5.14 CAP versus dbRDA

So, how does CAP differ from dbRDA for relating two sets of variables? First, dbRDA is directional. Each set of variables has a role as either predictor variables (**X**) or response variables (**Q**), while for CAP (when there are multiple variables in **X**), the two sets of variables are essentially treated *symmetrically*. Canonical correlation analysis (on the basis of Euclidean distances) finds linear combinations of **Y** and linear combinations of **X** that are maximally correlated with one another. In contrast, RDA finds linear combinations of **X** that are best at explaining or predicting **Y** (the latter set of variables are *not* sphericised). Note that canonical correlation sphericises both of the data clouds using **X**^{0} and **Y**^{0} in the calculations. This ensures that the correlations among the variables *within* either set are taken into account. In contrast, RDA uses sphericised **X**^{0} (by constructing matrix **H**), but the variables in **Y** are *not* sphericised. Now, CAP generalises canonical correlation to any resemblance measure by replacing **Y**^{0} with **Q**$^0 _m$, while dbRDA generalises RDA to any resemblance measure by replacing **Y** with **Q**. Furthermore, no matter how many variables occurred in the original **Y** matrix, over-parameterisation in CAP can be avoided by a prudent choice for *m* (generally obtained using diagnostics).

The decision of which method to use should always be based on the conceptual goals of the analysis. When there are multiple **X** variables, CAP can be helpful for exploring relationships between these variables and a multivariate data cloud expressed by a resemblance matrix. In contrast, dbRDA is more appropriate when one wishes to explicitly model the variability in the multivariate data cloud using a set of **X** predictor variables. Table 5.1 provides a summary of the differences between these two approaches.

Table 5.1. Summary of differences between CAP and dbRDA for the analysis of the relationships between two sets of variables. Matrices are defined in the text and are also defined in the index of mathematical notation.

	CAP	*dbRDA*
Analysis of:	$\mathbf{Q}_m^0$ and $\mathbf{X}^0$	$\mathbf{Q}$ and $\mathbf{X}^0$
Standardisation:	PCO axes ($\mathbf{Q}^0$) are unit-normed (sphericised) to SSCP = $\mathbf{I}$. $\mathbf{X}$ variables are sphericised.	PCO axes ($\mathbf{Q}$) are standardised to SSCP = Λ. $\mathbf{X}$ variables are sphericised.
Roles of data clouds:	Symmetric: $\mathbf{Q}_m^0 \leftrightarrow \mathbf{X}^0$	Directional: Project $\mathbf{Q}$ (response data cloud) onto $\mathbf{X}^0$ (predictors)
Avoiding over-parametersation:	A subset of m PCO axes are used; m chosen using 'leave-one-out' diagnostics. All $\mathbf{X}$ variables are used.	All PCO axes are used. A subset of $\mathbf{X}$ variables can be chosen (optionally) using model selection criteria.
Purpose:	If $q > 1$, explore correlations between two data clouds. If $q = 1$, $\mathbf{Q}_m^0$ can be used to predict X, a single gradient.	Use $\mathbf{X}$ to explicitly model, explain or predict variation in $\mathbf{Q}$.
Test-statistic:	$tr(\mathbf{Q}_m^{0\prime}\mathbf{H}\mathbf{Q}_m^0)$	$tr(\mathbf{HQH})$ (best to calculate directly as $tr(\mathbf{HGH})$ for when there are negative eigenvalues)
Interpretation of eigenvalues:	Squared canonical correlation (δ^2)	A portion of the explained variation (γ^2), which can be expressed as a proportion of the total explained variation ($\Sigma\gamma^2$).
Interpretation of axes:	A linear combination of $\mathbf{Q}_m^0$ variables having maximum correlation with a linear combination of $\mathbf{X}$ variables.	A linear combination of $\mathbf{X}$ variables that explains the greatest amount of variation in the response data cloud $\mathbf{Q}$.
Default for vector overlay:	Linear combinations of $\mathbf{X}^0$ having maximum correlation with CAP axes.	Direct projection of $\mathbf{X}^0$ onto dbRDA axes.

Revision #2
Created 17 August 2022 10:51:01 by Arden
Updated 11 December 2024 03:11:28 by Abby Miller