Multivariate hypothesis testing and implementation in R

Testing for object grouping or treatment effects

Analysis of similarity (ANOSIM), Multi Response Permutation Procedure (MRPP), and PERMANOVA are multivariate versions of Analysis of Variance (ANOVA). ANOSIM compares the distances between samples within *a priori* defined groups to distances between samples across defined groups. However the actual distances are not used, ANOSIM is non-parametric and also is based on ranks of distances rather than the actual distances. The ANOSIM statistic (R) is scaled between -1 and 1 with 1 indicating strong differences in composition. Multiple Response Permutation Procedure (MRPP) is a similar method to ANOSIM, but is not rank based. For both approaches, significance of the statistic is determined by permutation of group labels. PERMANOVA is the most closely related to ANOVA of the three as it is based on a pseudo-F-distribution. PERMANOVA is implemented in R using the function adonis(). Being more closely related to ANOVA, this approach allows for incorporation of more complex model designs.

Overview of direct, canonical, or constrained ordination

Recall that indirect ordination evaluated the inherent structure of multivariate data (objects x descriptors) by reducing its dimensionality. Relationships between potential explanatory variables and the objects or descriptors were identified via correlation after the ordination analysis was completed. In this way, indirect ordination is thought of as a visualization and hypothesis-generating tool. Direct ordination techniques still provide a reduction in dimensionality of data, but it simultaneously evaluates *a priori* hypotheses about the link between objects or descriptors and explanatory variables. Direct ordination is useful because it evaluates the amount of variation in objects or descriptors that can be related to a set of explanatory variables and the significance of these relationships can be evaluated using permutation tests. In many ways direct ordination, especially the two approaches we'll cover in this class (redundancy analysis, RDA, and canonical correspondence analysis, CCA) is a multiple regression of multivariate data.

Redundancy Analysis (RDA)

RDA is essentially the combination of multiple regression and Principle Components Analysis (PCA). Euclidean distance amongst objects is conserved as in PCA, but the ordination axes are constrained by the explanatory variables provided. In other words, each RDA canonical ordination axis corresponds to a direction, in the multivariate scatter of objects, which is maximally related to a linear combination of the explanatory variables. If Y is a matrix of descriptor values for some number of objects and X is a matrix of explanatory variables for those same objects, one can conceptualize RDA as a two-step process. First, each descriptor in Y is regressed against all explanatory variables in X. Second, the fitted values from all of the multiple regressions (call this Y') is subjected to PCA to generate canonical ordination axes. The residuals between Y and Y' (call this Y_{res}) can also be subjected to PCA and this represents an unconstrained ordination of the unexplained variation in Y. Often, the descriptors in Y and variables in X are centered/standardized prior to analysis. The use of PCA in the RDA process makes it somewhat ill-suited for some ecological data (e.g. species composition data), but multiple transformations can be used to make RDA of these data possible.

RDA, and other direct ordination techniques, generate tri-plots! The interpretation of these depends on the scaling applied and can be a bit tricky. When scaled for distance (Type I), distances among objects are approximations of their Euclidean distances. The position of an object along a descriptor approximates the value of the object along that descriptor. The angles among descriptors are meaningless, and angles between explanatory variables and descriptors

represent their correlations. Correlation (Type II) scaling contrasts with the distance-based scaling. Distances amongst objects are not reflective of Euclidean distances. The position of an object along a descriptor AND AN EXPLANATORY VARIABLE vector approximates its value. Angles amongst descriptors and explanatory variables reflect their correlation. Distance-based scaling is most useful when explanatory variables are categorical and correlation-based scaling is most useful when explanatory variables are continuous.

Canonical Correspondence Analysis (CCA)
The math behind CCA is essentially identical to RDA, but the chi-squared distances are conserved rather than Euclidean distances and a weighted multiple regression is used instead of the conventional multiple regression in RDA. One potentially unfortunate side-effect of CCA is that differences amongst rare species can have a disproportionate effect on the CCA ordination. Differences between interpretation of ordinations based upon scaling discussed for RDA also apply for CCA.