Linguistics in the age of biostatistics: 
how data properties matter, and the case of Bayesian clustering

Erich Round, U. Queensland, Australia

Modern computational methods open up new directions of enquiry utilizing cross-linguistic datasets whose size would overwhelm traditional pencil-and-paper analyses. However, where techniques are not originally designed for linguistic research, questions will arise as to how performance might be affected by properties of linguistic datasets, whose principles of design have yet to alter much from the paper-and-pencil era. Issues such as non-independence of variables may be innocuous in some cases (Pagel & Meade 2006), but precisely when and whether this generalizes requires ongoing investigation.

In a series of experiments I examine the performance of the Bayesian clustering algorithm \textsc{Structure} (Pritchard et al. 2000) by making controlled variations to <10% of a large cross-linguistic dataset of typological variables (Reesink et al. 2009—121 languages; 155 binary variables; 88% of cells complete). I find that in terms of both the number of clusters inferred, and the languages they group together, \textsc{Structure} is indeed sensitive to the inclusion/exclusion of variables which have:

i. A high proportion of the same value across languages;

ii. A high number of missing values;

iii. A tendency to pair-wise correlation with another variable

iv. A relatively high level of correlation with all variables in the dataset

I consider two ramifications. Firstly, typological linguistic datasets do differ in some respects from the kinds of biological datasets which serve as test data for biostatistical techniques. Consequently, it will be desirable to test these techniques with specifically linguistic data, to best ascertain how they perform in the linguistic domain. Secondly, a typologist’s decision over whether or not to build variables X, Y or Z into a dataset may hinge precisely on issues such as (i–iv), which can then influence the final statistical result — a manifestation of the “researcher degrees of freedom” problem (Simmons et al. 2011). Thus, to aid the evaluation of cutting-edge research, it will be desirable when datasets are designed, for the accompanying decision-making process to be documented and published, since those decisions may bear upon the eventual results, even before language coding begins.


