Subselect 0.99: Selecting variable subsets in multivariate linear models

A. Pedro Duarte Silva and Jorge Cadima and Manuel Minhoto and Jorge Orestes Cerdeira

The subselect package combines, in a unified framework, a set of search routines that look for k-variable subsets that are good surrogates for a full p-variable data set. In version 0.8, presented at User!2004, no assumptions were made about the intended use of the data, and the criteria implemented measured the quality of the surrogates through different functions of the original covariance or correlation matrices.

The new version, 0.99, extends the package incorporating criteria that are more relevant when it is known that the data was collected with a particular type of analysis in view. Different kinds of statistical methodologies are considered within the framework of a multivariate linear model X = A Ψ + U, where X is the (n*p) data matrix of original variables, A is a known (n*q) design matrix, Ψ an (q*p) matrix of unknown parameters and U and (n*p) matrix of residual vectors. The new criteria are several descriptive indices, related to traditional test statistics, that measure the contribution of each subset to an "effect" characterized by the violation of a linear hypothesis of the form C Ψ= 0, where C is a known coefficient matrix of rank r. All these indices are functions of the r positive eigenvalues of a product H T-1 where H and T are matrices of "effect" and "total" squared and cross-product deviations associated with X.

Important cases within this framework include traditional canonical correlation analysis, in which the columns of A are observations on a fixed set of variables related to X, and C a (q*q) identity. If A consists on a single (dependent) variable, then the problem reduces to the traditional selection problem in linear regression analysis, using the R^2 coefficient as comparison criterion. Variable selection in generalized linear models can also be easily accommodated by specifying H and T in terms of appropriate Fisher information matrices. Linear Discriminant Analysis can be addressed by making A a matrix of group indicators, Ψ a matrix of group specific population means and the hypothesis C Ψ = 0 equivalent to the equality of all population means across groups. Searches for the variable subsets that best characterize a multivariate effect found by a multi-way MANOVA or MANCOVA analysis can also be easily addressed.

All the previous features and options of the subselect package are applicable to these new problems and criteria. In particular, for a moderate number of original variables, say less than 20 or 30, it is often possible to conduct an exhaustive search through all subsets using efficient adaptations of the classical Furnival and Wilson algorithm for variable selection. For larger data sets, several effective meta-heuristics are provided through reliable and updated implementations of simulated annealing, genetic and restricted local improvement algorithms. Furthermore, it is possible to forcibly include or exclude variables from the chosen subsets, specify the number of solutions to keep in each dimension, control several tuning parameters in the random search routines, specify a limit for the time spent in an exhaustive search and so on.

Key words: Variable selection algorithms. Heuristics. Linear Models. Generalized linear models. Discriminant analysis. Canonical correlation analysis