4.7 Assumptions & diagnostics

Thus far, we have only done examples for a univariate response variable in Euclidean space, using DISTLM to fit linear models, but with tests being done by permutation. However, the fact that any resemblance measure can be used as the basis of the analysis in dbRDA yields considerable flexibility in terms of modeling. In traditional regression and RDA, the fitted values are a linear combination of the variables in **X**. So, the relationship between the multivariate data **Y** and predictor variables **X** is assumed to be linear for the purposes of modeling. In dbRDA, however, the relationship being modeled between **Y** and **X** variables is more complex and depends on the chosen resemblance measure. Another way to describe dbRDA is that it is a traditional RDA done on the PCO axes[83] from a resemblance matrix, rather than on the original **Y** variables. Thus, in dbRDA, we effectively assume a linear relationship between the PCO axes derived from the resemblance matrix and the variables in **X** for purposes of modeling (Legendre & Anderson (1999)). In many cases, this is quite appropriate provided the resemblance measure used as the basis of the analysis is a sensible one for the data. By “sensible”, we mean that the resemblance measure describes multivariate variation in a way that emphasises the features of the data that are of interest (e.g., changes in composition, relative abundance, etc.) for specific hypotheses postulated by the user. Note, for example, that if one were to perform dbRDA on a chi-squared distance matrix, then this will assume unimodal relationships between **Y** and **X**, as is done in *canonical correspondence analysis* (CCA, ter Braak (1986a) , ter Braak (1986b))[84]. Clearly dbRDA goes beyond what either RDA or CCA can provide, by allowing any resemblance measure (e.g., Bray-Curtis, Manhattan, etc.) to define multivariate variation. In addition, once a given resemblance measure has been chosen, many other kinds of non-linear relationships between the PCO axes and **X** can be modeled by introducing polynomials of the variables in **X**, if desired (e.g., Makarenkov & Legendre (2002)).

Although dbRDA does provide quite impressive flexibility with respect to the response variables (**Y**), it pays to spend some time with the **X** variables to examine their distributions and the relationships among them, as these are being treated in the same way in dbRDA as they would for any linear multiple regression model. Although DISTLM does not make any explicit assumptions about the distributions of the **X** variables, they should nevertheless be reasonable for purposes of linear modeling – they should not be heavily skewed or contain extreme outliers. It is a very good idea, therefore, to examine the **X** variables using a draftsman plot in PRIMER and to transform them (individually if necessary) to avoid skewness before proceeding.

The issue of *multi-collinearity* – strong inter-correlations among the **X** variables – is also something to watch out for in dbRDA, as it is for RDA or multiple regression (e.g., Neter, Kutner, Nachtsheim *et al. (1996)*). If two variables in **X** are very highly co-linear (with correlation $|r| \geq 0.95$, for example), then they contain effectively the same information and are redundant for purposes of the analysis. A redundant variable should be dropped before proceeding (keeping in mind that the variable which is retained for modeling may of course simply be acting as a proxy for the one that was dropped). Once again, PRIMER’s **Draftsman Plot** tool will provide useful direct information

about multi-collinearity among the variables in **X**. See pp. 122-123 of chapter 11 in Clarke & Gorley (2006) , for example, which demonstrate how to use the **Draftsman Plot** tool to identify skewness and multi-collinearity for a set of environmental variables.

In traditional multiple regression, the errors are assumed to be independent and identically distributed (i.i.d.) as normal random variables. DISTLM uses permutations to test hypotheses, however, so normality is not assumed. For a permutation test in regression, if we consider that the null hypothesis is true and **Y** and **X** are not related to one another, then the matching of a particular row of **Y** (where rows identify the samples, as in Fig. 4.2) with a particular row of **X** does not matter, and we can order the 1 to *N* rows of **Y** (or, equivalently, the rows and columns of matrix **D**) in any way we wish (e.g., Manly (2006)). Thus, all that is assumed is that the sample rows are *exchangeable* under a true null hypothesis. For conditional tests, we assume that the residuals obtained after fitting covariates are exchangeable under a true null hypothesis. This means, more particularly, that we assume that the linear model being used to fit the covariate(s) to multivariate data in the space of the resemblance measure is appropriate and that the errors (estimated by the residuals from this model) have homogeneous dispersions in that space, so they are exchangeable.

---

[83] Being careful, that is, to do all computations on the axes associated with positive and negative eigenvalues separately, and combining them only when they are squared (e.g., as sums of squares). Axes corresponding to the negative eigenvalues contribute *negatively* in the squared terms. See chapter 3 regarding negative eigenvalues and McArdle & Anderson (2001) for more details.

[84] If the user performs dbRDA on the basis of the chi-squared distance measure, the results will produce patterns highly similar to those obtained using CCA. Any differences will be due to the intrinsic weights used in CCA. See ter Braak (1987) , Legendre & Legendre (1998) and Legendre & Gallagher (2001) for details regarding the differences between CCA and RDA and the algebraic formulations for these two approaches.

Revision #7
Created 11 August 2022 16:52:58 by Arden
Updated 5 September 2022 12:53:07 by Arden