4.16 DISTLM versus BEST/ BIOENV

On the face of it, the DISTLM routine might be thought of as playing a similar role to PRIMER's BEST routine in the analysis of multivariate species data. More particularly, the BEST (BIOENV or BVSTEP) procedure in PRIMER is designed to find a combination of environmental variables which, together, result in resemblances among samples whose rank order *best matches* the rank order of the inter-sample resemblances arising from biological (species) data[96]. There are some important differences, however, between the BEST/BIOENV approach and models of a multivariate data cloud obtained using DISTLM.

First, DISTLM actually formally fits a *linear* model of the predictor (environmental) variables to the response (species) data cloud, in the space defined by the chosen resemblance measure. This means that the dissimilarities (or similarities) themselves are important here. They will define the shape and structure of the data cloud, so (as in PERMANOVA), the resemblance measure the user chooses is quite important. Indeed, although our ability to view the structure of this data cloud using unconstrained ordination (such as PCO) is necessarily imperfect, such an ordination should nevertheless be used to provide some information on whether there are any gross outliers, for example, which would make a linear modeling approach inappropriate. DISTLM does not assume anything specific about the shape of the data cloud, and *any* resemblance measure that is deemed appropriate for the nature of the data and hypotheses of interest can be used to construct it, but outliers or "high leverage" points[97] in that space, if present, will tend to have a strong influence on the results.

The advantages of fitting a formal model using DISTLM are fairly clear. First, we achieve a direct quantitative *partitioning* of the multivariate variability that is *explained* by each of several environmental variables. Thus, we can determine *how much* of the variability is attributable to individual predictor variables (either acting alone or in pre-defined sets), and we can determine explicitly how much *overlap* there is in this explained variation. Of course, in order to do this, we have to be explicit about what we mean by "variation", so that is where (and why) the choice of resemblance measure becomes so important.

Of course, DISTLM also has some clear limitations. First, despite the flexibility afforded by being able to choose any resemblance measure we wish as the basis of the analysis (so the models are usually not at all linear with respect to the original **Y** variables in the majority of cases), these models *are* strictly linear in the **X** variables. We can use polynomials of the **X** variables to get around this to some extent, but this is not the only potential issue. Another is that DISTLM's reliance on the traditional partitioning approach means that we can run out of degrees of freedom if there are more predictor variables than there are samples. More particularly, in order to get sensible results, the largest possible full model is restricted to having $q \leq N - 1$, at most. This is a simple consequence of it being possible to perfectly fit a linear model with $(N - 1)$ parameters (variables) to N points ($R^2 = 1.0). Although we can use criteria that are not strictly monotonic on $R^2 with increases in predictor variables (such as adjusted $R^2, *AIC*, *AIC*$_c$ or *BIC*), which will certainly help to find parsimonious models, all of the models fit by DISTLM partition the total variation using a linear function of the X's and so will necessarily have this restriction of an upper bound on q.

In contrast, the BEST/BIOENV procedure arises out of the purely non-parametric approach inherent in the majority of the routines already available in PRIMER, such as non-metric MDS and ANOSIM. The differences between DISTLM and BIOENV are therefore directly analogous to many of the differences between PERMANOVA and ANOSIM already discussed in chapter 1. In essence, the BEST/BIOENV procedure does not attempt to model the data cloud at all, but rather tries to find the best possible *rank-order match* between the inter-point dissimilarities and the inter-point distances derived from sets of environmental variables. The criterion used for this matching is either a Spearman or a Kendall rank correlation, so it is only the rank orders of the resemblances that are being considered. There are several advantages to this approach. First, we can have as many variables as we want in either of the original matrices. The “matching” is being done on the basis of rank resemblances only, so there is simply no limit to how many original variables may be used to calculate either the species resemblances or environmental distance matrices. Second, the rank correlation (whether we use Spearman, weighted Spearman or Kendall) yields a criterion for the success of the match which (unlike R^2) is not monotonically related to the number of variables in the environmental data matrix at all. In fact, the inclusion of variables that do nothing to enhance the match will clearly cause a decrease in rank correlation. This criterion has intuitive appeal for identifying parsimonious sets of environmental variables that produce patterns among samples that are similar to the patterns produced among those same samples using the biotic data. Furthermore, the permutation test associated with the BEST/BIOENV routine includes the selection step with each permutation. This is really rather neat and allows the user validly to test the significance of the relationship between the two matrices given that some effort has gone into selecting environmental variables that will provide a good match.

The limitations of the BEST/BIOENV approach become apparent, however, when we realise that, once a purportedly “useful” set of environmental variables have been selected, we are not in a position to say *how much* of the variation inherent in the species resemblance matrix is “explained” by these variables, either individually or collectively. Such a variance is a function of the precise measurement scale of the resemblances, i.e. is a “metric” concept that cannot be captured by a non-(para)metric approach. The rank correlation between the two resemblance matrices does provide a valuable non-parametric index of how closely the collective set of environmental variables captures the multivariate pattern of the species variables (on a scale from $\rho \approx 0$ to 1), and this is an index with an absolute validity in comparisons across different transformations, resemblance measures, etc. (as with the similarly rank-based ANOSIM R statistic). However, it does not directly provide a quantitative measure of the *relative importance* of the individual environmental variables that have been selected; this can only be inferred by comparing the match (ρ) to the multivariate species cloud for different subsets of these environmental variables. Most tellingly, it cannot provide sequential (partial) tests, i.e. of the statistical significance of adding (or deleting) an explanatory variable from the current set. In other words, by going “non-parametric” (BEST/BIOENV), we relinquish our ability to explicitly measure and model multivariate variation. On the other hand, if we want to create such a model (DISTLM), then we must define *what we mean* by “multivariate variation” and decide *how* we are going to model it. This requires some decisions (e.g., which resemblance measure shall I use?) and some model assumptions (e.g., fitting linear combinations of predictor variables, and that the residual variability is additive and homogeneous across the different levels of the predictor variables). In short, we believe that DISTLM retains much of the flexibility of the non-parametric approach by allowing any (reasonable)data[98] resemblance measure to define what we mean by “multivariate variation”.

However, in order to take the step towards formally modeling this variation, we are forced to let go of the fully non-parametric (and completely assumption-free) setting. Nevertheless, by using permutation procedures for the marginal and sequential tests, these additional assumptions can, however, entirely avoid being distributional.

---

data[96] See chapter 11 in Clarke & Warwick (2001) and chapter 11 in Clarke & Gorley (2006) for more details regarding these routines.

data[97] For a discussion of outliers and high leverage points in multiple regression, see for example Neter, Kutner, Nachtsheim *et al. (1996)* .

data[98] By “reasonable” we generally mean a measure that fulfills at least the first 3 of the 4 criteria of a metric distance measure (see the section Negative eigenvalues in chapter 3) and also one which is meaningful to the researcher for interpretation.

Revision #4
Created 15 August 2022 12:59:20 by Arden
Updated 5 September 2022 14:09:01 by Arden