UNIVERSITY OF SYDNEY

SCHOOL OF MATHEMATICS AND STATISTICS

Statistics Seminar

Friday, 21 April, 2.00pm

Eastern Avenue Lecture Theater

Modelling high-dimensional abundance data using generalised estimating equations

Dr David Warton
University of New South Wales.

Abstract

Multivariate abundance data (where abundances are measured simultaneously for many different types of organism) are commonly collected in the environmental sciences - whether testing for environmental impacts, analysing a controlled experiment, or in exploratory ecological research. Particular difficulties in analysing this type of data are high dimensionality (more variables than observations), overdispersion of abundance counts, and a high incidence of zero abundance counts.

In this talk I will introduce multivariate abundance data, describe typical data properties, outline an efficient approach to data analysis, and some recent innovations. Results that will be discussed include the following: (a) The most common approach to addressing overdispersion in count data - specifying that the variance is proportional to the mean - has a poor fit to most multivariate abundance datasets, and alternatives are necessary; (b) Despite the high incidence of zeros, there is typically no evidence of zero-inflation in model-fitting; (c) Generalised estimating equations (GEEs) can be applied to account for the correlation between variables, although the GEE methodology requires modification to deal with high dimensionality; (d) One particularly promising approach for dealing with high dimensionality is to use regularisation - shrinking the correlation matrix towards the identity, where the shrinkage parameter is estimated by cross-validation; (e) Ecologists are often interested in questions of composition rather than questions of abundance per se, and this can be simply accommodated via reparameterisation of the mean model.

Enquiries about the Statistics Seminar should be directed to
Marc Raimondo (`marcr@maths.usyd.edu.au`)