and/or excluded from the k-subsets, and may specify an initial population.

For each cardinality k, the total number of calls to the procedure which computes the criterion values is popsize + nger x popsize/2. These calls are the dominant computational effort in each iteration of the algorithm.

In order to improve computation times, the bulk of computations are carried out by a Fortran routine. Further details about the Genetic Algorithm can be found in Reference 1 and in the comments to the Fortran code (in the src subdirectory for this package). For datasets with a very large number of variables (currently p > 400), it is necessary to set the force argument to TRUE for the function to run, but this may cause a session crash if there is not enough memory available.

The function checks for ill-conditioning of the input matrix (specifically, it checks whether the ratio of the input matrix’s smallest and largest eigenvalues is less than tolval). For an ill-conditioned input matrix, the search is restricted to its well-conditioned subsets. The function trim.matrix may be used to obtain a well-conditioned input matrix.

In a general descriptive (Principal Components Analysis) setting, the three criteria Rm, Rv and Gcd can be used to select good k-variable subsets. Arguments H and r are not used in this context. See references [1] and [2] and the Examples for a more detailed discussion.

In the setting of a multivariate linear model, $X = A\Psi + U$, criteria Ccr12, Tau2, Xi2 and Zeta2 can be used to select subsets according to their contribution to an effect characterized by the violation of a reference hypothesis, $C\Psi = 0$ (see reference [3] for further details). In this setting, arguments
mat and H should be set respectively to the usual Total (Hypothesis + Error) and Hypothesis, Sum of Squares and Cross-Products (SSCP) matrices. Argument r should be set to the expected rank of H. Currently, for reasons of computational efficiency, criterion Ccr12 is available only when r ≤ 3. Particular cases in this setting include Linear Discriminant Analysis (LDA), Linear Regression Analysis (LRA), Canonical Correlation Analysis (CCA) with one set of variables fixed and several extensions of these and other classical multivariate methodologies.

In the setting of a generalized linear model, criterion Wald can be used to select subsets according to the (lack of) significance of the discarded variables, as measured by the respective Wald’s statistic (see reference [4] for further details). In this setting arguments mat and H should be set respectively to FI and FI %*% b %*% t(b) %*% FI, where b is a column vector of variable coefficient estimates and FI is an estimate of the corresponding Fisher information matrix.

The auxiliary functions lmHmat, ldaHmat glhHmat and glmHmat are provided to automatically create the matrices mat and H in all the cases considered.

Value
A list with five items:

subsets A popsize x kmax x length(kmin:kmax) 3-dimensional array, giving for each cardinality (dimension 3) and each subset in the final population (dimension 1) the list of variables (referenced by their row/column numbers in matrix mat) in the subset (dimension 2). (For cardinalities smaller than kmax, the extra final positions are set to zero).

values A popsize x length(kmin:kmax) matrix, giving for each cardinality (columns), the (ordered) criterion values of the popsize (rows) subsets in the final generation.

bestvalues A length(kmin:kmax) vector giving the best values of the criterion obtained for each cardinality. If improvement is TRUE, these values result from the final restricted local search algorithm (and may therefore exceed the largest value for that cardinality in values).

bestsets A length(kmin:kmax) x kmax matrix, giving, for each cardinality (rows), the variables (referenced by their row/column numbers in matrix mat) in the best k-subset that was found.

call The function call which generated the output.

References


See Also

rm.coef, rv.coef, gcd.coef, tau2.coef, xi2.coef, zeta2.coef, ccr12.coef, genetic, anneal, eleaps, trim.matrix, lmHmat, ldaHmat, glhHmat, glmHmat.

Examples

```r
## 1) For illustration of use, a small data set with very few iterations
## of the algorithm. Escoufier's 'RV' criterion is used to select variable
## subsets of size 3 and 4.
##
data(swiss)
genetic(cor(swiss), 3, 4, popsize = 10, nger = 5, criterion = "Rv")

## For cardinality k =
##[[1]] 4
## there is not enough genetic diversity in generation number
##[[1]] 3
## for acceptable levels of consanguinity (couples differing by at least 2 genes).
## Try reducing the maximum acceptable number of clones (maxclone) or
## increasing the population size (popsize)
## Best criterion value found so far:
##[[1]] 0.9557145
#$subsets
## , , Card.3
##
##    Var.1 Var.2 Var.3 Var.4
##Solution 1 1 2 3 0
##Solution 2 1 2 3 0
##Solution 3 1 2 3 0
##Solution 4 3 4 6 0
##Solution 5 3 4 6 0
##Solution 6 3 4 5 0
##Solution 7 3 4 5 0
##Solution 8 1 3 6 0
##Solution 9 1 3 6 0
##Solution 10 1 3 6 0
##
## , , Card.4
##
##    Var.1 Var.2 Var.3 Var.4
##Solution 1 2 4 5 6
##Solution 2 1 2 5 6
##Solution 3 1 2 3 5
##Solution 4 1 2 4 5
##Solution 5 1 2 4 5
##Solution 6 1 4 5 6
##Solution 7 1 4 5 6
```
## Solution 8 1 4 5 6
## Solution 9 1 3 4 5
## Solution 10 1 3 4 5
#
## $values
## card.3 card.4
## Solution 1 0.9141995 0.9557145
## Solution 2 0.9141995 0.9485699
## Solution 3 0.9141995 0.9455508
## Solution 4 0.9034868 0.9433203
## Solution 5 0.9034868 0.9433203
## Solution 6 0.9020271 0.9428967
## Solution 7 0.9020271 0.9428967
## Solution 8 0.8988192 0.9428967
## Solution 9 0.8988192 0.9357982
## Solution 10 0.8988192 0.9357982
##
## $bestvalues
## Card.3 Card.4
## 0.9141995 0.9557145
##
## $bestsets
## Var.1 Var.2 Var.3 Var.4
## Card.3 1 2 3 0
## Card.4 2 4 5 6
##
## $call
## genetic(mat = cor(swiss), kmin = 3, kmax = 4, popsize = 10, nger = 5, criterion = "Rv")

---

##
## 2) An example of subset selection in the context of Multiple Linear Regression. Variable 5 (average car price) in the Cars93 MASS library data set is regressed on 13 other variables. The six-variable subsets of linear predictors are chosen using the "CCR1_2" criterion which, in the case of a Linear Regression, is merely the standard Coefficient of Determination, R^2 (as are the other three criteria for the multivariate linear hypothesis, "XI_2", "TAU_2" and "ZETA_2").
##
library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,5])

names(Cars93[,5,drop=FALSE])
## [1] "Price"

colnames(CarsHmat)
## 

### [1] "MPG.city" "MPG.highway" "EngineSize"
### [4] "Horsepower" "RPM" "Rev.per.mile"
### [7] "Fuel.tank.capacity" "Passengers" "Length"
### [10] "Wheelbase" "Width" "Turn.circle"
### [13] "Weight"

```r
genetic(CarsHmat$mat, kmin=6, H=CarsHmat$H, r=1, crit="CCR12")
```

---

#### (Partial results only)

#### $subsets

<table>
<thead>
<tr>
<th>Solution</th>
<th>Var.1</th>
<th>Var.2</th>
<th>Var.3</th>
<th>Var.4</th>
<th>Var.5</th>
<th>Var.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

#### (...)

<table>
<thead>
<tr>
<th>Solution</th>
<th>Var.1</th>
<th>Var.2</th>
<th>Var.3</th>
<th>Var.4</th>
<th>Var.5</th>
<th>Var.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

#### $values

<table>
<thead>
<tr>
<th>Solution</th>
<th>Solution 2</th>
<th>Solution 3</th>
<th>Solution 4</th>
<th>Solution 5</th>
<th>Solution 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7310150</td>
<td>0.7310150</td>
<td>0.7310150</td>
<td>0.7310150</td>
<td>0.7310150</td>
</tr>
<tr>
<td></td>
<td>0.7310150</td>
<td>0.7271056</td>
<td>0.7271056</td>
<td>0.7271056</td>
<td>0.7271056</td>
</tr>
<tr>
<td></td>
<td>0.7310150</td>
<td>0.7271056</td>
<td>0.7271056</td>
<td>0.7271056</td>
<td>0.7271056</td>
</tr>
<tr>
<td></td>
<td>0.7271056</td>
<td>0.7270257</td>
<td>0.7270257</td>
<td>0.7270257</td>
<td>0.7270257</td>
</tr>
</tbody>
</table>

#### (...)

<table>
<thead>
<tr>
<th>Solution</th>
<th>Solution 2</th>
<th>Solution 3</th>
<th>Solution 4</th>
<th>Solution 5</th>
<th>Solution 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7228800</td>
<td>0.7228800</td>
<td>0.7228800</td>
<td>0.7228800</td>
<td>0.7228800</td>
</tr>
<tr>
<td></td>
<td>0.7228463</td>
<td>0.7228463</td>
<td>0.7228463</td>
<td>0.7228463</td>
<td>0.7228463</td>
</tr>
<tr>
<td></td>
<td>0.7228463</td>
<td>0.7228463</td>
<td>0.7228463</td>
<td>0.7228463</td>
<td>0.7228463</td>
</tr>
</tbody>
</table>

#### $bestvalues

| Card.6  | 0.731015 |
## $bestsets
## Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
## 4   5   9  10  11  12
##
## $call
## genetic(mat = CarsHmat$mat, kmin = 6, criterion = "CCR12"; H = CarsHmat$H,
##   r = 1)

### 3) An example of subset selection in the context of a Canonical
### Correlation Analysis. Two groups of variables within the Cars93
### MASS library data set are compared. The goal is to select 4- to
### 6-variable subsets of the 13-variable 'X' group that are optimal in
### terms of preserving the canonical correlations, according to the
### "ZETA_2" criterion (Warning: the 3-variable 'Y' group is kept
### intact; subset selection is carried out in the 'X'
### group only). The 'tolsym' parameter is used to relax the symmetry
### requirements on the effect matrix H which, for numerical reasons,
### is slightly asymmetric. Since corresponding off-diagonal entries of
### matrix H are different, but by less than tolsym, H is replaced
### by its symmetric part: (H+t(H))/2.

library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,4:6])
names(Cars93[,4:6])
colnames(CarsHmat$mat)
## [1] "MPG.city"  "MPG.highway"  "EngineSize"
## [4] "Horsepower" "RPM"        "Rev.per.mile"
## [7] "Fuel.tank.capacity" "Passengers" "Length"
## [10] "Wheelbase"  "Width"     "Turn.circle"
## [13] "Weight"

genetic(CarsHmat$mat, kmin=5, kmax=6, H=CarsHmat$H, r=3, crit="zeta2", tolsym=1e-9)

### (PARTIAL RESULTS ONLY)
###
### $subsets
###
### Solution 1   4   5   9  10  11  0
### Solution 2   4   5   9  10  11  0
### Solution 3   4   5   9  10  11  0
### Solution 4   4   5   9  10  11  0
### Solution 5   4   5   9  10  11  0
## Solution 6 4 5 9 10 11 0
## Solution 7 4 5 9 10 11 0
## Solution 8 3 4 9 10 11 0
## Solution 9 3 4 9 10 11 0
## Solution 10 3 4 9 10 11 0
## (...
## Solution 87 3 4 6 9 10 11
## Solution 88 3 4 6 9 10 11
## Solution 89 3 4 6 9 10 11
## Solution 90 2 3 4 10 11 12
## Solution 91 2 3 4 10 11 12
## Solution 92 2 3 4 10 11 12
## Solution 93 2 3 4 10 11 12
## Solution 94 2 3 4 10 11 12
## Solution 95 2 3 4 10 11 12
## Solution 96 2 3 4 10 11 12
## Solution 97 1 3 4 6 10 11
## Solution 98 1 3 4 6 10 11
## Solution 99 1 3 4 6 10 11
## Solution 100 1 3 4 6 10 11
##
## $values
##
## card.5   card.6
##
## Solution 1 0.5018922 0.5168627
## Solution 2 0.5018922 0.5168627
## Solution 3 0.5018922 0.5168627
## Solution 4 0.5018922 0.5168627
## Solution 5 0.5018922 0.5168627
## Solution 6 0.5018922 0.5168627
## Solution 7 0.5018922 0.5096500
## Solution 8 0.4966191 0.5096500
## Solution 9 0.4966191 0.5096500
## Solution 10 0.4966191 0.5096500
## (...
## Solution 87 0.4893824 0.5038649
## Solution 88 0.4893824 0.5038649
## Solution 89 0.4893824 0.5038649
## Solution 90 0.4893824 0.5035489
## Solution 91 0.4893824 0.5035489
## Solution 92 0.4893824 0.5035489
## Solution 93 0.4893824 0.5035489
## Solution 94 0.4893824 0.5035489
## Solution 95 0.4893824 0.5035489
## Solution 96 0.4893824 0.5035489
## Solution 97 0.4890986 0.5035386
## Solution 98 0.4890986 0.5035386
## Solution 99 0.4890986 0.5035386
glhHmat

Total and Effect Deviation Matrices for General Linear Hypothesis

Description

Computes total and effect matrices of Sums of Squares and Cross-Product (SSCP) deviations for a general multivariate effect characterized by the violation of a linear hypothesis. These matrices may be used as input to the variable selection search routines `anneal`, `genetic_improve` or `eleaps`.

Usage
## Default S3 method:
glhHmat(x,A,C,...)

## S3 method for class 'data.frame'
glhHmat(x,A,C,...)

## S3 method for class 'formula'
glhHmat(formula,C,data=NULL,...)

### Arguments

- **x**: A matrix or data frame containing the variables for which the SSCP matrix is to be computed.
- **A**: A matrix or data frame containing a design matrix specifying a linear model in which x is the response.
- **C**: A matrix or vector containing the coefficients of the reference hypothesis.
- **formula**: A formula of the form `x ~ A1 + A2 + ...` That is, the response is the set of variables whose subsets are to be compared and the right hand side specifies the columns of the design matrix.
- **data**: Data frame from which variables specified in 'formula' are preferentially to be taken.
- **...**: further arguments for the method.

### Details

Consider a multivariate linear model $x = A\Psi + U$ and a reference hypothesis $H_0: C\Psi = 0$, with $\Psi$ being a matrix of unknown parameters and $C$ a known coefficient matrix with rank $r$. It is well known that, under classical Gaussian assumptions, $H_0$ can be tested by several increasing functions of the $r$ positive eigenvalues of a product $T^{-1}H$, where $T$ and $H$ are total and effect matrices of SSCP deviations associated with $H_0$. Furthermore, whether or not the classical assumptions hold, the same eigenvalues can be used to define descriptive indices that measure an "effect" characterized by the violation of $H_0$ (see reference [1] for further details). Those SSCP matrices are given by $T = x'(I - P_\omega)x$ and $H = x'(P_\Omega - P_\omega)x$, where $I$ is an identity matrix and $P_\Omega = A(A'A)^{-1}A'$,

$$
P_\omega = A(A'A)^{-1}A' - A(A'A)^{-1}C'(C(A'A)^{-1}C')^{-1}C(A'A)^{-1}A'
$$

are projection matrices on the spaces spanned by the columns of $A$ (space $\Omega$) and by the linear combinations of these columns that satisfy the reference hypothesis (space $\omega$). In these formulae $M'$ denotes the transpose of $M$ and $M^-$ a generalized inverse. `glhHmat` computes the $T$ and $H$ matrices which then can be used as input to the search routines `anneal`, `genetic improve` and `eleaps` that try to select subsets of $x$ according to their contribution to the violation of $H_0$.

### Value

A list with four items:

- **mat**: The total SSCP matrix
The effect SSCP matrix

The expected rank of the H matrix which equals the rank of C. The true rank of H can be different from r if the x variables are linearly dependent.

The function call which generated the output.

References


See Also

`anneal`, `genetic`, `improve`, `eleaps`, `lmHmat`, `ldaHmat`.

Examples

```r
##----------------------------------------------------------------------------
## The following examples create T and H matrices for different analysis
## of the MASS data set "crabs". This data records physical measurements
## on 200 specimens of Leptograpsus variegatus crabs observed on the shores
## of Western Australia. The crabs are classified by two factors, sex and sp
## (crab species as defined by its colour: blue or orange), with two levels
## each. The measurement variables include the carapace length (CL),
## the carapace width (CW), the size of the frontal lobe (FL) and the size of
## the rear width (RW). In the analysis provided, we assume that there is
## an interest in comparing the subsets of these variables measured in their
## original and logarithmic scales.

library(MASS)
data(crabs)
lFL <- log(crabs$FL)
lRW <- log(crabs$RW)
lCL <- log(crabs$CL)
lCW <- log(crabs$CW)

# 1) Create the T and H matrices associated with a linear
# discriminant analysis on the groups defined by the sp factor.
# This call is equivalent to ldaHmat(sp ~ FL + RW + CL + CW + lFL +
# lRW + lCL + lCW, crabs)
Hmat1 <- glhHmat(cbind(FL,RW,CL,CW,lFL,lRW,lCL,lCW) ~ sp,c(0,1),crabs)
Hmat1
```

```
##$mat
##       FL      RW      CL      CW    lFL     lRW     lCL
##FL 2431.2422 1623.4509 4846.9787 5283.6093 162.7186 133.3604 158.8651
##RW 1623.4509 1317.7935 3254.5776 3629.6883 109.8772 107.2872 108.3357
##CL 4846.9787 3254.5776 10085.3040 11096.5141 326.2433 269.5647 330.9126
##CW 5283.6093 3629.6883 11096.5141 12331.5680 356.3179 300.7868 364.6208
```
glhHmat

# ICL 158.8651 108.3357 330.9126 364.6208 10.910730 9.130692 11.088706
# ICW 152.7872 106.4277 321.0253 357.0051 10.503303 8.970570 10.765175
## FL 152.78716
## RW 106.42775
## CL 321.02534
## CW 357.00510
## lFL 10.50330
## lRW 8.97057
## lCL 10.76517
## lCW 10.54334

### FL RW CL CW lFL lRW lCL
### FL 466.34580 247.52670 625.30650 518.41650 30.7408809 19.4543206 20.5494907
### RW 247.52670 131.38205 331.89975 275.16475 16.3166234 10.3259508 10.9072444
### CL 625.30650 331.89975 838.45125 695.12625 41.2193540 26.0856066 27.5540813
### CW 518.41650 275.16475 695.12625 576.30125 34.1733106 21.6265286 22.8439819
### lFL 30.74088 16.31662 41.21935 34.17331 2.0263971 1.2824024 1.3545945
### lRW 19.45432 10.32595 26.08561 21.62653 1.2824024 0.8115664 0.8572531
### lCL 20.54949 10.90724 27.55408 22.84398 1.3545945 0.8572531 0.9055117
### lCW 15.16136 8.049106 20.32932 16.85420 0.9994161 0.6324790 0.6680840

# 2) Create the T and H matrices associated with an analysis
# of the interactions between the sp and sex factors
Hmat2 <- glhHmat(cbind(FL, RW, CL, CW, lFL, lRW, lCL) ~ sp, C = c(0, 1), data = crabs)
Hmat2

# smat
### FL RW CL CW lFL lRW lCL
### FL 1960.3362 1398.52890 4199.1581 4747.5409 131.651804 115.607172 137.663744
### RW 1398.5289 1074.36105 3034.2793 3442.0233 95.176151 88.529040 100.659912
### CL 4199.1581 3034.27925 9135.6987 10314.2389 283.414814 251.877591 300.140085
### CW 4747.5409 3442.02325 10314.2389 11686.9387 320.883015 285.744945 339.253367
### lFL 131.6518 95.17615 283.4148 320.8830 9.065041 8.027569 9.509543
## lRW
## lCL
## lCW
## FL
## RW
## CL
## CW

### H

<table>
<thead>
<tr>
<th>FL</th>
<th>RW</th>
<th>CL</th>
<th>CW</th>
<th>lFL</th>
<th>lRW</th>
<th>lCL</th>
<th>lCW</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.645000</td>
<td>68.389500</td>
<td>153.733500</td>
<td>191.579500</td>
<td>5.4708199</td>
<td>5.1596883</td>
<td>5.2140868</td>
<td></td>
</tr>
<tr>
<td>68.389500</td>
<td>57.996450</td>
<td>130.370850</td>
<td>162.465450</td>
<td>4.6394276</td>
<td>4.3755782</td>
<td>4.4217098</td>
<td></td>
</tr>
<tr>
<td>153.733500</td>
<td>130.370850</td>
<td>293.062050</td>
<td>365.207850</td>
<td>10.4290197</td>
<td>9.8359098</td>
<td>9.9396095</td>
<td></td>
</tr>
<tr>
<td>191.579500</td>
<td>162.465450</td>
<td>365.207850</td>
<td>455.11445</td>
<td>12.9964281</td>
<td>12.2573068</td>
<td>12.3865353</td>
<td></td>
</tr>
</tbody>
</table>

### T

<table>
<thead>
<tr>
<th>FL</th>
<th>RW</th>
<th>CL</th>
<th>CW</th>
<th>lFL</th>
<th>lRW</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.470820</td>
<td>4.639428</td>
<td>9.835911</td>
<td>12.99643</td>
<td>0.3711311</td>
<td>0.3500245</td>
</tr>
<tr>
<td>4.639428</td>
<td>4.375578</td>
<td>9.939610</td>
<td>12.25731</td>
<td>0.3500245</td>
<td>0.3301182</td>
</tr>
<tr>
<td>9.835911</td>
<td>9.939610</td>
<td>9.939610</td>
<td>12.25731</td>
<td>0.3500245</td>
<td>0.3301182</td>
</tr>
<tr>
<td>12.99643</td>
<td>12.25731</td>
<td>12.99643</td>
<td>12.99643</td>
<td>0.3500245</td>
<td>0.3301182</td>
</tr>
</tbody>
</table>

### r

### call

```r
# glHmat.formula(formula = cbind(FL, RW, CL, CW, lFL, lRW, lCL, lCW) ~ sp * sex, C = c(0, 0, 0, 1), data = crabs)
```

### 3) Create the T and H matrices associated with an analysis

### of the effect of the sp factor after controlling for sex

C <- matrix(0,2,4)

```r
# C

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

Hmat3 <- glHmat(cbind(FL, RW, CL, CW, lFL, lRW, lCL, lCW) ~ sp*sex,C,crabs)
Hmat3
```r
glmHmat

```
**glmHmat**

*Input matrices for subselect search routines in generalized linear models*

**Description**

`glmHmat` uses a glm object (fitdglmmodel) to build an estimate of Fisher’s Information (FI) matrix together with an auxiliary rank-one positive-definite matrix (H), such that the positive eigenvalue of $FI^{-1}H$ equals the value of Wald’s statistic for testing the global significance of fitdglmmodel. These matrices may be used as input to the variable selection search routines `anneal`, `genetic`, `improve` or `eleaps`, using the minimization of Wald’s statistic as criterion for discarding variables.

**Usage**

```r
## S3 method for class 'glm'
glmHmat(fitdglmmodel, ...)
```

**Arguments**

- `fitdglmmodel`: A glm object containing the estimates, and respective covariance matrix, of a generalized linear model.
- `...`: Further arguments for the method.

**Details**

Variable selection in the context of generalized linear models is typically based on the minimization of statistics that test the significance of excluded variables. In particular, the likelihood ratio, Wald’s, Rao’s and some adaptations of such statistics, are often proposed as comparison criteria for variable subsets of the same dimensionality. All these statistics are asymptotically equivalent and can be converted into information criteria, like the AIC, that are also able to compare subsets of different dimensionalities (see references [1] and [2] for further details).

Among these criteria, Wald’s statistic has some computational advantages because it can always be derived from the same (concerning the full model) maximum likelihood and Fisher information estimates. In particular, if $W_{allv}$ is the value of the Wald statistic testing the significance of the full covariate vector, b and FI are coefficient and Fisher information estimates and H is an auxiliary rank-one matrix given by $H = FI \times b' \times b \times FI$, it follows that the value of Wald’s statistic for the excluded variables ($W_{exc}$) in a given subset is given by $W_{exc} = W_{allv} - tr(\sum_{indext} H_{indices})$, where $FI_{indices}$ and $H_{indices}$ are the portions of the FI and H matrices associated with the selected variables.

`glmHmat` retrieves the values of the FI and H matrices from a glm object. These matrices may then be used as input to the search functions `anneal`, `genetic`, `improve` and `eleaps`.

**Value**

A list with four items:
glmHmat

mat        An estimate (FI) of Fisher's information matrix for the full model variable-
           coefficient estimates
H          A product of the form (FI %*% b %*% t(b) %*% FI) where b is a vector of variable-
           coefficient estimates
r          The rank of the H matrix. Always set to one in glmHmat.
call       The function call which generated the output.

References

   Biometrics, Vol. 34, 318-327.

   alized Models I. Statistical and Computational Background, Computer Methods and Programs in 

See Also

anneal, genetic, improve, eleaps, glm.

Examples

###---------------------------------------------------------------
###---------------------------------------------------------------

### An example of variable selection in the context of binary response 
### regression models. We consider the last 100 observations of 
### the iris data set (versicolor an veriginica species) and try 
### to find the best variable subsets for models that take species 
### as the response variable.

data(iris)
iris2sp <- iris[iris$Species != "setosa",]

# Create the input matrices for the search routines in a logistic regression model
modelfit <- glm(Species ~ Sepal.Length + Sepal.Width + Petal.Length + 
Petal.Width,iris2sp,family=binomial)
Hmat <- glmHmat(modelfit)
Hmat

## $mat
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length  0.28340358  0.03263437  0.09552821 -0.01779067
## Sepal.Width   0.03263437  0.13941541  0.01086596  0.04759284
## Petal.Length  0.09552821  0.01086596  0.08847655 -0.01853044
## Petal.Width   -0.01779067  0.04759284 -0.01853044  0.03258730

## $H
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## `sep` Length 0.11643732 0.013349227 -0.063924853 -0.050181400
## `sep` Width 0.01334923 0.001530453 -0.007328813 -0.005753163
## `pet` Length -0.06392485 -0.007328813 0.035095164 0.027549918
## `pet` Width -0.05018140 -0.005753163 0.027549918 0.021626854

### $r$
### [1] 1

### $call$
### `glmHmat(fitdglmmmodel = modelfit)`

# Search for the 3 best variable subsets of each dimensionality by an exhaustive search

`eleaps(Hmat$mat,H=Hmat$H,r=1,criterion="Wald",nsol=3)`

### $subsets$
### , , Card.1
###     Var.1 Var.2 Var.3
### Solution 1  4 0 0
### Solution 2  1 0 0
### Solution 3  3 0 0

### , , Card.2
###     Var.1 Var.2 Var.3
### Solution 1  1 3 0
### Solution 2  3 4 0
### Solution 3  2 4 0

### , , Card.3
###     Var.1 Var.2 Var.3
### Solution 1  2 3 4
### Solution 2  1 3 4
### Solution 3  1 2 3

### $values$
### card.1 card.2 card.3
### Solution 1 4.894554 3.522885 1.060121
### Solution 2 5.147360 3.952538 2.224335
### Solution 3 5.161553 3.972410 3.522879

### $bestvalues$
### Card.1 Card.2 Card.3
### 4.894554 3.522885 1.060121

### $bestsets$
### Var.1 Var.2 Var.3
### Card.1 4 0 0
### Card.2 1 3 0
### Card.3 2 3 4
improve

Restricted Local Improvement search for an optimal k-variable subset

Description

Given a set of variables, a Restricted Local Improvement algorithm seeks a k-variable subset which is optimal, as a surrogate for the whole set, with respect to a given criterion.

Usage

improve(mat, kmin, kmax = kmin, nsol = 1, exclude = NULL, include = NULL, setseed = FALSE, criterion = "default", pcindices="first_k", initialsol = NULL, force = FALSE, H=NULL, r=0, tolval=1000*.Machine$double.eps,tolsym=1000*.Machine$double.eps)

Arguments

mat a covariance/correlation, information or sums of squares and products matrix of the variables from which the k-subset is to be selected. See the Details section below.

kmin the cardinality of the smallest subset that is wanted.

kmax the cardinality of the largest subset that is wanted.

nsol the number of different subsets (runs of the algorithm) wanted.

exclude a vector of variables (referenced by their row/column numbers in matrix mat) that are to be forcibly excluded from the subsets.

include a vector of variables (referenced by their row/column numbers in matrix mat) that are to be forcibly included from the subsets.

setseed logical variable indicating whether to fix an initial seed for the random number generator, which will be re-used in future calls to this function whenever setseed is again set to TRUE.
Character variable, which indicates which criterion is to be used in judging the quality of the subsets. Currently, the "Rm", "Rv", "Gcd", "Tau2", "Xi2", "Zeta2", "ccr12" and "Wald" criteria are supported (see the Details section, the References and the links rm.coef, rv.coef, gcd.coef, tau2.coef, xi2.coef, zeta2.coef and ccr12.coef for further details). The default criterion is "Rm" if parameter r is zero (exploratory and PCA problems), "Wald" if r is equal to one and mat has a "FisherI" attribute set to TRUE (generalized linear models), and "Tau2" otherwise (multivariate linear model framework).

either a vector of ranks of Principal Components that are to be used for comparison with the k-variable subsets (for the Gcd criterion only, see gcd.coef) or the default text first_k. The latter will associate PCs 1 to k that has been requested by the user.

vector, matrix or 3-d array of initial solutions for the restricted local improvement search. If a single cardinality is required, initialsol may be a vector of length k (accepted even if nsol > 1, in which case it is used as the initial solution for all nsol final solutions that are requested with a warning that the same initial solution necessarily produces the same final solution); a 1 x k matrix (as produced by the $bestsets output value of the algorithm functions anneal, genetic, or improve), or a 1 x k x 1 array (as produced by the $subsets output value), in which case it will be treated as the above k-vector; or an nsol x k matrix, or nsol x k x 1 3-d array, in which case each row (dimension 1) will be used as the initial solution for each of the nsol final solutions requested. If more than one cardinality is requested, initialsol can be a length(kmin:kmax) x kmax matrix (as produced by the $bestsets option of the algorithm functions) (even if nsol > 1, in which case each row will be replicated to produced the initial solution for all nsol final solutions requested in each cardinality, with a warning that a single initial solution necessarily produces identical final solutions), or a nsol x kmax x length(kmin:kmax) 3-d array (as produced by the $subsets output option), in which case each row (dimension 1) is interpreted as a different initial solution.

If the exclude and/or include options are used, initialsol must also respect those requirements.

a logical variable indicating whether, for large data sets (currently p > 400) the algorithm should proceed anyways, regardless of possible memory problems which may crash the R session.

Effect description matrix. Not used with the Rm, Rv or Gcd criteria, hence the NULL default value. See the Details section below.

Expected rank of the effects (H) matrix. Not used with the Rm, Rv or Gcd criteria. See the Details section below.

the tolerance level for the reciprocal of the 2-norm condition number of the correlation/covariance matrix, i.e., for the ratio of the smallest to the largest eigenvalue of the input matrix. Matrices with a reciprocal of the condition number smaller than tolval will activate a restricted-search for well conditioned subsets.

the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than
this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix $A$ becomes $(A + t(A))/2$.

Details

An initial $k$-variable subset (for $k$ ranging from $k_{\min}$ to $k_{\max}$) of a full set of $p$ variables is randomly selected and the variables not belonging to this subset are placed in a queue. The possibility of replacing a variable in the current $k$-subset with a variable from the queue is then explored. More precisely, a variable is selected, removed from the queue, and the $k$ values of the criterion which would result from swapping this selected variable with each variable in the current subset are computed. If the best of these values improves the current criterion value, the current subset is updated accordingly. In this case, the variable which leaves the subset is added to the queue, but only if it has not previously been in the queue (i.e., no variable can enter the queue twice). The algorithm proceeds until the queue is emptied.

The user may force variables to be included and/or excluded from the $k$-subsets, and may specify initial solutions.

For each cardinality $k$, the total number of calls to the procedure which computes the criterion values is $O(n_{sol} \times k \times p)$. These calls are the dominant computational effort in each iteration of the algorithm.

In order to improve computation times, the bulk of computations are carried out in a Fortran routine. Further details about the algorithm can be found in Reference 1 and in the comments to the Fortran code (in the src subdirectory for this package). For datasets with a very large number of variables (currently $p > 400$), it is necessary to set the `force` argument to TRUE for the function to run, but this may cause a session crash if there is not enough memory available.

The function checks for ill-conditioning of the input matrix (specifically, it checks whether the ratio of the input matrix’s smallest and largest eigenvalues is less than `tolval`). For an ill-conditioned input matrix, the search is restricted to its well-conditioned subsets. The function `trim.matrix` may be used to obtain a well-conditioned input matrix.

In a general descriptive (Principal Components Analysis) setting, the three criteria $R_m$, $R_v$ and $G_{cd}$ can be used to select good $k$-variable subsets. Arguments $H$ and $r$ are not used in this context. See references [1] and [2] and the Examples for a more detailed discussion.

In the setting of a multivariate linear model, $X = A\Psi + U$, criteria $C_{cr12}$, $Tau2$, $Xi2$ and $Zeta2$ can be used to select subsets according to their contribution to an effect characterized by the violation of a reference hypothesis, $C\Psi = 0$ (see reference [3] for further details). In this setting, arguments `mat` and `$H$ should be set respectively to the usual Total (Hypothesis + Error) and Hypothesis, Sum of Squares and Cross-Products (SSCP) matrices. Argument $r$ should be set to the expected rank of $H$. Currently, for reasons of computational efficiency, criterion $C_{cr12}$ is available only when $r \leq 3$. Particular cases in this setting include Linear Discriminant Analysis (LDA), Linear Regression Analysis (LRA), Canonical Correlation Analysis (CCA) with one set of variables fixed and several extensions of these and other classical multivariate methodologies.

In the setting of a generalized linear model, criterion Wald can be used to select subsets according to the (lack of) significance of the discarded variables, as measured by the respective Wald’s statistic (see reference [4] for further details). In this setting arguments `mat` and `$H$ should be set respectively to $FI$ and $FI \%*% b \%*% t(b) \%*% FI$, where $b$ is a column vector of variable coefficient estimates and $FI$ is an estimate of the corresponding Fisher information matrix.
The auxiliary functions `lmHmat`, `ldaHmat` `glhHmat` and `glmHmat` are provided to automatically create the matrices `mat` and `H` in all the cases considered.

**Value**

A list with five items:

- `subsets` An `nsol` x `kmax` x `length(kmin:kmax)` 3-dimensional array, giving for each cardinality (dimension 3) and each solution (dimension 1) the list of variables (referenced by their row/column numbers in matrix `mat`) in the subset (dimension 2). (For cardinalities smaller than `kmax`, the extra final positions are set to zero).
- `values` An `nsol` x `length(kmin:kmax)` matrix, giving for each cardinality (columns), the criterion values of the `nsol` (rows) solutions obtained.
- `bestvalues` A `length(kmin:kmax)` vector giving the best values of the criterion obtained for each cardinality.
- `bestsets` A `length(kmin:kmax)` x `kmax` matrix, giving, for each cardinality (rows), the variables (referenced by their row/column numbers in matrix `mat`) in the best k-subset that was found.
- `call` The function call which generated the output.

**References**


**See Also**

`rm.coef`, `rv.coef`, `gcd.coef`, `tau2.coef`, `xi2.coef`, `zeta2.coef`, `ccr12.coef`, `genetic`, `anneal`, `eleaps`, `trim.matrix`, `lmHmat`, `ldaHmat`, `glhHmat`, `glmHmat`.

**Examples**

```r
## --------------------------------------------------------------------
##
## 1) For illustration of use, a small data set with very few iterations
## of the algorithm.
## Subsets of 2 and of 3 variables are sought using the RM criterion.
##
## #
```
improve
data(swiss)
improve(cor(swiss),2,3,nsol=4,criterion="GCD")
## $subsets
## , , Card.2
## ## Var.1 Var.2 Var.3
## Solution 1 3 6 0
## Solution 2 3 6 0
## Solution 3 3 6 0
## Solution 4 3 6 0
## , , Card.3
## ## Var.1 Var.2 Var.3
## Solution 1 4 5 6
## Solution 2 4 5 6
## Solution 3 4 5 6
## Solution 4 4 5 6
##
## $values
## card.2 card.3
## Solution 1 0.8487026 0.925372
## Solution 2 0.8487026 0.925372
## Solution 3 0.8487026 0.925372
## Solution 4 0.8487026 0.925372
##
## $bestvalues
## Card.2 Card.3
## 0.8487026 0.9253720
##
## $bestsets
## Var.1 Var.2 Var.3
## Card.2 3 6 0
## Card.3 4 5 6
##
## $call
## improve(cor(swiss), 2, 3, nsol = 4, criterion = "GCD")

# 2) Forcing the inclusion of variable 1 in the subset
#
improve(cor(swiss),2,3,nsol=4,criterion="GCD",include=c(1))

## $subsets
## , , Card.2
## ## Var.1 Var.2 Var.3
## Solution 1 1 6 0
## Solution 2 1 6 0
## Solution 3 1 6 0
## Solution 4 1 6 0
#
## , , Card.3
#
## Var.1 Var.2 Var.3
## Solution 1 1 5 6
## Solution 2 1 5 6
## Solution 3 1 5 6
## Solution 4 1 5 6
#
## $values
card.2 card.3
## Solution 1 0.7284477 0.8048528
## Solution 2 0.7284477 0.8048528
## Solution 3 0.7284477 0.8048528
## Solution 4 0.7284477 0.8048528
#
## $bestvalues
card.2 card.3
## 0.7284477 0.8048528
#
## $bestsets
## Var.1 Var.2 Var.3
## Card.2 1 6 0
## Card.3 1 5 6
#
## $call
## improve(cor(swiss), 2, 3, nsol = 4, criterion = "GCD", include = c(1))

## 3) An example of subset selection in the context of Multiple Linear Regression. Variable 5 (average car price) in the Cars93 MASS library data set is regressed on 13 other variables. Three variable subsets of cardinalities 4, 5 and 6 are requested, using the "XI_2" criterion which, in the case of a Linear Regression, is merely the standard Coefficient of Determination, R^2 (as are the other three criteria for the multivariate linear hypothesis, "TAU_2", "CCR1_2" and "ZETA_2").

library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,5])

names(Cars93[,5,drop=FALSE])
## [1] "Price"
colnames(CarsHmat$mat)
## [1] "MPG.city" "MPG.highway" "EngineSize"
## [4] "Horsepower" "RPM" "Rev.per.mile"
### $subsets
### , , Card.4
###
### Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
### Solution 1 3 4 11 13 0 0
### Solution 2 3 4 11 13 0 0
### Solution 3 4 5 10 11 0 0
###
### , , Card.5
###
### Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
### Solution 1 3 4 8 11 13 0
### Solution 2 4 5 10 11 12 0
### Solution 3 4 5 10 11 12 0
###
### , , Card.6
###
### Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
### Solution 1 4 5 6 10 11 12
### Solution 2 4 5 8 10 11 12
### Solution 3 4 5 9 10 11 12
###
### $values
###
### card.4 card.5 card.6
### Solution 1 0.6880773 0.6899182 0.7270257
### Solution 2 0.6880773 0.7241457 0.7271056
### Solution 3 0.7143794 0.7241457 0.7310150
###
### $bestvalues
### Card.4 Card.5 Card.6
### 0.7143794 0.7241457 0.7310150
###
### $bestsets
###
### Var.1 Var.2 Var.3 Var.4 Var.5 Var.6
### Card.4 4 5 10 11 0 0
### Card.5 4 5 10 11 12 0
### Card.6 4 5 9 10 11 12
###
### $call
###
### improve(CarsHmat$mat, kmin = 4, kmax = 6, H = CarsHmat$H, r = 1, criterion = "xi2", nsol = 3)
## 4) A Linear Discriminant Analysis example with a very small data set.
We consider the Iris data and three groups, defined by species (setosa, versicolor and virginica). The goal is to select the 2- and 3-variable subsets that are optimal for the linear discrimination (as measured by the "TAU_2" criterion).

data(iris)
irisHmat <- ldaHmat(iris[1:4],iris$Species)
improve(irisHmat$mat,kmin=2,kmax=3,H=irisHmat$H,r=2,crit="ccr12")

## $subsets
## , , Card.2
## # Solution 1 2 3 0
## , , Card.3
## # Solution 1 2 3 4

## $values
## card.2 card.3
## Solution 1 0.8079476 0.8419635

## $bestvalues
## Card.2 Card.3
## 0.8079476 0.8419635

## $bestsets
## , Var.1 Var.2 Var.3
## Card.2 2 3 0
## Card.3 2 3 4

## $call
## improve(mat = irisHmat$mat, kmin = 2, kmax = 3, criterion = "tau2", H = irisHmat$H, r = 2)

## 5) An example of subset selection in the context of a Canonical Correlation Analysis. Two groups of variables within the Cars93 MASS library data set are compared. The goal is to select 4- to 6-variable subsets of the 13-variable 'X' group that are optimal in terms of preserving the canonical correlations, according to the "ZETA_2" criterion (Warning: the 3-variable 'Y' group is kept intact; subset selection is carried out in the 'X' group only). The 'tolsym' parameter is used to relax the symmetry requirements on the effect matrix H which, for numerical reasons, is slightly asymmetric. Since corresponding off-diagonal entries of
## matrix H are different, but by less than tolsym, H is replaced
## by its symmetric part: \((H + t(H))/2\).

library(MASS)
data(Cars93)
CarsHmat <- lmHmat(Cars93[,c(7:8,12:15,17:22,25)],Cars93[,4:6])

names(Cars93[,4:6])

colnames(CarsHmat$mat)
## [1] "MPG.city" "MPG.highway" "EngineSize"
## [4] "Horsepower" "RPM" "Rev.per.mile"
## [7] "Fuel.tank.capacity" "Passengers" "Length"
## [10] "Wheelbase" "Width" "Turn.circle"
## [13] "Weight"

improve(CarsHmat$mat, kmin=4, kmax=6, H=CarsHmat$H, r=3, crit="zeta2", tolsym=1e-9)

## $subsets
## , , Card.4
## 
## , , Card.5
## 
## , , Card.6

## $values
## card.4 card.5 card.6
## Solution 1  0.4626035  0.4875495  0.5071096

## $bestvalues
## Card.4 Card.5 Card.6
## 0.4626035  0.4875495  0.5071096

## $bestsets
## , , Card.4
## , , Card.5
## , , Card.6

## $call
## improve(mat = CarsHmat$mat, kmin = 4, kmax = 6, criterion = "zeta2",
## H = CarsHmat$H, r = 3, tolsym = 1e-09)
##
## Warning message:
##
## The effect description matrix (H) supplied was slightly asymmetric:
## symmetric entries differed by up to 3.63797880709171e-12.
## (less than the \( \text{\textquoteleft tolsym\textquoteprime\textquoteperiod} \) parameter).
## The H matrix has been replaced by its symmetric part.
## in: validnovcrit(mat, criterion, H, r, p, tolval, tolsym)
##
## 6) An example of variable selection in the context of a logistic
## regression model. We consider the last 100 observations of
## the iris data set (versicolor and virgincia species) and try
## to find the best variable subsets for the model that takes species
## as response variable.

data(iris)
iris2sp <- iris[iris$Species != "setosa",]
logrfit <- glm(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
iris2sp,family=binomial)
Hmat <- glmHmat(logrfit)
improve(Hmat$mat,1,3,H=Hmat$H,r=1,criterion="Wald")

## $subsets
## , , Card.1
## #  Var.1 Var.2 Var.3
## # Solution 1  4  0  0

## , , Card.2
## #  Var.1 Var.2 Var.3
## # Solution 1  1  3  0

## , , Card.3
## #  Var.1 Var.2 Var.3
## # Solution 1  2  3  4

## $values
## #  card.1 card.2 card.3
## # Solution 1  4.894554  3.522885  1.060121

## $bestvalues
## #  Card.1 Card.2 Card.3
## #  4.894554  3.522885  1.060121

## $bestsets
## #  Var.1 Var.2 Var.3
ldaHmat

## Card.1 4 0 0
## Card.2 1 3 0
## Card.3 2 3 4

## $call
## improve(mat = Hmat$mat, kmin = 1, kmax = 3, criterion = "Wald",
## H = Hmat$H, r = 1)
## -----------------------------------------------

## It should be stressed that, unlike other criteria in the
## subselect package, the Wald criterion is not bounded above by
## 1 and is a decreasing function of subset quality, so that the
## 3-variable subsets do, in fact, perform better than their smaller-sized
## counterparts.

### ldaHmat

**Total and Between-Group Deviation Matrices in Linear Discriminant Analysis**

**Description**

Computes total and between-group matrices of Sums of Squares and Cross-Product (SSCP) deviations in linear discriminant analysis. These matrices may be used as input to the variable selection search routines `anneal`, `genetic_improve` or `eleaps`.

**Usage**

```r
## Default S3 method:
ldaHmat(x, grouping,...)

## S3 method for class 'data.frame'
ldaHmat(x, grouping,...)

## S3 method for class 'formula'
ldaHmat(formula,data=NULL,...)
```

**Arguments**

- **x**: A matrix or data frame containing the discriminators for which the SSCP matrix is to be computed.
- **grouping**: A factor specifying the class for each observation.
- **formula**: A formula of the form `groups ~ x1 + x2 + ...` That is, the response is the grouping factor and the right hand side specifies the (non-factor) discriminators.
- **data**: Data frame from which variables specified in 'formula' are preferentially to be taken.
- **...**: further arguments for the method.
Value

A list with four items:

mat The total SSCP matrix
H The between-groups SSCP matrix
r The expected rank of the H matrix which equals the minimum between the number of discriminators and the number of groups minus one. The true rank of H can be different from r if the discriminators are linearly dependent.
call The function call which generated the output.

See Also

anneal, genetic, improve, eleaps, lda.

Examples

```r
## An example with a very small data set. We consider the Iris data and three groups, defined by species (setosa, versicolor and virginica).

data(iris)
irisHmat <- ldaHmat(iris[,1:4], iris$Species)
irisHmat

#mat
# Sepal.Length Sepal.Width Petal.Length Petal.Width
# Sepal.Length 102.168333 -6.322667 189.8730 76.92433
# Sepal.Width -6.322667 28.306933 -49.1188 -18.12427
# Petal.Length 189.873000 -49.118800 464.3254 193.04580
# Petal.Width 76.924333 -18.124267 193.0458 86.56993

#H
# Sepal.Length Sepal.Width Petal.Length Petal.Width
# Sepal.Length 63.21213 -19.95267 165.2484 71.27933
# Petal.Length 165.24840 -57.23960 437.1028 186.77400
# Petal.Width 71.27933 -22.93267 186.7740 80.41333

#r
#[1] 2

#call
# ldaHmat.data.frame(x = iris[,1:4], grouping = iris$Species)
```
Total and Effect Deviation Matrices for Linear Regression and Canonical Correlation Analysis

Description

Computes total and effect matrices of Sums of Squares and Cross-Product (SSCP) deviations, divided by a normalizing constant, in linear regression or canonical correlation analysis. These matrices may be used as input to the variable selection search routines `anneal`, `genetic improve` or `eleaps`.

Usage

```r
# Default S3 method:
lmHmat(x, y, ...)
# S3 method for class 'data.frame'
lmHmat(x, y, ...)
# S3 method for class 'formula'
lmHmat(formula, data = NULL, ...)
# S3 method for class 'lm'
lmHmat(fitdlmmodel, ...)
```

Arguments

- `x`: A matrix or data frame containing the variables for which the SSCP matrix is to be computed.
- `y`: A matrix or data frame containing the set of fixed variables, the association of `x` is to be measured with.
- `formula`: A formula of the form `y ~ x1 + x2 + ...`. That is, the response is the set of fixed variables and the right hand side specifies the variables whose subsets are to be compared.
- `data`: Data frame from which variables specified in `formula` are preferentially to be taken.
- `fitdlmmodel`: An object of class `lm`, as produced by R’s `lm` function.
- `...`: Further arguments for the method.

Details

Let `x` and `y` be two different groups of linearly independent variables observed on the same set of data units. It is well known that the association between `x` and `y` can be measured by their squared canonical correlations which may be found as the positive eigenvalues of certain matrix products. In particular, if \( T_x \) and \( H_{x/y} \) denote SSCP matrices of deviations from the mean, respectively for
the original x variables \( (T_x) \) and for their orthogonal projections onto the space spanned by the y’s \( (H_{x/y}) \), then the positive eigenvalues of \( T_x^{-1}H_{x/y} \) equal the squared correlations between x and y. Alternatively these correlations could also be found from \( T_y^{-1}H_{y/x} \) but here, assuming a goal of comparing x’s subsets for a given fixed set of y’s, we will focus on the former product. \( \text{lmHmat} \) computes a scaled version of \( T_x \) and \( H_{x/y} \) such that \( T_x \) is converted into a covariance matrix. These matrices can be used as input to the search routines \texttt{anneal}, \texttt{genetic improve} and \texttt{eleaps} that try to select x subsets based on several functions of their squared correlations with y. We note that when there is only one variable in the y set, this is equivalent to selecting predictors for linear regression based on the traditional coefficient of determination.

Value

A list with four items:

- mat: The total SSCP matrix divided by nrow(x)-1
- H: The effect SSCP matrix divided by nrow(x)-1
- r: The expected rank of the H matrix which, under the assumption of linear independence, equals the minimum between the number of variables in the x and y sets. The true rank of H can be different from r if the linear independence condition fails.
- call: The function call which generated the output.

See Also

\texttt{anneal}, \texttt{genetic}, \texttt{improve}, \texttt{eleaps}, \texttt{lm}.

Examples

```r
##------------------------------------------------------------------
## 1) An example of subset selection in the context of Multiple
## Linear Regression. Variable 5 (average price) in the Cars93 MASS
## library is to be regressed on 13 other variables. The goal is to
## compare subsets of these 13 variables according to their ability
## to predict car prices.

library(MASS)
data(Cars93)
CarsHmat1 <- lmHmat(Cars93[c(7:8,12:15,17:22,25)],Cars93[5])
CarsHmat1
```

```r
##$mat
##     MPG.city   MPG.highway  EngineSize  Horsepower
###MPG.city  31.582281  28.283427 -4.1391655 -197.979897
###MPG.highway  28.283427  28.427302 -3.4667602 -172.865475
###EngineSize -4.139165 -3.466760  1.0761220  39.776998
###Horsepower -197.979897 -172.865475  39.776998  2.743079
###RPM  1217.478962  997.335203 -339.163744  1.146634
###Rev.per.mile  1941.631019 1555.243104 -424.411816 -1.561070
###Fuel.tank.capacity -14.985799 -13.743654  2.583082  1.222536
```
<table>
<thead>
<tr>
<th></th>
<th>Passengers</th>
<th>Length</th>
<th>Wheelbase</th>
<th>Width</th>
<th>Turn.circle</th>
<th>Weight</th>
<th>RPM</th>
<th>Rev.per.mile</th>
<th>Fuel.tank.capacity</th>
<th>Passengers</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPG.city</td>
<td>1217.4790</td>
<td>1941.6310</td>
<td>11.8197055</td>
<td>3.3992286</td>
<td>2.6029453</td>
<td>356088.7097</td>
<td>347977.8927</td>
<td>1729.462686</td>
<td>339.8953717</td>
<td></td>
</tr>
<tr>
<td>MPG.highway</td>
<td>997.3352</td>
<td>1555.2431</td>
<td>173.892824</td>
<td>127.543712</td>
<td>817.013371</td>
<td>146589.3233</td>
<td>246518.7295</td>
<td>-3844.91583</td>
<td>-5004.31393</td>
<td></td>
</tr>
<tr>
<td>EngineSize</td>
<td>-339.1637</td>
<td>-424.4118</td>
<td>3.3992286</td>
<td>2.6029453</td>
<td>1.9237962</td>
<td>82.021973</td>
<td>45.36793</td>
<td>20.803062</td>
<td>14.2807394</td>
<td></td>
</tr>
<tr>
<td>RPM</td>
<td>-289.62133</td>
<td>150636.1325</td>
<td>-1217.093268</td>
<td>-972.580645</td>
<td>-5004.31393</td>
<td>-2795.094670</td>
<td>1729.462686</td>
<td>339.8953717</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rev.per.mile</td>
<td>146589.3233</td>
<td>246518.7295</td>
<td>-3844.91583</td>
<td>-5004.31393</td>
<td>-1173.328074</td>
<td>-150636.1325</td>
<td>1729.462686</td>
<td>339.8953717</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MPG.city</th>
<th>MPG.highway</th>
<th>EngineSize</th>
<th>Horsepower</th>
</tr>
</thead>
<tbody>
<tr>
<td>##</td>
<td>11.1644681</td>
<td>9.9885440</td>
<td>-2.07077758</td>
<td>-137.938111</td>
</tr>
<tr>
<td>MPG.city</td>
<td>11.1644681</td>
<td>9.9885440</td>
<td>-2.07077758</td>
<td>-137.938111</td>
</tr>
<tr>
<td>MPG.highway</td>
<td>9.9885440</td>
<td>8.9364770</td>
<td>-1.85266802</td>
<td>-123.409453</td>
</tr>
<tr>
<td>Variable</td>
<td>Coefficient 1</td>
<td>Coefficient 2</td>
<td>Coefficient 3</td>
<td>Coefficient 4</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>EngineSize</td>
<td>-2.0707776</td>
<td>-1.8526680</td>
<td>0.38408635</td>
<td>25.584662</td>
</tr>
<tr>
<td>Horsepower</td>
<td>-137.9381108</td>
<td>-123.4094525</td>
<td>25.58466246</td>
<td>1704.239046</td>
</tr>
<tr>
<td>RPM</td>
<td>9.8795182</td>
<td>8.8389345</td>
<td>-1.83244599</td>
<td>-122.062428</td>
</tr>
<tr>
<td>Rev_per.mile</td>
<td>707.3855787</td>
<td>632.8785101</td>
<td>-131.2053714</td>
<td>-8739.818920</td>
</tr>
<tr>
<td>Fuel.tank.capacity</td>
<td>-6.7879289</td>
<td>-6.0729671</td>
<td>1.25901874</td>
<td>83.865437</td>
</tr>
<tr>
<td>Passengers</td>
<td>-0.200865141</td>
<td>-0.1797085</td>
<td>0.03725632</td>
<td>2.481709</td>
</tr>
<tr>
<td>Wheelbase</td>
<td>-11.4130722</td>
<td>-10.2109633</td>
<td>2.1168849</td>
<td>141.009639</td>
</tr>
<tr>
<td>Width</td>
<td>-5.7581866</td>
<td>-5.1516920</td>
<td>1.06802435</td>
<td>71.142967</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>-4.2281864</td>
<td>-3.7828426</td>
<td>0.78424099</td>
<td>52.239662</td>
</tr>
<tr>
<td>Weight</td>
<td>-1275.6139645</td>
<td>-1141.2569026</td>
<td>236.5996884</td>
<td>15760.337110</td>
</tr>
</tbody>
</table>

## Additional Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient 1</th>
<th>Coefficient 2</th>
<th>Coefficient 3</th>
<th>Coefficient 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPM</td>
<td>9.8795182</td>
<td>8.8389345</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Rev_per.mile</td>
<td>707.3855787</td>
<td>632.8785101</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Fuel.tank.capacity</td>
<td>-6.7879289</td>
<td>-6.0729671</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Passengers</td>
<td>-0.200865141</td>
<td>-0.1797085</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Wheelbase</td>
<td>-11.4130722</td>
<td>-10.2109633</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Width</td>
<td>-5.7581866</td>
<td>-5.1516920</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>-4.2281864</td>
<td>-3.7828426</td>
<td>4.55772770</td>
<td>303.598201</td>
</tr>
<tr>
<td>Weight</td>
<td>-1275.6139645</td>
<td>-1141.2569026</td>
<td>236.5996884</td>
<td>15760.337110</td>
</tr>
</tbody>
</table>
lmHmat

```r
## [1] 1

lmHmat.data.frame(x = Cars93[c(7:8, 12:15, 17:22, 25)], y = Cars93[5])

2) An example of subset selection in the context of Canonical Correlation Analysis. Two groups of variables within the Cars93 MASS library data set are compared. The first group (variables 4th, 5th and 6th) relates to price, while the second group is formed by 13 variables that describe several technical car specifications. The goal is to select subsets of the second group that are optimal in terms of preserving the canonical correlations with the variables in the first group (Warning: the 3-variable "response" group is kept intact; subset selection is to be performed only in the 13-variable group).

library(MASS)
data(Cars93)
CarsHmat2 <- lmHmat(Cars93[c(7:8, 12:15, 17:22, 25)], Cars93[4:6])

names(Cars93[4:6])

CarsHmat2
```

```r
##$mat
##$mat
## Name       MPG.city   MPG.highway EngineSize Horsepower
## MPG.city   31.582281  28.283427 -4.1391655 -1.979799e+02
## MPG.highway 28.283427  28.427302 -3.4667602 -1.728655e+02
## EngineSize  -4.139165 -3.466760  1.0761220  3.977700e+01
## Horsepower  -197.979897 -172.865475 39.7769986  2.743079e+03
## RPM          1217.479062  997.335283 -339.1637447  1.146634e+03
## Rev.per.mile 1941.631019 1555.243104 -424.4118163 -1.561070e+04
## Fuel.tank.capacity -14.985799 -13.743654 2.5830820  1.222536e+02
## RPM            0.4017181  5.040907  -2.4339645
## Rev.per.mile  -13.743654  517.1327139 33.063850  1.0794764
## Length        -54.673329 -42.267765 33.063850  7.3626695
## Wheelbase     -25.567087 -22.375760 11.8197055  4.212964e+02
## Width         -15.302127 -12.902291 122.253612  0.5040907
## Turn.circle   -12.871061 -10.202782 122.253612  0.5040907
## Weight        -2795.094670 -2549.654628 33.063850  7.3626695
## RPM           1217.479062  997.335283 -339.1637447  1.146634e+03
## Rev.per.mile  -14.985799 -13.743654 2.5830820  1.222536e+02
```
<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wheelbase</td>
<td>-1903.7693</td>
<td>-2156.2932</td>
<td>16.94481</td>
<td>4.917718</td>
</tr>
<tr>
<td>Width</td>
<td>-1217.0933</td>
<td>-1464.3712</td>
<td>9.898282</td>
<td>1.923796</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>-972.5806</td>
<td>-1173.3281</td>
<td>7.096283</td>
<td>1.503740</td>
</tr>
<tr>
<td>Weight</td>
<td>-150636.1325</td>
<td>-215349.6757</td>
<td>1729.468268</td>
<td>339.095371</td>
</tr>
<tr>
<td>MPG.city</td>
<td>-54.6733</td>
<td>-25.567087</td>
<td>-15.302127</td>
<td>-12.071061</td>
</tr>
<tr>
<td>MPG.highway</td>
<td>-42.2677</td>
<td>-22.375760</td>
<td>-12.902291</td>
<td>-10.202782</td>
</tr>
<tr>
<td>EngineSize</td>
<td>11.8197</td>
<td>5.181942</td>
<td>3.399229</td>
<td>2.602945</td>
</tr>
<tr>
<td>Horsepower</td>
<td>421.2964</td>
<td>173.892824</td>
<td>127.543712</td>
<td>94.742520</td>
</tr>
<tr>
<td>RPM</td>
<td>-3844.9158</td>
<td>-1903.7692</td>
<td>-1217.0933</td>
<td>-972.5806</td>
</tr>
<tr>
<td>Rev.per.mile</td>
<td>-5004.3139</td>
<td>-2156.2932</td>
<td>-1464.3712</td>
<td>-1173.3280</td>
</tr>
<tr>
<td>Fuel.tank.capacity</td>
<td>33.0635</td>
<td>16.94481</td>
<td>9.898282</td>
<td>7.096283</td>
</tr>
<tr>
<td>Passengers</td>
<td>7.36267</td>
<td>4.917719</td>
<td>1.923796</td>
<td>1.503740</td>
</tr>
<tr>
<td>Length</td>
<td>213.22955</td>
<td>82.021973</td>
<td>45.367929</td>
<td>34.780622</td>
</tr>
<tr>
<td>Wheelbase</td>
<td>82.02197</td>
<td>46.507948</td>
<td>20.803062</td>
<td>15.899836</td>
</tr>
<tr>
<td>Width</td>
<td>45.36793</td>
<td>20.803062</td>
<td>14.280739</td>
<td>9.962015</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>34.78062</td>
<td>15.899836</td>
<td>9.962015</td>
<td>10.389434</td>
</tr>
<tr>
<td>Weight</td>
<td>6945.16129</td>
<td>3507.549088</td>
<td>1950.471599</td>
<td>1479.36535</td>
</tr>
<tr>
<td>MPG.city</td>
<td>-2795.0947</td>
<td>-2549.6546</td>
<td>-149.055525</td>
<td>-132.381671</td>
</tr>
<tr>
<td>MPG.highway</td>
<td>-110802504</td>
<td>9.9241995</td>
<td>-2.1555142</td>
<td>-132.381671</td>
</tr>
<tr>
<td>EngineSize</td>
<td>-2.4485655</td>
<td>-2.1555142</td>
<td>0.48131168</td>
<td>28.438641</td>
</tr>
<tr>
<td>Horsepower</td>
<td>-149.055525</td>
<td>-132.381670</td>
<td>28.43864077</td>
<td>1788.168412</td>
</tr>
<tr>
<td>RPM</td>
<td>116.9463486</td>
<td>90.2758380</td>
<td>-29.90735790</td>
<td>-935.019669</td>
</tr>
<tr>
<td>Rev.per.mile</td>
<td>850.6791690</td>
<td>744.7148717</td>
<td>-168.44221351</td>
<td>-9825.172173</td>
</tr>
<tr>
<td>Fuel.tank.capacity</td>
<td>-7.3863845</td>
<td>-6.5473387</td>
<td>1.41367337</td>
<td>88.391549</td>
</tr>
<tr>
<td>Passengers</td>
<td>-0.2756475</td>
<td>-0.2507147</td>
<td>0.85519028</td>
<td>3.036255</td>
</tr>
<tr>
<td>Length</td>
<td>-29.808749</td>
<td>-25.4286533</td>
<td>5.74148535</td>
<td>337.880225</td>
</tr>
<tr>
<td>Wheelbase</td>
<td>-12.4579187</td>
<td>-11.0288656</td>
<td>2.38966697</td>
<td>148.928877</td>
</tr>
<tr>
<td>Width</td>
<td>-6.8768553</td>
<td>-6.0641799</td>
<td>1.35405290</td>
<td>79.579186</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>-4.9652258</td>
<td>-4.3460777</td>
<td>0.97719452</td>
<td>57.833523</td>
</tr>
<tr>
<td>Weight</td>
<td>-1399.801940</td>
<td>-1239.6883974</td>
<td>268.43952022</td>
<td>16693.580681</td>
</tr>
<tr>
<td>RPM</td>
<td>116.946347</td>
<td>850.679171</td>
<td>-7.3863845</td>
<td>-0.2756475</td>
</tr>
<tr>
<td>Rev.per.mile</td>
<td>90.275838</td>
<td>744.714871</td>
<td>-6.5473387</td>
<td>-0.2507147</td>
</tr>
<tr>
<td>Fuel.tank.capacity</td>
<td>-29.90735790</td>
<td>-168.44221351</td>
<td>1.41367337</td>
<td>88.391549</td>
</tr>
<tr>
<td>Passengers</td>
<td>-395.019669</td>
<td>-9825.172173</td>
<td>88.3915487</td>
<td>3.03625516</td>
</tr>
</tbody>
</table>
### RM Coefficient for Variable Subset Selection

```r
rm.coef

Computes the RM coefficient for variable subset selection
```

```
#\textbf{RPM} 8930.289631 11941.01945 -51.6620352 -3.30491485
#\textbf{Rev.per.mile} 11941.019450 59470.19997 -490.0061258 -18.1789645
#\textbf{Fuel.tank.capacity} -51.662035 -490.00613 4.3742368 0.14814085
#\textbf{Passengers} -3.3049151 -18.17896 0.1481409 0.01208877
#\textbf{Length} -397.601848 -2033.81167 16.8646785 0.57474210
#\textbf{Wheelbase} -93.828737 -830.92582 7.3783050 0.24261242
#\textbf{Width} -84.771418 -472.37388 3.9523474 0.16370704
#\textbf{Turn.circle} -64.578815 -345.33527 2.8839031 0.09876958
#\textbf{Weight} -10423.776629 -93087.56026 826.3348263 28.56899347
```

### Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Wheelbase</th>
<th>Width</th>
<th>Turn.circle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>-29.0878749</td>
<td>-12.4579187</td>
<td>-6.8768553</td>
<td>-4.96522585</td>
</tr>
<tr>
<td>Wheelbase</td>
<td>-25.4265633</td>
<td>-11.0208656</td>
<td>-6.0641799</td>
<td>-4.34607767</td>
</tr>
<tr>
<td>Width</td>
<td>5.7414854</td>
<td>2.3890670</td>
<td>1.3540529</td>
<td>0.97719452</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>337.8802249</td>
<td>148.9288871</td>
<td>79.5791065</td>
<td>57.83352310</td>
</tr>
<tr>
<td>RPM</td>
<td>-397.6018484</td>
<td>-93.8287370</td>
<td>-84.7714184</td>
<td>-64.57881537</td>
</tr>
<tr>
<td>Rev.per.mile</td>
<td>-2033.8116669</td>
<td>-830.9258201</td>
<td>-472.3738765</td>
<td>-345.33527111</td>
</tr>
<tr>
<td>Fuel.tank.capacity</td>
<td>16.8646785</td>
<td>7.3738050</td>
<td>3.9523474</td>
<td>2.88390313</td>
</tr>
<tr>
<td>Passengers</td>
<td>0.5747421</td>
<td>0.2426124</td>
<td>0.1637070</td>
<td>0.09876958</td>
</tr>
<tr>
<td>Length</td>
<td>69.9185465</td>
<td>28.6482825</td>
<td>16.0342179</td>
<td>11.86931842</td>
</tr>
<tr>
<td>Wheelbase</td>
<td>28.6482825</td>
<td>12.4615297</td>
<td>6.6687394</td>
<td>4.89477408</td>
</tr>
<tr>
<td>Width</td>
<td>16.0342179</td>
<td>6.6687394</td>
<td>3.8217667</td>
<td>2.73004255</td>
</tr>
<tr>
<td>Turn.circle</td>
<td>11.8693184</td>
<td>4.8947741</td>
<td>2.7300425</td>
<td>2.01640426</td>
</tr>
<tr>
<td>Weight</td>
<td>3199.4701647</td>
<td>1393.7884808</td>
<td>751.2183342</td>
<td>546.92139008</td>
</tr>
</tbody>
</table>

### Summary Statistics

- **RPM**: 8930.289631
- **Rev.per.mile**: 11941.019450
- **Fuel.tank.capacity**: -51.662035
- **Passengers**: -3.3049151
- **Length**: -397.601848
- **Wheelbase**: -93.828737
- **Width**: -84.771418
- **Turn.circle**: -64.578815
- **Weight**: -10423.776629

### Correlation Coefficients

- **RPM**: 0.8930289631
- **Rev.per.mile**: 0.99876958
- **Fuel.tank.capacity**: 0.57474209
- **Passengers**: 0.14814085
- **Length**: -0.87685534
- **Wheelbase**: -0.96522585
- **Width**: -0.94076679
- **Turn.circle**: -0.88390313
- **Weight**: -0.83903135

### Additional Code

```r
rm.coef
```

```r
r
[r[1] = 3

## Call

```r
lmHmat.data.frame(x = Cars93[c(7:8, 12:15, 17:22, 25)], y = Cars93[4:6])
```
Description

Computes the RM coefficient, measuring the similarity of the spectral decompositions of a p-variable data matrix, and of the matrix which results from regressing all the variables on a subset of only k variables.

Usage

rm.coef(mat, indices)

Arguments

- **mat**: the full data set's covariance (or correlation) matrix
- **indices**: a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different k-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.

Details

Computes the RM coefficient that measures the similarity of the spectral decompositions of a p-variable data matrix, and of the matrix which results from regressing those variables on a subset (given by "indices") of the variables. Input data is expected in the form of a (co)variance or correlation matrix. If a non-square matrix is given, it is assumed to be a data matrix, and its correlation matrix is used as input.

The definition of the RM coefficient is as follows:

\[ RM = \sqrt{\frac{\text{tr}(X^t P_v X)}{X^t X}} \]

where \( X \) is the full (column-centered) data matrix and \( P_v \) is the matrix of orthogonal projections on the subspace spanned by a k-variable subset.

This definition is equivalent to:

\[ RM = \sqrt{\sum_{i=1}^{p} \frac{\lambda_i (r_i^2)}{\sum_{j=1}^{p} \lambda_j}} \]

where \( \lambda_i \) stands for the i-th largest eigenvalue of the covariance matrix defined by X and \( r \) stands for the multiple correlation between the i-th Principal Component and the k-variable subset.

These definitions are also equivalent to the expression used in the code, which only requires the covariance (or correlation) matrix of the data under consideration.

The fact that indices can be a matrix or 3-d array allows for the computation of the RM values of subsets produced by the search functions `anneal`, `genetic` and `improve` (whose output option `$subsets` are matrices or 3-d arrays), using a different criterion (see the example below).

Value

The value of the RM coefficient.
rv.coef

Computes the RV-coefficient applied to the variable subset selection problem

Description

Computes the RV coefficient, measuring the similarity (after rotations, translations and global re-sizing) of two configurations of n points given by: (i) observations on each of p variables, and (ii) the regression of those p observed variables on a subset of the variables.

Usage

rv.coef(mat, indices)
Arguments

mat the full data set’s covariance (or correlation) matrix
indices a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different \( k \)-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.

Details

Input data is expected in the form of a (co)variance or correlation matrix of the full data set. If a non-square matrix is given, it is assumed to be a data matrix, and its correlation matrix is used as input. The subset of variables on which the full data set will be regressed is given by indices.

The RV-coefficient, for a (column-centered) data matrix (with \( p \) variables/columns) \( X \), and for the regression of these columns on a \( k \)-variable subset, is given by:

\[
RV = \frac{\text{tr}(XX^t \cdot (P_vX)(P_vX)^t)}{\sqrt{\text{tr}(XX^t)^2 \cdot \text{tr}((P_vX)(P_vX)^t)^2}}
\]

where \( P_v \) is the matrix of orthogonal projections on the subspace defined by the \( k \)-variable subset.

This definition is equivalent to the expression used in the code, which only requires the covariance (or correlation) matrix of the data under consideration.

The fact that indices can be a matrix or 3-d array allows for the computation of the RV values of subsets produced by the search functions anneal, genetic and improve (whose output option $subsets$ are matrices or 3-d arrays), using a different criterion (see the example below).

Value

The value of the RV-coefficient.

References


Examples

# A simple example with a trivially small data set

data(iris3)
x<-iris3[,,1]
rv.coef(var(x),c(1,3))
## [1] 0.8659685

## An example computing the RVs of three subsets produced when the
## anneal function attempted to optimize the RM criterion (using an
## absurdly small number of iterations).
data(swiss)
rmresults<-anneal(cor(swiss),2,nsol=4,niter=5,criterion="Rm")
rv.coef(cor(swiss),rmresults$subsets)

## Card.2
##Solution 1 0.8389669
##Solution 2 0.8663006
##Solution 3 0.8093862
##Solution 4 0.7529066

tau2.coef

---

Computes the Tau squared coefficient for a multivariate linear hypothesis

Description

Computes the Tau squared index of "effect magnitude". The maximization of this criterion is equivalent to the minimization of Wilk's lambda statistic.

Usage

tau2.coef(mat, H, r, indices,
tolval=10*.Machine$double.eps, tolsym=1000*.Machine$double.eps)

Arguments

mat the Variance or Total sums of squares and products matrix for the full data set.

H the Effect description sums of squares and products matrix (defined in the same way as the mat matrix).

r the Expected rank of the H matrix. See the Details below.

indices a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different k-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.

tolval the tolerance level to be used in checks for ill-conditioning and positive-definiteness of the 'total' and 'effects' (H) matrices. Values smaller than tolval are considered equivalent to zero.

tolsym the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes (A+t(A))/2.
Details

Different kinds of statistical methodologies are considered within the framework, of a multivariate linear model:

\[ X = A\Psi + U \]

where \( X \) is the (nxp) data matrix of original variables, \( A \) is a known (nxp) design matrix, \( \Psi \) an (qxp) matrix of unknown parameters and \( U \) an (nxp) matrix of residual vectors. The \( \tau^2 \) index is related to the traditional test statistic (Wilk’s lambda statistic) and measures the contribution of each subset to an Effect characterized by the violation of a linear hypothesis of the form \( C\Psi = 0 \), where \( C \) is a known coefficient matrix of rank \( r \). The Wilk’s lambda statistic (\( \lambda \)) is given by:

\[ \Lambda = \frac{\det(E)}{\det(T)} \]

where \( E \) is the Error matrix and \( T \) is the Total matrix. The index \( \tau^2 \) is related to the Wilk’s lambda statistic (\( \Lambda \)) by:

\[ \tau^2 = 1 - \lambda^{(1/r)} \]

where \( r \) is the rank of \( H \) the Effect matrix.

The fact that indices can be a matrix or 3-d array allows for the computation of the \( \tau^2 \) values of subsets produced by the search functions \texttt{anneal}, \texttt{genetic}, \texttt{improve} and \texttt{eleaps} (whose output option \$subsets are matrices or 3-d arrays), using a different criterion (see the example below).

Value

The value of the \( \tau^2 \) coefficient.

Examples

```r
## 1) A Linear Discriminant Analysis example with a very small data set.
## We considered the Iris data and three groups,
## defined by species (setosa, versicolor and virginica).

data(iris)
irisHmat <- ldaHmat(iris[1:4],iris$Species)
tau2.coef(irisHmat$mat,H=irisHmat$H,r=2,c(1,3))
## [1] 0.8003044

## 2) An example computing the value of the tau_2 criterion for two
## subsets produced when the anneal function attempted to optimize
## the xi_2 criterion (using an absurdly small number of iterations).
xireresults<-anneal(irisHmat$mat,2,nso=2,niter=2,criterion="xi2", H=irisHmat$H,r=2)
tau2.coef(irisHmat$mat,H=irisHmat$H,r=2,xireresults$subsets)
## Card.2```

```
trim.matrix

Given an ill-conditioned square matrix, deletes rows/columns until a well-conditioned submatrix is obtained.

Description

This function seeks to deal with ill-conditioned matrices, for which the search algorithms of optimal k-variable subsets could encounter numerical problems. Given a square matrix mat which is assumed positive semi-definite, the function checks whether it has reciprocal of the 2-norm condition number (i.e., the ratio of the smallest to the largest eigenvalue) smaller than tolval. If not, the matrix is considered well-conditioned and remains unchanged. If the ratio of the smallest to largest eigenvalue is smaller than tolval, an iterative process is begun, which deletes rows/columns (using Jolliffe’s method for subset selections described on pg. 138 of the Reference below) until a principal submatrix with reciprocal of the condition number larger than tolval is obtained.

Usage

trim.matrix(mat, tolval = 10 * .Machine$double.eps)

Arguments

mat  a symmetric matrix, assumed positive semi-definite.
tolval  the tolerance value for the reciprocal condition number of matrix mat.

Details

For the given matrix mat, eigenvalues are computed. If the ratio of the smallest to the largest eigenvalue is less than tolval, matrix mat remains unchanged and the function stops. Otherwise, an iterative process is begun, in which the eigenvector associated with the smallest eigenvalue is considered and its largest (in absolute value) element is identified. The corresponding row/column are deleted from matrix mat and the eigendecomposition of the resulting submatrix is computed. This iterative process stops when the ratio of the smallest to largest eigenvalue is not smaller than tolval.

The function checks whether the input matrix is square, but not whether it is positive semi-definite. This trim.matrix function can be used to delete rows/columns of square matrices, until only non-negative eigenvalues appear.
Value

Output is a list with four items:

- `trimmedmat` is a principal submatrix of the original matrix, with the ratio of its smallest to largest eigenvalues no smaller than `tolval`. This matrix can be used as input for the search algorithms in this package.

- `numbers.discriminated` is a list of the integer numbers of the original variables that were discarded.

- `names.discriminated` is a list of the original column numbers of the variables that were discarded.

- `size` is the size of the output matrix.

Note

When the `trim.matrix` function is used to produce a well-conditioned matrix for use with the `anneal`, `genetic`, `improve` or `eleaps` functions, care must be taken in interpreting the output of those functions. In those search functions, the selected variable subsets are specified by variable numbers, and those variable numbers indicate the position of the variables in the input matrix. Hence, if a trimmed matrix is supplied to functions `anneal`, `genetic`, `improve` or `eleaps`, variable numbers refer to the trimmed matrix.

References


Examples

```r
# a trivial example, for illustration of use: creating an extra column, # as the sum of columns in the "iris" data, and then using the function # trim.matrix to exclude it from the data's correlation matrix
data(iris)
lindepir<-cbind(apply(iris[,-5],1,sum),iris[,-5])
colnames(lindepir)[1]<-"Sum"
cor(lindepir)
## Sum Sepal.Length Sepal.Width Petal.Length Petal.Width
##Sum 1.0000000 0.9409143 -0.2230928 0.9713793 0.9538850
##Sepal.Length 0.9409143 1.0000000 -0.1175698 0.8717538 0.8179411
##Sepal.Width -0.2230928 -0.1175698 1.0000000 -0.4284401 -0.3661259
##Petal.Length 0.9713793 0.8717538 -0.4284401 1.0000000 0.9628654
##Petal.Width 0.9538850 0.8179411 -0.3661259 0.9628654 1.0000000
trim.matrix(cor(lindepir))
##$trimmedmat
## Sepal.Length Sepal.Width Petal.Length Petal.Width
##Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411
##Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259
```

wald.coef

Wald statistic for variable selection in generalized linear models

### Description

Computes the value of Wald's statistic, testing the significance of the excluded variables, in the context of variable subset selection in generalized linear models.
Usage

wald.coef(mat, H, indices, 
tolval=10*.Machine$double.eps, tolsym=1000*.Machine$double.eps)

Arguments

mat  An estimate (FI) of Fisher’s information matrix for the full model variable-coefficient estimates
H   A matrix product of the form FI %*% b %*% t(b) %*% FI where b is a vector of variable-coefficient estimates
indices  a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different k-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.
tolval  the tolerance level to be used in checks for ill-conditioning and positive-definiteness of the Fisher Information and the auxiliar (H) matrices. Values smaller than tolval are considered equivalent to zero.
tolsym  the tolerance level for symmetry of the Fisher Information and the auxiliar (H) matrices. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes (A+t(A))/2.

Details

Variable selection in the context of generalized linear models is typically based on the minimization of statistics that test the significance of excluded variables. In particular, the likelihood ratio, Wald’s, Rao’s and some adaptations of such statistics, are often proposed as comparison criteria for variable subsets of the same dimensionality. All these statistics are asymptotically equivalent and can be converted into information criteria, like the AIC, that are also able to compare subsets of different dimensionalities (see references [1] and [2] for further details).

Among these criteria, Wald’s statistic has some computational advantages because it can always be derived from the same (concerning the full model) maximum likelihood and Fisher information estimates. In particular, if \( W_{\text{allv}} \) is the value of the Wald statistic testing the significance of the full covariate vector, b and FI are coefficient and Fisher information estimates and H is an auxiliary rank-one matrix given by \( H = FI %*% b %*% t(b) %*% FI \), it follows that the value of Wald’s statistic for the excluded variables \( W_{\text{exc}} \) in a given subset is given by \( W_{\text{exc}} = W_{\text{allv}} - \text{tr}(FI_{\text{indices}}^{-1}H_{\text{indices}}) \), where \( FI_{\text{indices}} \) and \( H_{\text{indices}} \) are the portions of the FI and H matrices associated with the selected variables.

The FI and H matrices can be retrieved (from a glm object) by the \texttt{glmHmat} function and may be used as input to the search functions \texttt{anneal}, \texttt{genetic}, \texttt{improve} and \texttt{eleaps}. The Wald function computes the value of Wald statistic from these matrices for a subset specified by indices.

The fact that indices can be a matrix or 3-d array allows for the computation of the Wald statistic values of subsets produced by the search functions \texttt{anneal}, \texttt{genetic}, \texttt{improve} and \texttt{eleaps} (whose output option \$subsets are matrices or 3-d arrays), using a different criterion (see the example below).
Value

The value of the Wald statistic.

References


Examples

```r
library(MASS)
data(crabs)
LFL <- log(crabs$FL)
LRW <- log(crabs$RW)
LCL <- log(crabs$CL)
LCW <- log(crabs$CW)
logrfit <- glm(sex ~ FL + RW + CL + CW + LFL + LRW + LCL + LCW, crabs,family=binomial)
## Warning message:
## fitted probabilities numerically 0 or 1 occurred in: glm.fit(x = X, y = Y, 
## weights = weights, start = start, etastart = etastart,
##Hmat <- glmHmat(logrfit)
wald.coef(lHmat$mat,lHmat$H,c(1,6,7),tolsym=1E-06)
## [1] 2.286739
## Warning message:
## The covariance/total matrix supplied was slightly asymmetric:
## symmetric entries differed by up to 6.57252030578093e-14.
## (less than the 'tolsym' parameter).
## It has been replaced by its symmetric part.
## in: validmat(mat, p, tolval, tolsym)
```

## 2) An example computing the value of the Wald statistic in a logistic model for five subsets produced when a probit model was originally
library(MASS)
data(crabs)
lFL <- log(crabs$FL)
lRW <- log(crabs$RW)
lCL <- log(crabs$CL)
lCW <- log(crabs$CW)
probfit <- glm(sex ~ FL + RW + CL + CW + lFL + lRW + lCL + lCW, crabs, family=binomial(link=probit))
## Warning message:
## fitted probabilities numerically 0 or 1 occurred in: glm.fit(x = X, y = Y, 
## weights = weights, start = start, etastart = etastart)

pHmat <- glmHmat(probfit)
probresults <- eleaps(pHmat$mat,kmin=3,kmax=3,nsol=5,criterion="Wald",H=pHmat$H, r=1,tolsym=1E-10)
## Warning message:
## The covariance/total matrix supplied was slightly asymmetric:
## symmetric entries differed by up to 3.14069889205964e-12.
## (less than the 'tolsym' parameter).
## It has been replaced by its symmetric part.
## in: validmat(mat, p, tolval, tolsym)

logrfit <- glm(sex ~ FL + RW + CL + CW + lFL + lRW + lCL + lCW, crabs, family=binomial)
## Warning message:
## fitted probabilities numerically 0 or 1 occurred in: glm.fit(x = X, y = Y, 
## weights = weights, start = start, etastart = etastart)

lHmat <- glmHmat(logrfit)
wald.coef(lHmat$mat,H=lHmat$H,probresults$subsets,tolsym=1e-06)
## Card.3
## Solution 1 2.286739
## Solution 2 2.595165
## Solution 3 2.585149
## Solution 4 2.669059
## Solution 5 2.690954
## Warning message:
## The covariance/total matrix supplied was slightly asymmetric:
## symmetric entries differed by up to 6.57252030578093e-14.
## (less than the 'tolsym' parameter).
## It has been replaced by its symmetric part.
## in: validmat(mat, p, tolval, tolsym)

xi2.coef Computes the Xi squared coefficient for a multivariate linear hypothesis
xi2.coef

Description
Computes the Xi squared index of "effect magnitude". The maximization of this criterion is equivalent to the maximization of the traditional test statistic, the Bartlett-Pillai trace.

Usage
```r
xi2.coef(mat, H, r, indices,
tolval=10*.Machine$double.eps, tolsym=1000*.Machine$double.eps)
```

Arguments
- `mat`: the Variance or Total sums of squares and products matrix for the full data set.
- `H`: the Effect description sums of squares and products matrix (defined in the same way as the `mat` matrix).
- `r`: the Expected rank of the `H` matrix. See the Details below.
- `indices`: a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different `k`-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.
- `tolval`: the tolerance level to be used in checks for ill-conditioning and positive-definiteness of the 'total' and 'effects' (`H`) matrices. Values smaller than `tolval` are considered equivalent to zero.
- `tolsym`: the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (`H`) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes \((A + (A)')/2\).

Details
Different kinds of statistical methodologies are considered within the framework, of a multivariate linear model:

\[ X = A\Psi + U \]

where \(X\) is the \((nxp)\) data matrix of original variables, \(A\) is a known \((nxp)\) design matrix, \(\Psi\) an \((qxp)\) matrix of unknown parameters and \(U\) an \((nxp)\) matrix of residual vectors. The Xi squared index is related to the traditional test statistic (Bartlett-Pillai trace) and measures the contribution of each subset to an Effect characterized by the violation of a linear hypothesis of the form \(C\Psi = 0\), where \(C\) is a known coefficient matrix of rank \(r\). The Bartlett-Pillai trace \((P)\) is given by: \(P = tr(HT^{-1})\) where \(H\) is the Effect matrix and \(T\) is the Total matrix. The Xi squared index is related to Bartlett-Pillai trace \((P)\) by:

\[ \xi^2 = \frac{P}{r} \]

where \(r\) is the rank of \(H\) matrix.

The fact that `indices` can be a matrix or 3-d array allows for the computation of the Xi squared values of subsets produced by the search functions `anneal`, `genetic`, `improve` and `eleaps` (whose
output option $\$subsets$ are matrices or 3-d arrays), using a different criterion (see the example below).

**Value**

The value of the $\zeta^2$ coefficient.

**Examples**

```
## ---------------------------------------------------------------
## 1) A Linear Discriminant Analysis example with a very small data set.
## We considered the Iris data and three groups,
## defined by species (setosa, versicolor and virginica).

data(iris)
irisHmat <- ldaHmat(iris[,1:4],iris$Species)
xi2.coef(irisHmat$mat,H=irisHmat$H,r=2,c(1,3))
## [1] 0.4942503

## ---------------------------------------------------------------
## 2) An example computing the value of the $\xi_2$ criterion for two subsets
## produced when the anneal function attempted to optimize the $\tau_2$
## criterion (using an absurdly small number of iterations).

tauresults<-anneal(irisHmat$mat,2,nsol=2,niter=2,criterion="tau2",
                   H=irisHmat$H,r=2)
xi2.coef(irisHmat$mat,H=irisHmat$H,r=2,tauresults$subsets)

## Card.2
##Solution 1 0.5718811
##Solution 2 0.5232262

## ---------------------------------------------------------------
```

---

### zeta2.coef

**Computes the Zeta squared coefficient for a multivariate linear hypothesis**

**Description**

Computes the Zeta squared index of "effect magnitude". The maximization of this criterion is equivalent to the maximization of the traditional test statistic, the Lawley-Hotelling trace.

**Usage**

```
zeta2.coef(mat, H, r, indices, tolval=10*.Machine$double.eps, tolsym=1000*.Machine$double.eps)
```
Arguments

mat
the Variance or Total sums of squares and products matrix for the full data set.

H
the Effect description sums of squares and products matrix (defined in the same way as the mat matrix).

r
the Expected rank of the H matrix. See the Details below.

indices
a numerical vector, matrix or 3-d array of integers giving the indices of the variables in the subset. If a matrix is specified, each row is taken to represent a different k-variable subset. If a 3-d array is given, it is assumed that the third dimension corresponds to different cardinalities.

tolval
the tolerance level to be used in checks for ill-conditioning and positive-definiteness of the ‘total’ and ‘effects’ (H) matrices. Values smaller than tolval are considered equivalent to zero.

tolsym
the tolerance level for symmetry of the covariance/correlation/total matrix and for the effects (H) matrix. If corresponding matrix entries differ by more than this value, the input matrices will be considered asymmetric and execution will be aborted. If corresponding entries are different, but by less than this value, the input matrix will be replaced by its symmetric part, i.e., input matrix A becomes (A+(A))/2.

Details

Different kinds of statistical methodologies are considered within the framework, of a multivariate linear model:

\[ X = A \Psi + U \]

where \( X \) is the (nxp) data matrix of original variables, \( A \) is a known (nxp) design matrix, \( \Psi \) an (qxp) matrix of unknown parameters and \( U \) an (nxp) matrix of residual vectors. The \( \zeta^2 \) index is related to the traditional test statistic (Lawley-Hotelling trace) and measures the contribution of each subset to an Effect characterized by the violation of a linear hypothesis of the form \( C \Psi = 0 \), where \( C \) is a known coefficient matrix of rank \( r \). The Lawley-Hotelling trace is given by: \( V = tr(H E^{-1}) \) where \( H \) is the Effect matrix and \( E \) is the Error matrix. The index \( \zeta^2 \) is related to Lawley-Hotelling trace (\( V \)) by:

\[ \zeta^2 = \frac{V}{V + r} \]

where \( r \) is the rank of \( H \) matrix.

The fact that indices can be a matrix or 3-d array allows for the computation of the \( \zeta^2 \) values of subsets produced by the search functions anneal, genetic, improve and eleaps (whose output option $subsets are matrices or 3-d arrays), using a different criterion (see the example below).

Value

The value of the \( \zeta^2 \) coefficient.

Examples

```r
## ---------------------------------------------------------------
```
## 1) A Linear Discriminant Analysis example with a very small data set.  
## We considered the Iris data and three groups,  
## defined by species (setosa, versicolor and virginica).

```r
data(iris)
irisHmat <- ldaHmat(iris[1:4], iris$Species)
zeta2.coef(irisHmat$mat, H=irisHmat$H, r=2, c(1,3))
## [1] 0.9211501
```

## 2) An example computing the value of the zeta_2 criterion for two  
## subsets produced when the anneal function attempted to optimize  
## the ccr1_2 criterion (using an absurdly small number of iterations).

```r
ccr1results<-anneal(irisHmat$mat, 2, nsol=2, niter=2, criterion="ccr12",  
H=irisHmat$H, r=2)
zeta2.coef(irisHmat$mat, H=irisHmat$H, r=2, ccr1results$subsets)
```

```
Card.2
Solution 1 0.9105021
Solution 2 0.9161813
```
Index

* Genetic algorithms  
genetic, 29

* Leaps and Bounds algorithm  
eleaps, 15

* Local search algorithms  
  improve, 47

* Simulated annealing  
anneal, 2

* datasets  
farm, 25

* manip  
anneal, 2  
ccr12.coef, 13  
eleaps, 15  
genetic, 29  
glhHmat, 38  
glmHmat, 44  
  improve, 47  
ldaHmat, 57  
  lmHmat, 59  
rm.coef, 65  
rv.coef, 67  
tau2.coef, 69  
trim.matrix, 71  
wald.coef, 73  
xi2.coef, 76  
zeta2.coef, 78

  anneal, 2, 6, 14, 18, 28, 33, 38–40, 44, 45, 50,  
  57–60, 66, 68, 70, 72, 74, 77, 79

  ccr12.coef, 3, 6, 13, 15, 18, 30, 33, 48, 50

  eleaps, 6, 15, 33, 38–40, 44, 45, 50, 57–60, 70, 72, 74, 77, 79

  farm, 25

  gcd.coef, 3, 6, 15, 16, 18, 27, 30, 33, 48, 50

  genetic, 3, 6, 14, 18, 28, 29, 33, 38–40, 44, 45, 48, 50, 57–60, 66, 68, 70, 72, 74, 77, 79

  glhHmat, 5, 6, 17, 18, 32, 33, 38, 50

  glm, 45

  glmHmat, 5, 6, 17, 18, 32, 33, 43, 50, 74

  improve, 3, 14, 28, 30, 31, 38–40, 44, 45, 47,  
  48, 57–60, 66, 68, 70, 72, 74, 77, 79

  lda, 58

  ldaHmat, 5, 6, 17, 18, 32, 33, 40, 50, 57

  leaps (eleaps), 15

  lm, 60

  lmHmat, 5, 6, 17, 18, 32, 33, 40, 50, 59

  rm.coef, 3, 6, 15, 18, 30, 33, 48, 50, 65

  rv.coef, 3, 6, 15, 18, 30, 33, 48, 50, 67

  tau2.coef, 3, 6, 15, 18, 30, 33, 48, 50, 69

  trim.matrix, 4, 6, 16, 18, 31, 33, 49, 50, 71, 72

  wald.coef, 15, 18, 73

  xi2.coef, 3, 6, 15, 18, 30, 33, 48, 50, 76

  zeta2.coef, 3, 6, 15, 18, 30, 33, 48, 50, 78