DETRENDED CORRESPONDENCE ANALYSIS (DCA)

Steven M. Holland

Department of Geology, University of Georgia, Athens, GA 30602-2501

May 2008
revised December 2019

Introduction

Correspondence analysis (CA) was first introduced to ecology in the early 1970s under the name of reciprocal averaging (RA). It quickly gained popularity because of its better recovery of simulated one-dimensional gradients, compared with principal components analysis. Correspondence analysis was also more attractive than PCA on theoretical grounds, because it allowed for realistic modal responses of species along environmental gradients, where a species might be most abundant at one position along the gradient and decline in abundance in either direction along the gradient. This was in contrast to the linear relationship assumed by PCA, where a species was assumed to increase in abundance steadily in one direction along a gradient.

Correspondence analysis had a flaw, though: it tended to distort one-dimensional gradients into an arch on the second analysis. As a result, positions of samples along the gradient tended to become more compressed at the end points of the gradient. Detrended correspondence analysis (DCA) was developed to overcome these problems by flattening the arch and rescaling the positions of samples along an axis. Simulations showed that DCA performed markedly better than CA/RA and PCA.

Computation

Correspondence analysis / reciprocal averaging can be calculated with an eigenanalysis approach and with a reciprocal averaging approach. The latter is more intuitive, so the reciprocal averaging approach will be presented here. Beginning with a matrix of n rows of samples and p columns of taxa, each sample is assigned an arbitrarily chosen score (x_i). Scores for each taxon (y_j) are calculated as a weighted average, where the abundance of a taxon (a_{ij}) is multiplied by the sample score, and these are summed across all samples and divided by the total abundance for that taxon.

$$y_j = \frac{\sum_{i=1}^{n} a_{ij} x_i}{a_{+j}}$$

These taxon scores are used to calculate a new set of sample scores following the same procedure, in which the abundance of a taxon is multiplied by the taxon score. These are summed across all taxa and divided by the total abundance within each sample.

$$x_i = \frac{\sum_{j=1}^{p} a_{ij} y_j}{a_{i+}}$$

Sample scores are centered and standardized such that their mean is zero and their variance is one.

$$\sum_{i=1}^{n} a_{i+} x_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} a_{i+} x_i^2 = 1$$

This procedure of alternately calculating sample and taxon scores is repeated until the scores stabilize. This initial pass produces the CA axis 1 scores for both samples and taxa. The same procedure is performed to produce higher-order axes orthogonal (uncorrelated) to lower-order axes.

Detrended correspondence analysis begins with a correspondence analysis, but follows it with steps to detrend and rescale axes. In the simplest approach to detrending, the axis is first divided into an arbitrary number of equal-length segments. Within each segment, the scores on the next higher-order axis are recentered such that their mean is zero. In effect, if an arch is present, it is flattened onto the lower-order axis. The detrending process is sensitive to the number of segments and the default value is 26, which has empirically produced acceptable results. Others have proposed using polynomial regressions for rescaling and others have used a sliding moving average window, which is what the algorithm in R does. The rescaling of an axis is accomplished by equalizing the weighted variance of taxon scores along the axis segments.

Caution

Detrended correspondence analysis has come under intense criticism from some ecologists, and its application now is less common, even though many ecologists report that it performs better than other methods. The principal criticism is that DCA may not fully eliminate distortions of the underlying data structure, and that it can sometimes introduce new distortions (Kenkel and Orloci 1986; Minchin 1987). The most common distortion of a two-dimensional gradient can be envisioned by taking a piece of paper and twisting it, such that axis 1 is preserved reasonably undistorted, but the second axis of variation is expressed on DCA axis 2 at one end, and on DCA axis 3 at the opposite end. This produces what has been called the DCA wedge, consisting of a tapering of sample points in axis 1-2 space and an opposing wedge in axis 1-3 space.

Other criticisms of DCA stemmed from the brute-force inelegance of the method and that it assumes a chi-square distance measure that can overemphasize the importance of rare species. In large samples with multiple rare species, which may occur solely because the sample is large, the chi-square distance can exaggerate the how distinctive the sample is.

A crucial test of any ordination method is how it performs, that is, how well it recovers ecological gradients. Several ecological studies have simulated two-dimensional gradients and subjected them to a variety of ordination methods to compare their results (Kenkel and Orlóci 1986; Minchin 1987). Principal components analysis and correspondence analysis typically display severe distortions of a horseshoe and an arch, respectively, so they should be used in ecological data only with considerable caution. Detrended correspondence analysis and non-metric multidimensional scaling (NMS) commonly perform much better and display minimal distortions in most cases. Because DCA produced stronger distortions than NMS in some of these early simulations, ecologists embraced NMS as the method of choice. However, more exhaustive similar explorations of the success of DCA and NMS not only

revealed similarly common distortions in NMS, they also showed either DCA or NMS might perform better in a particular situation, with about an equal probability of success (Patzkowsky and Holland 2012). Moreover, several authors have reported from field data that DCA ordinations are more interpretable than those from NMS. The stronger condemnations of DCA that have appeared in the ecological literature may be overstated, and I encourage everyone to try both DCA and NMS to evaluate which produces more interpretable results for any given data set.

CA and DCA in R

In the code below, R commands are indicated with a **`bold monospaced font`**. The second and subsequent lines of code of a multiline command are indented. R output is shown in `gray`.

Correspondence analysis and detrended correspondence analysis can each be run using the **vegan** package (https://cran.r-project.org/web/packages/vegan), written by Jari Oksanen and colleagues. As such, the first step is to install the **vegan** library, a one-time operation, except for installing updates.

1) The **vegan** library will need to be loaded whenever its commands are needed; here, it is needed to perform two data transformations, run the detrended correspondence analysis and other ordinations, and compare the configurations of two ordinations.

```
library(vegan)
```

2) This demonstration will use a data set on Late Ordovician benthic marine macroinvertebrate communities from central Kentucky (Holland and Patzkowsky 2004), available on the GEOL 8370 website (http://strata.uga.edu/8370/data/). The data set consists of three files: the first contains the counts of species within samples, the second contains additional properties of each sample (such as age, depositional environment, etc.), and the third contains the properties of the species (such as phylum, class, order).

```
ky <- read.table('KentuckyCounts.txt', header=TRUE,
  row.names=1, sep=',')

sampleProperties <- read.table('KentuckySampleProperties.txt',
  header=TRUE, row.names=1, sep=',')

speciesProperties <- read.table('KentuckyTaxonProperties.txt',
  header=TRUE, row.names=1, sep=',')
```

3) Because the counts table contains raw counts of species within samples, two aspects of the data must be corrected, because they can seriously distort the ordination. The first is sample size, because each sample contains a different number of individuals. When sample size varies substantially, the ordination may reflect that instead of the differences in species composition among the samples. To address this, apply a **percent data transformation** to every row, which converts each abundance to a percentage of that sample, using the **`decostand()`**

function, specifying that the transformation is based on **total**, that is, the total abundance in each row:

```
kyTotal <- decostand(ky, 'total')
```

This transformation would not be applied if one wanted to preserve total abundance, for example, if equally sized areas were counted. For fossil data, one would also need to know that preservation and taphonomy were the same across all the samples, such that total abundance was not caused by differences in preservation.

Next, it is often desirable (although not always) to give the species equal weight, where the changes in abundance among samples for a rare species will have the same importance as the changes in abundance for an abundant species. If this correction is not made, abundant species will have a stronger effect on the ordination than rare species. This requires a **percent maximum data transformation**, which is accomplished by using **decostand()** to standard species abundances by their maximum (**max**) value.

```
kyMax <- decostand(kyTotal, 'max')
```

This transformation would not be applied if one desired that the ordination would be dominated by the abundant species, with progressively rarer species having diminishing influence. I find that ordinations are sometimes more interpretable if this second data transformation is not applied; the first (percent) transformation should always be applied.

4) The detrended correspondence analysis is performed on the doubly-transformed data using the **decostand()** function.

```
kyDCA <- decorana(kyMax)
```

Usually, the defaults of **decorana()** are fine. One option that helps in some cases is to turn on downweighting of rare species with the **iweigh** argument, which can alleviate problems caused by the chi-square distance metric.

```
kyDCAdownweighted <- decorana(kyMax, iweigh=1)
```

Reciprocal averaging / correspondence analysis can be performed by setting the **ira** argument to 1; the default of 0 is for detrended correspondence analysis.

```
kyRA <- decorana(kyMax, ira=1)
```

Other arguments are available, such as for the number of segments in rescaling, but changing these is not recommended without a clear understanding of their effects.

5) The object returned by any of these methods belongs to the class "decorana", and what it contains can be viewed with **names()**.

```
names(kyDCA)
```

The most useful objects it contains are **rproj** and **cproj**, which contain the sample scores (**rproj**) and species scores (**cproj**) for the first four axes. These are the only axes that are calculated.

6) The **`summary()`** function displays the function call, eigenvalues, axis lengths, species scores, and sample scores.

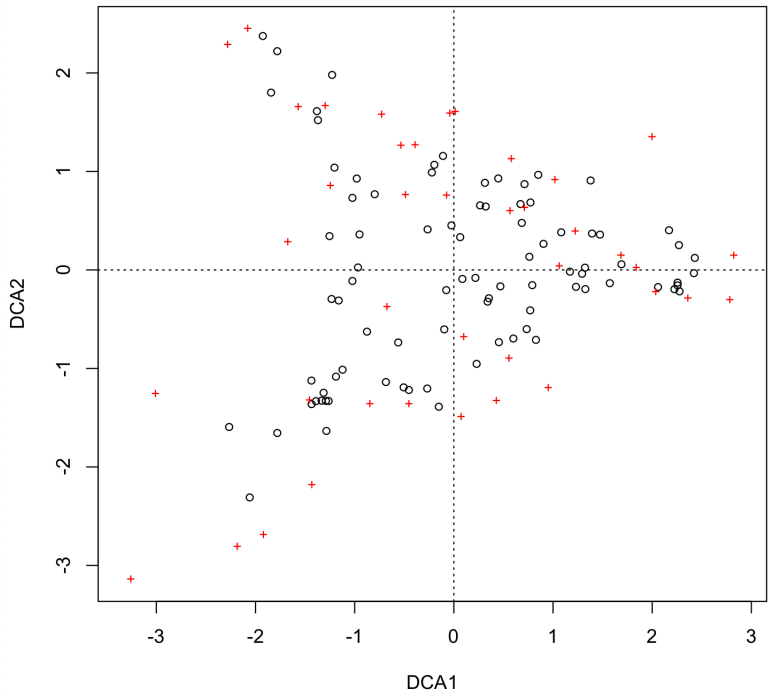
```
summary(kyDCA)
```

Because detrending and rescaling deform the original correspondence analysis configuration, the standard interpretation of eigenvalues in terms of explained variance is destroyed. Consequently, these eigenvalues should not be used or reported.

The axis lengths are useful, because they describe the amount of species turnover on each azis, indicating their relative importance. In general, axis lengths tend to get shorter from axis 1 to axis 4, as each axis explains progressively less variation. In this example, axes 1 and 2 each have lengths about 4.7, suggesting that they reflect similar amounts of species turnover, whereas axes 3 and 4 have lengths of less than 3.9, indicating that they reflect substantially less species turnover. As a result, our analyses will focus on axes 1 and 2.

7) A simple plot of species and sample scores can be made with **`plot()`**.

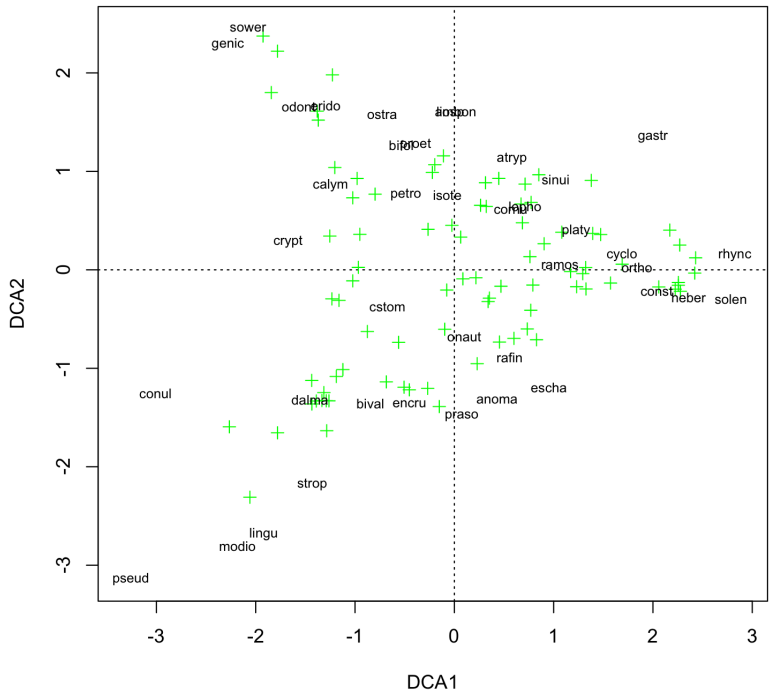
```
plot(kyDCA)
```



Samples are shown by black open circles, and species are indicated by red crosses. This isn't particularly useful because we'd like to know which samples and which species are indicated by the symbols.

The **points()** and **text()** commands allow for some customization of this plot.

```
plot(kyDCA, display=c('none'), cols=c(1, 2))
points(kyDCA, display=c('sites'), choices=1:2, pch=3,
  col='green')
text(kyDCA, display=c('species'), choices=1:2, cex=0.7)
```



This example shows samples with green crosses and species with their 5-letter codes (the column names in the data). Rather than use the decorana-specific plotting functions, I find that the basic plotting commands give me greater flexibility with a more familiar syntax than specifying **display** and **choices**.

8) To make richer plots or to perform other analyses, it is generally useful to extract the species and sample scores from the decorana object. The **scores()** function makes this simple; one just specifies whether scores are needed for species or samples (sites), as well as which axes are desired.

```
kySpecies <- scores(kyDCA, display=c('species'), choices=c(1,
  2))
kySamples <- scores(kyDCA, display=c('sites'), choices=c(1,
  2))
```

Alternatively, one can extract the scores directly from **rproj** and **cproj**.

```
kySpecies <- kyDCA$cproj[, 1:2]
kySamples <- kyDCA$rproj[, 1:2]
```

9) Plots of scores are often most useful when points (samples or species) can be coded by external data. For example, samples could be coded by facies or age. Species might be coded by higher taxon or life habit. In both cases, we are trying to see what controls where species or samples plot in the ordination space.

Let's tackle coding the samples by lithofacies first, as this will give us a sense of whether different communities are found in different depositional environments. If that variation exists, we can see whether it is expressed along axis 1 (the principal source of variation in community composition) or on a higher-order axis (a less important source of variation).

There are many ways to make this plot. For simplicity in function calls, extract vectors to hold the axis 1 and axis 2 sample scores, and a vector holding the lithofacies for each of the samples. It will also be useful to know how many different facies there are.

```
dca1 <- kySamples[, 1]
dca2 <- kySamples[, 2]
facies <- sampleProperties$Lithofacies
numFacies <- length(levels(facies))
```

To color-code the samples based on the facies they occur in, we find what facies are present using **`levels()`**. Based on that output, create a vector of colors for each facies (i.e., Deep Subtidal will be blue, Offshore will be black, and so on).

```
levels(facies)
```

```
[1] "Deep Subtidal"    "Offshore"         "Sand Shoal"
"Shallow Subtidal"
```

```
colors <- c('blue', 'black', 'red', 'orange')
```

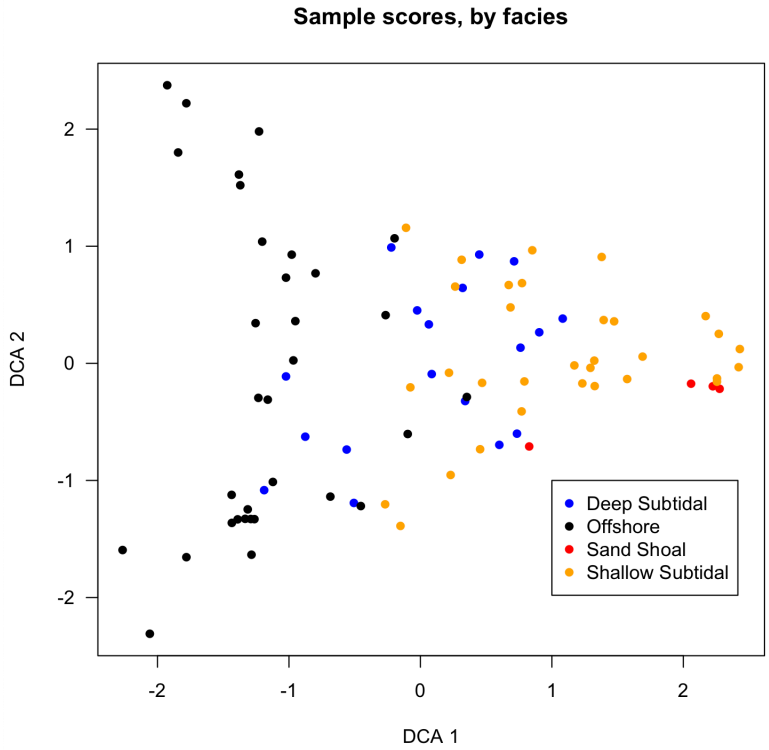
Finally, create an empty plot and then loop through the facies, adding the points for each, finishing by adding a legend.

```
plot(dca1, dca2, type='n', xlab='DCA 1', ylab='DCA 2',
   las=1, main='Sample scores, by facies')

for (i in 1:numFacies) {
  faciesPoints <- facies==levels(facies)[i]
  points(dca1[faciesPoints], dca2[faciesPoints],
    col=colors[i], pch=16)
}

legend(1, -1, levels(facies), pch=16, col=colors)
```

If you are not comfortable with loops, this could also be done using one of the **`apply()`**-family of functions or by calling points individually for each facies.



This plot demonstrates that the samples sort by facies, with overlap, with deeper-water facies to the left and shallower-water facies. Moreover, it is the most important source of variation in community composition because the sorting occurs along axis 1 and not axis 2 or a higher-order axis. One could code the samples by other variables, such as age, to see how those sources of variation are expressed. One could also make similar plots for axes 3 and 4.

10) Species scores can be similarly explored by plotting them coded by external data. For example, perhaps we are interested in where higher taxa plot, such as whether trilobites or brachiopods tend to occur together, or whether they occur in particular facies (as expressed along axis 1, as we just showed). To do so, follow an approach similar to what was done for samples.

First, create vectors of species scores and phylum, and count the number of phyla.

```
dca1 <- kySpecies[, 1]
dca2 <- kySpecies[, 2]
phylum <- speciesProperties$Phylum
numPhyla <- length(levels(phylum))
```

Next, create a vector of colors, one for each phylum.

```
levels(phylum)
colors <- c('black', 'blue', 'lightblue', 'green', 'orange',
  'red', 'brown', 'purple')
```

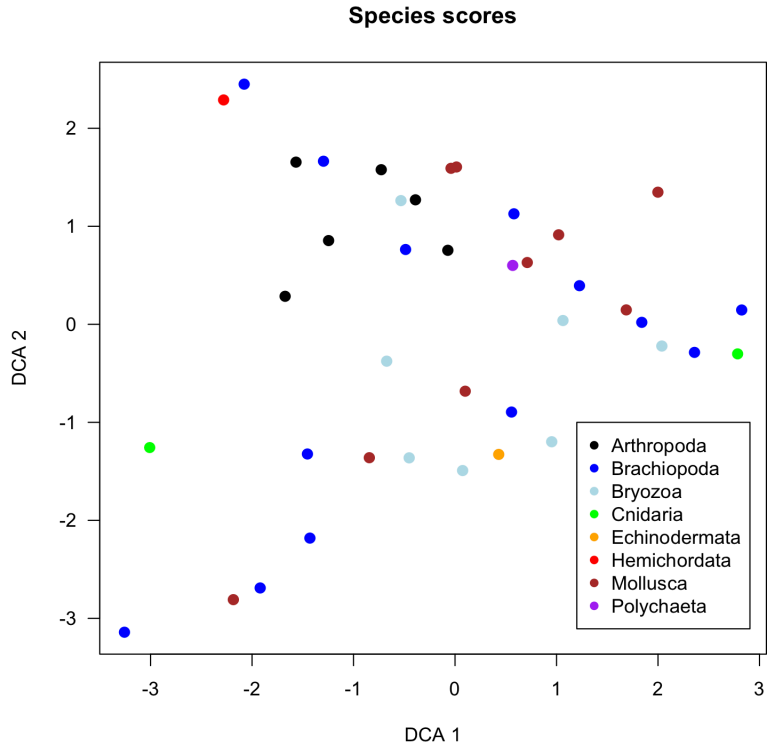
Last, make an empty plot and add the points for each phylum within a loop, finishing with a legend. With so many colors, they will be easier to distinguish if the points are made larger with **cex**.

```
plot(dca1, dca2, type='n', xlab='DCA 1', ylab='DCA 2', las=1,
  main='Species scores')

for (i in 1:numPhyla) {
   phylumPoints <- phylum==levels(phylum)[i]

   points(dca1[phylumPoints], dca2[phylumPoints],
     col=colors[i], pch=16, cex=1.3)
}

legend(1.2, -1, levels(phylum), pch=16, col=colors)
```



This plot shows several things. Brachiopods occur throughout axis 1 (facies) and axis 2, but bryozoans and mollusks tend to be at higher axis 1 scores (shallower-water facies). Trilobites tend to occur at lower axis 1 scores (deeper-water facies) and at higher axis 2 scores. Higher axis 2 scores correspond to younger strata (see Holland and Patzkowsky 2004). Several of the phyla are represented by too few taxa to see any patterns: cnidarians, echinoderms, hemichordates, and polychaetes.

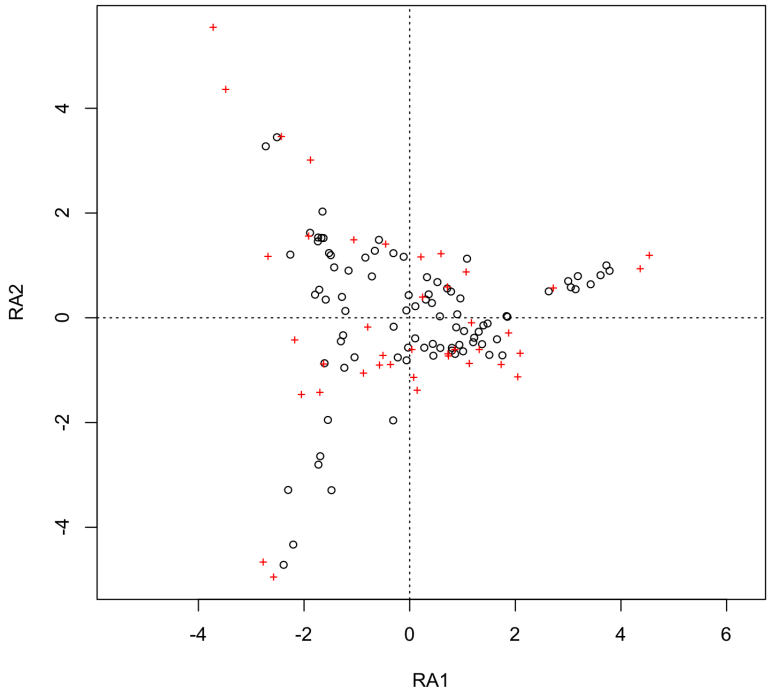
Through a combination of plots of species and sample scores, all coded by external data, it is possible to understand the underlying variables that control the species compositions of samples.

11) When different types of ordinations are applied to the same data, we want to know how they affect the configuration of points. For example, we might want to contrast a correspondence analysis with the detrended version. We might want to know the effects of a particular data transformation on an ordination. We might want to compare DCA to NMS. A technique called procrustes analysis will let us understand how two ordinations compare. All that is needed are the sample scores for both ordinations.

Let's explore the first example, comparing a DCA to a plain correspondence analysis. First, run the correspondence analysis and extract the sample scores. Run it on the doubly-transformed data, so that any differences will be attributable only to the ordination.

```
kyRA <- decorana(kyMax, ira=1)
raSamples <- scores(kyRA, display=c('sites'), choices=c(1, 2))
```

A simple plot of the ordination looks like this:

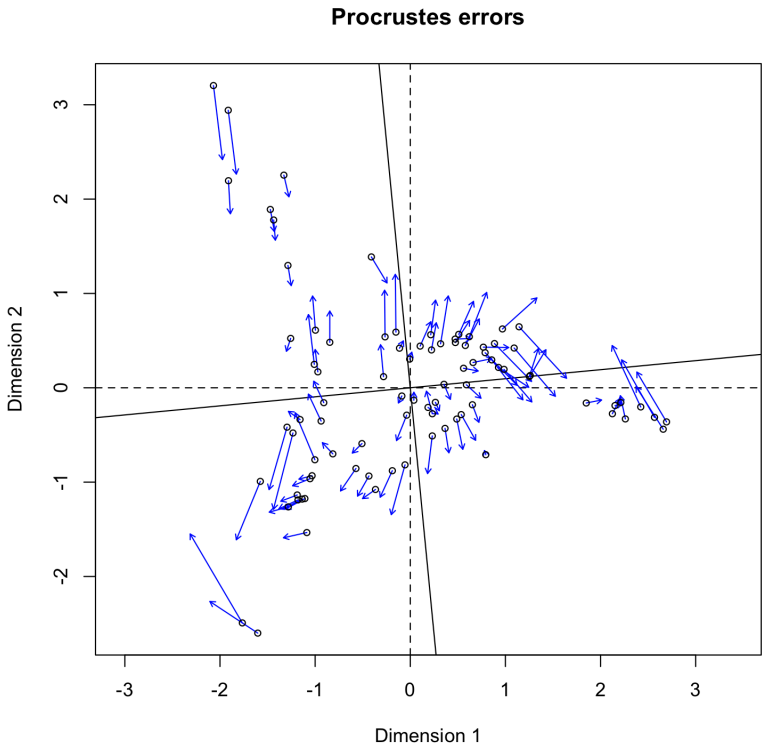


The overall configuration looks similar to the DCA, but more exaggerated. It's interesting to note in this case that the overall triangular shape (or wedge) is present in both ordinations; it is a genuine aspect of the data. This is important to recognize because not all DCA wedges are artifacts, as some authors automatically assume; that's not the case here and in some other studies.

The **`procrustes()`** function, also part of the vegan package, will allow us to quantitatively compare the two ordination configurations. The first argument to the function is the target

ordination, and the second argument is the ordination that will be rotated, inverted, translated, and rescaled to best fit the target ordination. We are interested in how the points move from the correspondence analysis to the DCA, so DCA will be the target ordination (the first argument).

```
comparison <- procrustes(kySamples, raSamples)
plot(comparison)
```

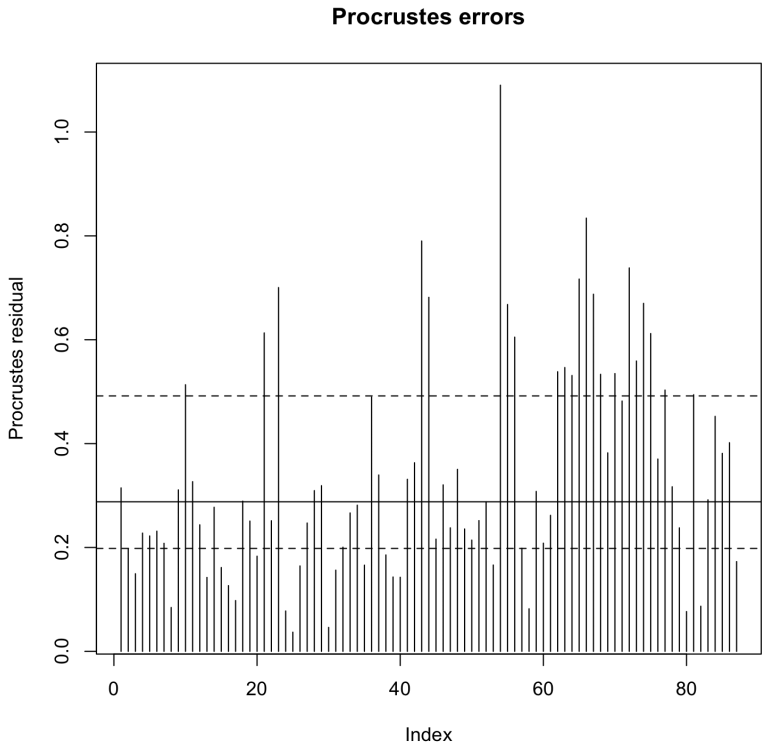


This plot shows the location of the scores for the rotated ordination (**raSamples**) with black circles, with the arrows pointing to the location of the corresponding score for the target ordination (**kySamples**). On the left side, the arrows point vertically, towards the x-axis, reflecting how the DCA ordination is more vertically compressed relative to the RA/CA ordination. Points near the center show an opposite pattern with outward-pointing vectors. The dashed lines represent the axes for the target ordination (**kySamples**), and the solid lines are the axes for the rotated ordination (**raSamples**).

The procrustes object contains several useful components (see the help page). An overall measure of how well the rotated points fit the target ordination is provided by **ss**, the sum of squares of the deviations. If one is comparing several ordinations, this metric can be used to measure how similar two ordinations are to one another. The **summary()** function shows many of these values of the procrustes object, including how much rotation, translation, and scaling was necessary.

A second type of plot of the procrustes object helps one to see how much individual samples (numbered by their row) have shifted position between the two ordinations.

```
plot(comparison, kind=2)
```



Note that some points (e.g., 25, 58) have low residuals, meaning they are in almost the same position in the two ordinations, whereas others (e.g., 54) have large residuals, indicating that their position changes greatly. The residuals shown in this plot are the distances between the target ordination (**`X`**) and the rotated ordination (**`Yrot`**):

```
residuals <- sqrt((comparison$Yrot[, 1] - comparison$X[, 1])^2
  + (comparison$Yrot[, 2] - comparison$X[, 2])^2)
```

Closing comments

Detrended correspondence analysis remains a valuable tool for ordinating ecological data. Because of the similarity in performance of DCA and NMS, I recommend running both to see which better reveals the patterns in the data. Sometimes it is NMS, and sometimes it is DCA.

References

Hill, M.O., and H.G. Gauch, Jr., 1980. Detrended correspondence analysis: an improved ordination technique. Vegetatio 42: 47-58.

Holland, S.M., and M.E. Patzkowsky, 2004. Ecosystem structure and stability: middle Upper Ordovician of central Kentucky, USA. Palaios 19:316–331.

Jackson, D.A., and K.M. Somers, 1991. Putting things in order: the ups and downs of detrended correspondence analysis. American Naturalist 137: 704-712.

Kenkel, N.C., and L. Orlóci, 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67: 919-928.

Legendre, P., and L. Legendre, 1998. Numerical Ecology. Elsevier: Amsterdam, 853 p.

McCune, B., and J.B. Grace, 2002. Analysis of Ecological Communities. MjM Software Design: Gleneden Beach, Oregon, 300 p.

Minchin, P.R., 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69: 89-107.

Patzkowsky, M.E., and S.M. Holland, 2012. Stratigraphic Paleobiology: Understanding the Distribution of Fossil Taxa in Time and Space. University of Chicago Press, Chicago.

Wartenberg, D., S. Ferson, F.J. Rohlf, 1987. Putting things in order: a critique of detrended correspondence analysis. American Naturalist 129: 434-448.