Ecological Archives **E089-105-A2**

Michael W. Palmer, Daniel J. McGlinn, Lars Westerberg, and Per Milberg. 2008. Indices for detecting differences in species composition: some simplifications of RDA and CCA. *Ecology* 89:1769–1771.

Appendix B: Simplification of CCA eigenvalue.

Notation (modified after Legendre and Legendre 1998):

As with Appendix A, we assume for convenience that we are dealing with fixed plots sampled twice, although the results pertain to split plot designs and balanced independent samples with no loss of generality.

Samples are indexed by the letter i: 1, 2, …n

Pairs of samples or 'plots' are indexed by the letter k: 1, 2, …n/2

Species are indexed by the letter j: 1, 2, …r

Elements y_{ij} of the response matrix $\mathbf{Y}$ $(n \times r)$ denote the abundance in sample i (rows) of species j (columns)

y_{+j} = sample totals (sum across rows to get column totals)

y_{i+} = species totals (sum across columns to get row totals)

y_{++} = grand total for sample by species matrix

The explanatory variable **x** is a column vector of length n and is a dummy variable where x_i equals 0 or 1 for times 1 and 2, respectively.

The response matrix **Y** is divided into two $\frac{n}{2} \times r$ submatrices **G** and **H**, where **G** contains the samples corresponding to $x_i = 0$ and where **H** contains the samples corresponding to $x_i = 1$.

p is the proportion of the total abundance in samples where $x_i = 1$, i.e. $p = h_{++} / y_{++}$.

S stands for a dispersion matrix except without dividing by degrees of freedom (see Legendre and Legendre 1998, p596).

D stands for a diagonal matrix of species totals, y_{i+}

Proof:

The CCA eigenvector equation (Legendre and Legendre 1998, p596) is:

$$(\mathbf{S}_{\overline{\mathbf{Q}}'\mathbf{D}^{1/2}\mathbf{x}^*}\mathbf{S}^{-1}_{\mathbf{x}^{*\prime}\mathbf{D}\mathbf{x}^*}\mathbf{S}_{\mathbf{x}^{*\prime}\mathbf{D}^{1/2}\overline{\mathbf{Q}}} - \lambda_a\mathbf{I})\mathbf{u}_a = 0 \qquad \text{(B.1)}$$

This equation is similar to equation A.1 for the RDA test statistic, but important differences exist in how the component matrices are scaled and weighted. First, **Y** must be scaled by its contribution to χ^2 (Legendre and Legendre 1998) into:

$$\overline{\mathbf{Q}} = \left[\overline{q}_{ij}\right] = \frac{y_{ij}y_{+,+} - y_{i+}y_{+j}}{y_{++}\sqrt{y_{i+}y_{+j}}} \qquad \text{(B.2)}$$

Without loss of generality, we define the sum of all abundances as 1, i.e. $y_{++} = 1$. Then:

$$\overline{\mathbf{Q}} = \left[\overline{q}_{ij}\right] = \left[\frac{y_{ij} - y_{i+}y_{+j}}{\sqrt{y_{i+}y_{+j}}}\right] \qquad \text{(B.3)}$$

The explanatory variable must also be transformed into a weighted standardized variable by subtracting the weighted average and dividing by the maximum likelihood estimator of the standard deviation:

$$\overline{x} = \frac{\sum_i y_{i+}x_i}{y_{++}} = \frac{\sum_i (g_{i+})(0) + \sum_i (h_{i+})(1)}{y_{++}} = \frac{h_{++}}{y_{++}} = \frac{h_{++}}{1} = p \qquad \text{(B.4)}$$

$$\text{var}(\mathbf{x}) = \sigma^2_{\mathbf{x}} = \frac{\sum_i \left[y_{i+}(x_i - \overline{x})^2\right]}{y_{++}} = \frac{\sum_i \left[g_{i+}(\overline{x}^2)\right] + \sum_i \left[h_{i+}(1-\overline{x})^2\right]}{1} \qquad \text{(B.5)}$$

$$\text{var}(\mathbf{x}) = (p)^2\sum_i g_{i+} + (1-p)^2\sum_i h_{i+} = (p)^2(1-p) + (1-p)^2\,p = p(1-p) \qquad \text{(B.6)}$$

$$\mathrm{sd}(\mathbf{x}) = \sigma_{\mathbf{x}} = \sqrt{p(1-p)} \tag{B.7}$$

Therefore the standardized value $\mathbf{x}^*$ is:

$$x_i^* = \frac{x_i - \bar{x}}{\sigma_{\mathbf{x}}} = \begin{cases} \dfrac{-p}{\sqrt{p(1-p)}}, x_i = 0 \\ \dfrac{1-p}{\sqrt{p(1-p)}}, x_i = 1 \end{cases} \tag{B.8}$$

Also because CCA relies upon weighted multiple regression instead of conventional multiple regression as in RDA, the square root of a diagonal matrix of weights **D** must be applied to $\mathbf{x}^*$ everywhere it occurs in B.1 (Legendre and Legendre 1998, p. 595). Therefore the inverse of the variance of the weighted standardized **x** matrix is a scalar equal to:

$$\mathbf{S}^{-1}_{\mathbf{x}^*\mathbf{'D}\mathbf{x}^*} = \frac{1}{\sigma^2_{\mathbf{x}^*}} = \frac{1}{1} = 1 \tag{B.9}$$

Furthermore, in our case, $\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{'D}^{1/2}\mathbf{x}^*}$ and $\mathbf{S}_{\mathbf{x}^*\mathbf{'D}^{1/2}\overline{\mathbf{Q}}}$ are column and row vectors respectively. The *j* elements of $\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{'D}^{1/2}\mathbf{x}^*}$ are equal to the *j* elements of $\mathbf{S}_{\mathbf{x}^*\mathbf{'D}^{1/2}\overline{\mathbf{Q}}}$, i.e.:

$$(\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{'D}^{1/2}\mathbf{x}^*})' = \mathbf{S}_{\mathbf{x}^*\mathbf{'D}^{1/2}\overline{\mathbf{Q}}} = \sum_i [\overline{q}_{ij}(x_i^* \sqrt{y_{i+}})] \text{ where } j = 1, 2, \ldots r \tag{B.10}$$

Expression B.10 simplifies further by expanding its components and partitioning matrix **Y** into the two sets of independent samples: **G** and **H**. Inserting equation B.3 and simplifying yields:

$$\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{'D}^{1/2}\mathbf{x}^*} = \sum_i \left[\frac{y_{ij} - y_{i+}y_{+j}}{\sqrt{y_{i+}y_{+j}}} \cdot \left(x_i^* \sqrt{y_{i+}} \right) \right] = \sum_i \left[\frac{y_{ij} - y_{i+}y_{+j}}{\sqrt{y_{+j}}} \cdot x_i^* \right] \tag{B.11}$$

$$\mathbf{S}_{\overline{\mathbf{Q}}'\mathbf{D}^{1/2}\mathbf{x}^*} = \frac{\sum_i (y_{ij} - y_{i+}y_{+j}) \cdot x_i^*}{\sqrt{y_{+j}}} \tag{B.12}$$

Simplifying the numerator and partitioning **Y** gives:

$$\sum_i (y_{ij} - y_{i+}y_{+j}) \cdot x_i^* = \sum_k (g_{kj} - g_{k+}y_{+j}) \cdot \left[\frac{-p}{\sqrt{p(1-p)}}\right] + \sum_k (h_{kj} - h_{k+}y_{+j}) \cdot \left[\frac{1-p}{\sqrt{p(1-p)}}\right] \tag{B.13}$$

$$= \frac{\sum_k (g_{kj} - g_{k+}y_{+j}) \cdot (-p) + \sum_k (h_{kj} - h_{k+}y_{+j}) \cdot (1-p)}{\sqrt{p(1-p)}} \tag{B.14}$$

$$= \frac{\sum_k \left\{(h_{kj} - h_{k+}y_{+j}) - p\left[(g_{kj} - g_{k+}y_{+j}) + (h_{kj} - h_{k+}y_{+j})\right]\right\}}{\sqrt{p(1-p)}} \tag{B.15}$$

$$= \frac{\sum_k h_{kj} - y_{+j}\sum_k h_{k+} - p\left[\sum_k g_{kj} - y_{+j}\sum_k g_{k+} + \sum_k h_{kj} - y_{+j}\sum_k h_{k+}\right]}{\sqrt{p(1-p)}} \tag{B.16}$$

$$= \frac{h_{+j} - y_{+j}h_{++} - p\left[g_{+j} + h_{+j} - y_{+j}(g_{++} + h_{++})\right]}{\sqrt{p(1-p)}} \tag{B.17}$$

$$= \frac{h_{+j} - y_{+j}h_{++} + y_{+j}h_{++} - p(g_{+j} + h_{+j})}{\sqrt{p(1-p)}} \tag{B.18}$$

$$= \frac{h_{+j} - p(g_{+j} + h_{+j})}{\sqrt{p(1-p)}} = \frac{h_{+j} - py_{+j}}{\sqrt{p(1-p)}} \tag{B.19}$$

Combining equation B.12 and B.19 gives:

$$\mathbf{S}_{\overline{\mathbf{Q}}'\mathbf{D}^{1/2}\mathbf{x}^*} = \frac{h_{+j} - py_{+j}}{\sqrt{p(1-p)y_{+j}}} \tag{B.20}$$

The CCA eigenvector equation (B.1) becomes $\lambda = \mathbf{S}_{\mathbf{x}^{*\prime}\mathbf{D}^{1/2}\overline{\mathbf{Q}}}\mathbf{S}_{\overline{\mathbf{Q}}'\mathbf{D}^{1/2}\mathbf{x}^*}\mathbf{S}^{-1}_{\mathbf{x}^{*\prime}\mathbf{D}\mathbf{x}^*}$ after rearranging (as in appendix A) and combining this expression with B.9 and B.20 we see that:

$$\lambda = \sum_j \left\{ \frac{h_{+j} - p y_{+j}}{\sqrt{p(1-p) y_{+j}}} \right\}^2 = \frac{1}{p(1-p)} \sum_j \frac{(h_{+j} - p y_{+j})^2}{y_{+j}} \qquad \text{(B.21)}$$

To express B.21 in terms of 'raw' y, we divide each y-term by the grand total, i.e. y_{++}:

$$\lambda = \frac{1}{p(1-p)} \sum_j \frac{\left\{ \frac{1}{y_{++}} h_{+j} - p \frac{1}{y_{++}} y_{+j} \right\}^2}{\frac{1}{y_{++}} y_{+j}} = \frac{1}{y_{++} p(1-p)} \sum_j \frac{(h_{+j} - p y_{+j})^2}{y_{+j}} \qquad \text{(B.22)}$$

LITERATURE CITED

Legendre, P., and L. Legendre. 1998. Numerical Ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.