Subset Selection

Many modelling problems involve choosing the best subset of features, variables or attributes. For example, when fitting a linear regression model you might be interested in the subset of variables that best describe the data.

There are a number of different ways that you can define the best subset, and a number of different strategies that can be adopted when searching for it. Again, using linear regression as an example, two widely used subset selection techniques are forward selection (G02EEF) and stepwise selection (G02EFF).

If you know *a-priori* the size of the subset you are interested in, for example, you have m variables and want the best regression model consisting of p variables, $p \leq m$, then a more general approach to picking the best subset might be to try all possible combinations of p variables and select the model that fit the data the best. However, there are $\frac{m!}{p!(m-p)!}$ possible combinations, which gets quickly out of hand as m increases, for example, $m = 20$ and $p = 4$ gives 4845 possible combinations. This general subset selection problem can be described as follows:

Given $\Omega = \{x_{\mathrm{i}} : i \in \mathbb{Z}, 1 \leq i \leq m\}$, a set of m unique features and a scoring mechanism $f(S)$ defined for all $S \subseteq \Omega$, then the optimal subset of size p is defined as the solution to

$$\underset{S \subseteq \Omega}{\mathrm{maximize}} f(S) \qquad \text{subject to} \qquad |S| = p \tag{1}$$

where $|S|$ denotes the cardinality of S, the number of elements in the set.

When maximising equation (1), rather than try all possible combinations, Ridout[2], based on some earlier work by Narendra and Fukunaga[1] developed a more efficient branch and bound algorithm for solving this particular subset selection problem. This algorithm is a general purpose algorithm (and so is not restricted to linear regression), the only restriction being on the scoring mechanism, f, which must satisfy:

$$f(S_i) \leq f(S_j) \quad \text{for all } S_j \subseteq \Omega \text{ and } S_i \subseteq S_j$$

The algorithm of Ridout[2] is now available in the NAG library, as either reverse (H05AAF) or direct communication (H05ABF). The terms reverse and direct communication are used in the NAG documentation to describe how a user supplied function (in this case, the scoring mechanism f) is passed to the NAG routine. More details on this, and the branch and bound algorithm, can be found in the individual routine documentation.

References

[1] P M Narendra and K Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 9:917–922, 1977.

[2] M S Ridout. Algorithm AS 233: An improved branch and bound algorithm for feature subset selection. *Journal of the Royal Statistics Society, Series C (Applied Statistics) (Volume 37)*, 1:139–147, 1988.