On the canonical correlation analysis of bi-allelic genetic markers

Jan Graffelman

Universitat Politècnica de Catalunya, Spain, email: jan.graffelman@upc.edu

Keywords: biplot, generalized inverse, Hardy-Weinberg equilibrium, linkage disequilibrium.

Multivariate analysis is becoming increasingly relevant in genetics, due to the automated generation of large databases of genetic markers, single nucleotide polymorphisms (SNPs) in particular. Most SNPs are bi-allelic, and individuals can be characterized generically as AA, AB or BB. Such genotype data can be coded in an indicator matrix. Additional indicators can be defined to indicate whether an individual is a carrier or a non-carrier of a particular allele.

Genetic markers are usually expected to be in Hardy-Weinberg equilibrium which can be assessed by a chi-square or exact test. Such a test concerns the correlation between the two indicators for the *same* marker (within marker correlation).

Correlation between two *different* markers is referred to as linkage disequilibrium in genetics. If the data is represented by indicator matrices, then linkage disequilibrium can be studied by a canonical correlation analysis of two indicator matrices. Generalized inverses can be used to cope with the singularity of covariance matrices. By using the carrier-indicators as supplementary variables, such a canonical analysis is also informative about Hardy-Weinberg equilibrium. Biplots [2] can be used to visualize the results.

In the light of the larger number of markers obtained in genotyping studies, Carroll's [1] generalized canonical correlation analysis can be used to study multiple markers simultaneously.

The various forms of the canonical analysis of genetic markers will be illustrated with several examples in the talk.

References

[1] Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 67th Annual Convention of the American Psychological Association*, 227–228.

[2] Graffelman, J. (2005). Enriched biplots for canonical correlation analysis. *Journal of Applied Statistics* **32**(2), 173–188.