Improving Computational Efficiency in Identifying Parsimonious Statistical Models

Joseph Valentin, Ken Aho, John Edwards, Dewayne Derryberry, and Teri Peterson
Idaho State University

Many authors have argued that identifying parsimonious statistical models (those that are neither overfit nor underfit) while considering curvature and/or interaction terms among predictors is inadvisable because of the huge number of potential models. For example, the complete second order model set will contain $\sum_{j=0}^{k} \binom{k}{j} 2^{(j^2+j)/2}$ models for consideration where k is the number of predictors in the model. To address this difficulty, we present a stepwise algorithm, developed for the R statistical environment, in which the number of considered models is quadratic in *k*. This is in contrast with conventional stepwise model selection functions (e.g., StepAIC and step) which consider a model set cubic in *k*. Our new approach, termed Greedy, uses one of 3 measures of statistical parsimony for its model set, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and its predicted residual error sum of squares (PRESS) statistic. We found that, when considering large and/or high dimensional datasets, the Greedy algorithm identified the same optimal (minimum AIC) model as conventional stepwise approaches, or one with essentially equal parsimony, while having dramatically smaller computational run times. **This work was made possible by the NSF Idaho EPSCoR Program and by the National Science Foundation under award number IIA-1301792.**