Fast forward selection for correlated clustered data with a large number of predictor variables

Jakub Stoklosa[a], **Heloise Gibb**[b] and **David I. Warton**[a]

[a]School of Mathematics and Statistics, Evolution & Ecology Research Centre.
The University of New South Wales
Sydney, Australia
j.stoklosa@unsw.edu.au

[b]Department of Zoology
La Trobe University
Bundoora, Australia

Keywords: Generalized Estimating Equations; Model Selection; Multivariate Analysis; Negative Binomial Regression; Score Statistics; Trait/Environment Relationship.

Abstract: We propose a new variable selection criterion designed for use with forward selection algorithms; the score information criterion (SIC). The proposed criterion is based on generalized estimating equations (GEE) and score statistics which incorporate correlated response data. The main advantages of SIC, as compared to competing GEE information criteria, are: (1) it is much faster to compute than existing model selection criteria when the number of predictor variables added to a model is large; (2) it incorporates the correlation between variables into its quasi-likelihood. Our motivating example arises in ecology, and involves selecting from a large number of interaction terms in order to explain environmental-trait association in an arthropod community. In addition to applying SIC to the arthropod data, we also show by theory and simulation that SIC has a number of desirable properties, such as model selection consistency and efficiency.