The `subselect` package - Selecting variable subsets in an exploratory data analysis.

Cadima, J.[*] Cerdeira, J.O.[†] Duarte Silva, A.P.[‡]& Minhoto, M.[§]

February 13, 2004

Identifying a subset of a large set of variables which can adequately replace the full data set is a problem that has been studied in different contexts and which is of widespread concern to analysts of large data sets.

The R package subselect provides functions which measure the quality of a given subset of k variables according to three criteria that are relevant for exploratory data analysis. These criteria measure: the similarity between subspaces spanned by a given subset of variables and a given subset of Principal Components of the full data set (via the mean of the squared canonical correlations of both sets of variables - Yanai's GCD); the similarity of the configurations of points obtained using all the original variables or only those in a given subset (via Escoufier's RV-coefficient); and the quality of a given variable subset as a predictor of all the original variables (McCabe's second criterion for Principal Variables). The subselect package also provides three different algorithms to search for optimal k-subsets, with respect to each of these criteria: a simulated annealing algorithm, a genetic algorithm and a modified local search algorithm.

New features of the package include an algorithm for a complete search (in the spirit of Furnival and Wilson's leaps-and-bounds algorithm for subset selection in multiple regression), which is viable for medium-sized data sets.

The issue of identifying a broad array of k-subsets that are maximal according to those three criteria taken simultaneously is also considered in the context of multi-criteria optimization.

[*]Departamento de Matemática, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Lisbon, Portugal.
[†]Departamento de Matemática, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Lisbon, Portugal.
[‡]Faculdade de Ciências Económicas e Empresariais, Universidade Católica Portuguesa, Oporto, Portugal.
[§]Departamento de Matemática, Universidade de Évora, Évora, Portugal.