Correspondence analysis of raw data

Michael Greenacre

Department of Economics and Business

Universitat Pompeu Fabra

Ramon Trias Fargas, 25-27

08005 Barcelona

SPAIN

E-mail: michael@upf.es

NOTE TO READERS OF THE ELECTRONIC PDF VERSION:

This article contains embedded dynamic graphics in Figure 3, which can be seen by simply clicking on the figure to see it in motion. To enable this feature you should use the latest version of the free Acrobat Reader, version 9, to open the PDF file.

Abstract: Correspondence analysis has found extensive use in ecology, archeology, linguistics and the social sciences as a method for visualizing the patterns of association in a table of frequencies or nonnegative ratio-scale data. Inherent to the method is the expression of the data in each row or each column relative to their respective totals, and it is these sets of relative values (called profiles) that are visualized. This 'relativization' of the data makes perfect sense when the margins of the table represent samples from sub-populations of inherently different sizes. But in some ecological applications sampling is performed on equal areas or equal volumes so that the absolute levels of the observed occurrences may be of relevance, in which case relativization may not be required. In this paper we define the correspondence analysis of the raw 'unrelativized' data and discuss its properties, comparing this new method to regular correspondence analysis and to a related variant of non-symmetric correspondence analysis.

Keywords: Abundance data, biplot, Bray-Curtis dissimilarity, profile, size and shape, visualization.

1. Introduction

Correspondence analysis (CA) and its variants – multiple, joint, subset and canonical correspondence analysis – have found acceptance and application by a wide variety of researchers in different disciplines, notably the social and environmental sciences (for an up-to-date account, see Greenacre, 2007). The method has also appeared in the major statistical software packages, for example SPSS, Minitab, Stata, SAS, Statistica and XLSTAT, and it is freely available in several implementations in R (R Development Core Team, 2007) – for example, the **vegan** package by Oksanen et al. (2006) and the **ca** package by Nenadić and Greenacre (2007). The method is routinely applied to a table of non-negative data to obtain a spatial map of the important dimensions in the data, where proximities between points and other geometric features of the map indicate associations between rows, between columns and between rows and columns.

In the social science context where the method originated, CA is typically applied to a cross-tabulation, or *contingency table*, between two or more categorical variables based on a random sample of respondents. For example, the respondents could be cross-classified according to a demographic variable such as education level as well as the categories of response to a survey question. Because the demographic groups are inherently of different sizes, a valid comparison between these groups is achieved by expressing the response frequencies relative to their respective totals, a process which we call *relativization,* or *profiling*. It is these vectors of relative frequencies, or *profiles*, that are visualized in the resulting CA maps.

The application of CA in ecology is radically different in several respects. Abundance data, for example, are collected from a number of physical samples, usually areas or volumes, in which various species are identified and counted. The resulting table is not a statistical contingency table, since each individual counted in the table is not obtained by random sampling. The application of CA, however, to such abundance data is justified because of the method's links to the Gaussian model in gradient analysis, and to ecological concepts such as niche theory and coenoclines (see, for example, Gauch (1982)).

Because ecological samples are often of the same physical size (fixed size quadrat in botany, or fixed volume in marine research) the question arises whether it is necessary to profile the abundances with respect to the total abundance in each sample. The frequently used Bray-Curtis index, for example, aggregates absolute differences between raw abundances in two samples, expressing this sum relative to the totals of the two samples. If two samples were exactly in proportion across the species but one more abundant overall than the other, Bray-Curtis would give a positive dissimilarity whereas in CA the distance would be zero – in this sense Bray-Curtis takes the 'size' of the abundance values into account as well as their 'shape', whereas CA only considers their 'shape'. In this article we are interested in an alternative version of CA that does analyze raw 'unrelativized' data on samples that are of physically equal sizes. In the process we show that this analysis has an interesting dual problem which can be linked with a variant of so-called "non-symmetric correspondence analysis" (NSCA). We compare the new method's properties with those of regular CA, illustrating these on a set of benthic data from the North Sea.

2. Correspondence analysis of relative and raw measurements

The data that use as an illustration are available on the website of Greenacre (2007) and are also given as an appendix in *Ecological Archives*. These are benthic abundance data of 92 species (columns of the table) from equal volume samples taken at 11 locations in a grid around the Ekofisk oilfield and also two reference locations that can be regarded as unpolluted (hence, 13 rows, which we refer to as 'sites'). These are typical ecological abundance data, with some species occurring at very low levels at some sites and others at most or all of the sites. Regular

CA would visualize them in their relative form; that is, the profiles of the rows of 92 values are calculated relative to total abundances of the sites, chi-square distances are computed between the site profiles, and these profiles are then weighted proportionally to their respective site abundances in the dimension-reduction step to achieve the low-dimensional map of the sites. Species can then be displayed as unit profiles (i.e., vectors of 91 zeros and a 1 in the column corresponding to the respective species) and the resulting map is a well-defined *biplot*, called the *asymmetric map* by Greenacre (2007) – see Figure 1. In this version of the asymmetric map sites are at weighted averages of the species points, related to coenocline theory, but we can also have the alternative asymmetric map where species are at weighted averages of the (unit) site points, which relates to gradient analysis, the Gaussian model and niche theory.

The algebraic definition of the analysis leading up to the map in Figure 1 is as follows, assuming the raw data matrix is denoted by **N**. The notation is simplified if we regard the initial data matrix as the matrix $\mathbf{P} = [p_{ij}]$ where $p_{ij} = n_{ij}/n$, where n is the grand total of the table (CA is invariant with respect to the grand total). Let the row and column totals of **P** be the vectors **r** and **c** respectively – these are the weights, or *masses*, associated with the rows and columns. Let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses. The row profiles are then rows of the matrix $\mathbf{D}_r^{-1}\mathbf{P}$ and the computational algorithm to obtain Figure 1, using the singular value decomposition (SVD), is as follows:

1. Center the row profiles with respect to their average $\mathbf{c}^\mathsf{T}$, then pre-multiply by $\mathbf{D}_r^{1/2}$ to weight the profiles by their masses, and post-multiply by $\mathbf{D}_c^{-1/2}$ to engender the chi-square metric between rows:

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-1/2} \ . \qquad (1)$$

2. Calculate the SVD: $\mathbf{S} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\mathsf{T}$ where $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$. (2)

3. Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\sigma$. (3)

4. Standard coordinates of columns: $\boldsymbol{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$. (4)

(For the distinction between principal and standard coordinates, see Greenacre (2007).) Figure 1 is the joint display of **F** and **Γ**, using their first two columns corresponding to the two major principal axes.

In order to analyze the raw values rather than the relative values, no division by row totals is performed and the chi-square distance is determined by the averages of the raw values across rows. There are no differential weights of the rows, because using the raw values already includes the absolute level of each site in the description vector. To distinguish the two alternatives here, we call this analysis *CA-raw* as opposed to the regular CA, *CA-relative*, defined above by (1)–(4). To put CA-raw on a comparable scale to CA-relative, each row of **P** is divided by $(1/I)$, where $(1/I)$ is the constant row mass (as opposed to CA-relative, where each row of **P** is divided by the variable row mass); that is, the equivalent of the matrix of row profiles $\mathbf{D}_r^{-1}\mathbf{P}$ is $I\,\mathbf{P}$. This is equivalent to taking the original rows of the matrix and dividing them all by the constant n/I, the grand total of the matrix averaged over the I rows. Some of these transformed rows will sum to more than 1 and others to less, but the (unweighted) average row is still $\mathbf{c}^\mathsf{T}$, since $(1/I)\mathbf{1}^\mathsf{T}(I\mathbf{P}) = \mathbf{c}^\mathsf{T}$. Hence, in this scale, the centering is the same as CA-relative and the chi-square metric is still based on the inverse of the matrix $\mathbf{D}_c$ – this shows that the essential difference between CA-raw and CA-relative is in the definition and weighting of the points, not in the form of the distance function between them. Hence, to perform CA-raw, the only change necessary in the algorithmic scheme (1)–(4) is to replace $\mathbf{D}_r$ throughout by $(1/I)\mathbf{I}$.

3. Further properties of CA-raw

In CA-raw the matrix which is decomposed by the SVD is (from (1), replacing **r** by a constant vector of 1/*I*'s):

$$\mathbf{S}^* = (1/I)^{1/2}(I\mathbf{P} - \mathbf{1}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2} \ .$$

which can be written equivalently as:

$$\mathbf{S}^* = (1/I)^{-1/2}(\mathbf{P}\mathbf{D}_c^{-1} - (1/I)\mathbf{1}\mathbf{1}^{\mathsf{T}})\mathbf{D}_c^{1/2} \qquad (5)$$

Comparing (5) to (1) (the columns may be turned into rows by transposing (5) to make this comparison clearer), it appears that CA-raw of the sites can be equivalently thought of as a CA of the species profiles, weighted by their masses, but with respect to a uniform profile across sites, and using the corresponding chi-square metric with equal dimension weights. Since non-symmetric correspondence analysis (NSCA) on species profiles analyzes the species profiles in the Euclidean metric – see Dray, Chessel and Thioulouse (2003: Table 1) – this shows that the difference between CA-raw and this version of NSCA is in the centering of the species profiles. In species-profiles NSCA, the profiles are centered with respect to the average species profile (**r** in our notation), whereas CA-raw centers with respect to the uniform profile (1/*I*)**1**. The uniform profile is the natural centre because under the hypothesis of no difference between the sites, based on equal-size sampling, we expect a uniform distribution of all species across the sites. So this is the natural centre of the data.

CA-raw also turns out to have min{*I*–1, *J*–1} dimensions, because the uniform center is orthogonal to the centered profiles:

$$\mathbf{1}^{\mathsf{T}}(\mathbf{P}\mathbf{D}_c^{-1} - (1/I)\mathbf{1}\mathbf{1}^{\mathsf{T}}) = \mathbf{c}^{\mathsf{T}}\mathbf{D}_c^{-1} - (1/I)I\mathbf{1}^{\mathsf{T}} = \mathbf{1}^{\mathsf{T}} - \mathbf{1}^{\mathsf{T}} = \mathbf{0}^{\mathsf{T}} \qquad (7)$$

As a consequence, the site coordinates have arithmetic average 0 and in the alternative asymmetric map, the species displayed in principal coordinates are at weighted averages of the sites in standard coordinates, as in CA-relative, thus still conforming to niche theory:

- SVD of $\mathbf{S}^* = (1/I)^{-1/2}(\mathbf{PD}_c^{-1} - (1/I)\mathbf{11}^\mathsf{T})\mathbf{D}_c^{1/2} = \mathbf{UD}_\alpha\mathbf{V}^\mathsf{T}$ (8)
- Standard coordinates of sites: $\mathbf{\Phi} = I^{1/2}\mathbf{U}$, $\mathbf{1}^\mathsf{T}\mathbf{\Phi} = \mathbf{0}^\mathsf{T}$ as a result of (7) (9)
- Principal coordinates of species: $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{VD}_\alpha$ (10)
- From (8)-(10), $\mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}^\mathsf{T}\mathbf{\Phi} - \frac{1}{I}\mathbf{11}^\mathsf{T}\mathbf{\Phi} = \mathbf{D}_c^{-1}\mathbf{P}^\mathsf{T}\mathbf{\Phi}$ (11)

(Notice that this barycentric (i.e., weighted average) relationship does not hold in NSCA).

As a further consequence of (5), the null distribution of the inertia in CA-raw can be derived simply as follows. Under the hypothesis of no difference between the sites, the observed frequencies of a species are compared to the expected (uniform) frequencies in the usual chi-square test of fit, giving a chi-square statistic with (*I*–1) degrees of freedom. Summing these for each species gives the sum of *J* independent chi-squares with (*I*–1) degrees of freedom, that is a chi-square with *J*(*I*–1) degrees of freedom. This is exactly the inertia of CA-raw (i.e., the sum of squares of the elements of $\mathbf{S}^*$ in (5)) multiplied by the grand total *n* of the abundance table:

$$\begin{aligned} n \times \sum_i \sum_j (s_{ij}^*)^2 &= nI \sum_i \sum_j c_j \left(\frac{p_{ij}}{c_j} - \frac{1}{I} \right)^2 \\ &= \sum_j \sum_i \left(np_{ij} - \frac{nc_j}{I} \right)^2 / \left(\frac{nc_j}{I} \right) \end{aligned} \qquad (12)$$

Where np_{ij} is the observed abundance of species *j* at site *i*, and nc_j is the total abundance of the *j*-th species, which under the null hypothesis is distributed equally across the *I* sites.

4. Illustration

As is well known, in CA-relative an $I \times J$ table has dimensionality $K = \min\{I–1, J–1\}$, which in this example is 12. As shown in Section 3, this property carries over to CA-raw. In each case the *inertia*, defined as the sum of squares of the matrix **S** in (1) (CA-relative) or of $\mathbf{S}^*$ in (5) (CA-raw) measures the total variance in the table, and is equal to the sum of the squared singular values – in this example the inertias in the two cases are 0.783 and 0.972 respectively. The two-dimensional solutions are shown in Figures 1 and 2. For purposes of comparison, we have shown both maps in the asymmetric solution where rows (sites) are in principal coordinates, in order to show chi-square distances between sites in both cases, and columns (species) in standard coordinates – hence, referring to (9) and (10), the coordinates used for CA-raw are $I^{1/2}\mathbf{UD}_\alpha$ for sites and $\mathbf{D}_c^{-1/2}\mathbf{V}$ for species (i.e., the scaling by singular values is simply interchanged between row and column points). All computations are performed in R (R development core team, 2009), making use of the **ca** package for CA developed by Nenadić and Greenacre (2007) – the R script for all the computations are provided in *Ecological Archives*.

In the CA-raw of Figure 2 where sites are compared with respect to absolute abundance levels (or, equivalently, where the species profiles are compared to a uniform profile), the very high value for *Myriochele oculata* in site S24 dominates the analysis, contributing 57.0% to the two-dimensional solution (compared to 31.8% in the case of CA-relative). Another high-abundance species, *Chaetezona setosa*, with particular high-abundance in site S15, similarly dominates the second axis (contributing 27.6% to the two-dimensional solution).

The test statistic (equal to the inertia 0.9722 in the CA-raw analysis multiplied by the grand total, n=9595) is equal to 9328, which is in the very extreme right tail of the chi-square distribution with 92×12=1104 degrees of freedom. We corroborated the null distribution by generating

10000 abundance tables with the 92 species randomly and uniformly distributed over the 13 sites and the average of the 10000 null statistics was 1104.17, with variance 1978.2. The mean is spot on, while the lower than theoretical variance of 2208 is due to the many rare species in this data set, whose contributions to total chi-square tend to be less than the theoretical component of 2(*I*–1) for each species.

Both maps are biplots and projecting the sites onto directions given by the species points, or vice versa, would lead to approximate values in the relative or raw biomass matrix, as the case may be. The success of the recovery is measured by the percentage of inertia explained, 57.5% (CA-relative) and 68.5% (CA-raw) (see Greenacre (1993, 2007: chap. 13) for a discussion of biplots in the CA context).

Figure 3 is a dynamic graphic showing a smooth transition from CA-relative to CA-raw. The way the animation is constructed is explained in the caption. Showing the change dynamically can help in comparing how the two approaches differ in practice, since the faster moving points are those that are exhibiting the biggest changes.

4. Discussion and conclusions

The analysis of raw site data (as opposed to site profiles) in the CA framework is a simple variation of the usual algorithm. This analysis makes sense only if the data at each site emanate from equal-sized samples, as is often the case in ecological research. We have shown that the visualization of the raw site abundances can be achieved very simply using an almost identical algorithm to regular CA, where the raw data are divided by a constant, the average abundance (or biomass as the case may be) per site, and equal weights are allocated to the sites.

From the species point of view, CA of raw site data is equivalent to a CA of the species profiles centered with respect to a uniform expected profile, which then engenders a chi-square distance with uniform dimension-weighting, that is a Euclidean distance. Because inter-species distances are Euclidean, the analysis can also be thought of as a variant of non-symmetric correspondence analysis of the species profiles, centered with respect to the uniform center rather than the usual NSCA centering by the average species profile. The inter-site distance measure in the CA of raw abundances can be considered as an alternative to Bray-Curtis dissimilarities based on raw abundance data (often power-transformed), with the advantage that it is a true Euclidean-embeddable metric.

As a final remark, we note a curiosity about NSCA. While CA-raw of site abundances has an interesting dual problem described above, the dual problem of NSCA of site profiles (i.e., sites predicting species) is almost a regular CA of the species profiles, but with species weighted in the dimension reduction by the squares of their masses. This can be seen by expressing the "**S**" matrix which is decomposed by the SVD as:

$$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}-\mathbf{1}\mathbf{c}^{\mathsf{T}}) = \mathbf{D}_r^{-1/2}(\mathbf{P}\mathbf{D}_c^{-1}-\mathbf{r}\mathbf{1}^{\mathsf{T}})\mathbf{D}_c \qquad (13)$$

The left-hand side is the matrix decomposed in site-profiles NSCA, and the right-hand side is the matrix decomposed in species-profiles CA, with the usual centering and chi-square metric, but with the squares of the species masses (in regular CA the final matrix on the right-hand side would be $\mathbf{D}_c^{1/2}$, not $\mathbf{D}_c$ – *cf.* matrix expression (1), but transposed for column profiles). Thus a row-profiles NSCA solution is achieved exactly through a column-profiles CA with the squares of the column masses, and vice versa. It seems then that squaring the masses of the species, which accentuates the higher-abundance species, does the same job in a CA that NSCA achieves by using the Euclidean metric between site profiles instead of the chi-square metric.

Acknowledgments

This research has been supported by the Fundación BBVA, Madrid, Spain, and the author expresses his gratitude to the Foundation's director, Prof. Rafael Pardo. Partial support of Spanish Ministry of Education and Science grant MEC-SEJ2006-14098 is also acknowledged. The first version of this paper was written while the author was on leave at Stanford University: thanks to Trevor Hastie and colleagues of the Department of Statistics. The paper was radically improved by comments on the first version from an anonymous referee.

References

Dray, S., Chessel, D, and Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, **84**, 3078–3089.

Gauch, H. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press, U.K.

Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics*, **20**, 251 – 269.

Greenacre, M. J. (2007). *Correspondence Analysis in Practice. Second Edition.* London: Chapman & Hall / CRC Press. Published in Spanish translation by the Fundación BBVA, Madrid, 2008, freely downloadable from URL http://www.fbbva.es/TLFU/tlfu/ing/publicaciones/fichalibro/index.jsp?codigo=300

Nenadić, O. and Greenacre, M. J. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: the **ca** package. *Journal of Statistical Software*, **20** (1). URL http://www.jstatsoft.org/v20/i03/

Oksanen J., Kindt R., Legendre P. & O'Hara R.B. (2006). **vegan**: Community Ecology Package version 1.8-3. URL http://cran.r-project.org/

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

Figure 1: "CA-relative": regular correspondence analysis of benthos data (i.e., displaying relative abundances), with rows in principal coordinates and columns in standard coordinates (row principal asymmetric map) – thus sites are at weighted averages of species points. The 11 species with labels are amongst the most abundant and each make a contribution of more than 1% to the solution (their font sizes are monotonically related to their contribution), the remaining 81 species, including all the rare ones, contribute very little individually, about 15% collectively, and are indicated by dots. Total inertia = 0.783; inertia explained in the two-dimensional map = 57.5%.

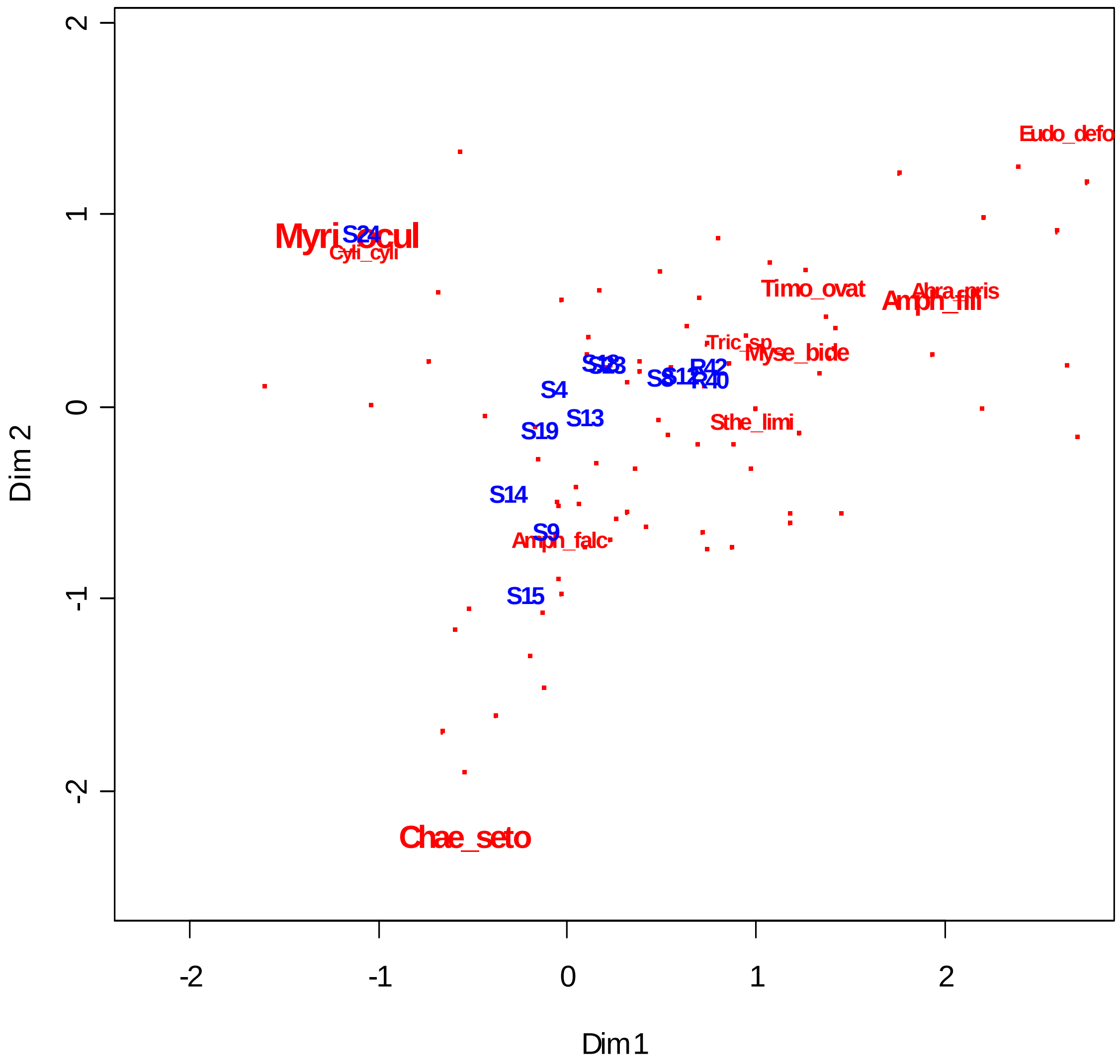


Figure 2: “CA-raw”: correspondence analysis of raw site abundances (i.e., no profiling of rows), with rows in principal coordinates and columns in standard coordinates. Note that sites are not at weighted averages of species points, but species would be at weighted averages of sites if we displayed the alternative asymmetric map, as shown in (11). Font sizes of labeled points again indicate level of contribution to the solution, which is more concentrated into the two most abundant species *Myriochele oculata* and *Chaetesona setosa*, found in very large numbers in sites S24 and S15 respectively. Total inertia = 0.972; inertia explained in the two-dimensional map is 68.5%.

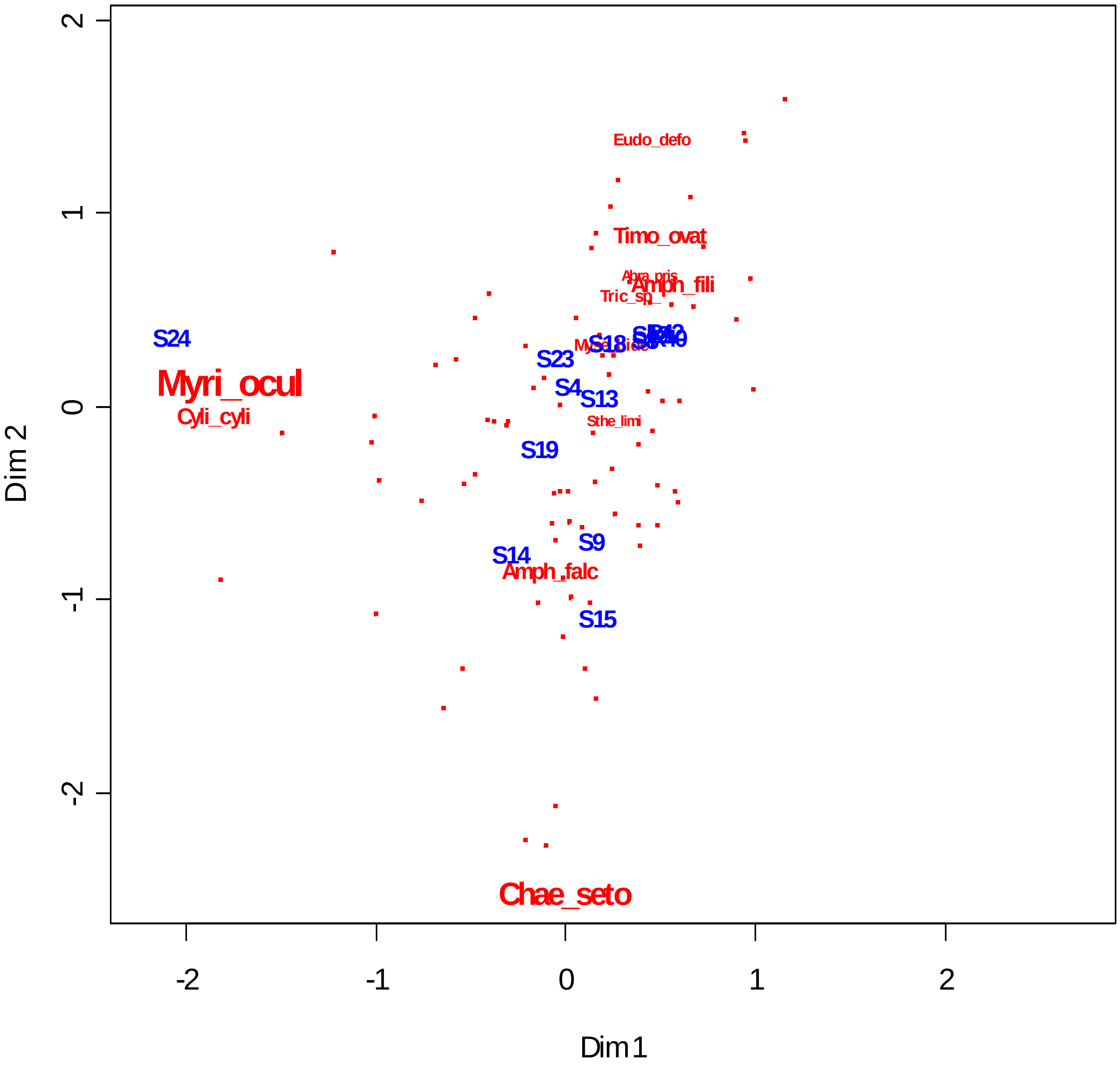
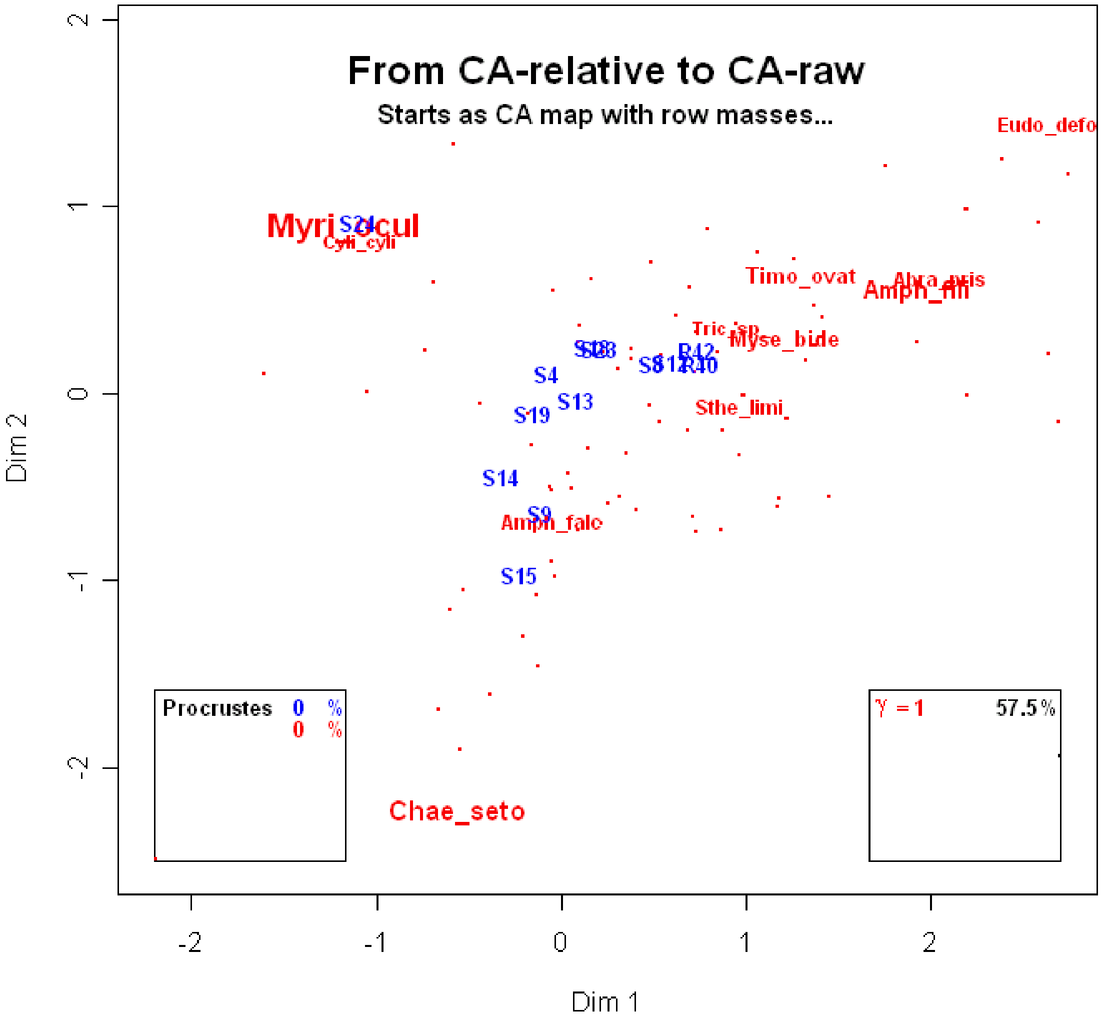


Figure 3: The transition from CA-relative to CA-raw visualized dynamically. This animation, which can be viewed onscreen in the PDF version of the paper, or alternatively as a flash video in *Ecological Archives*, shows the smooth change from Figure 1 to Figure 2 as the weights of the sites are changed in small steps from the usual CA masses **r** to uniform masses $(1/I)\mathbf{1}$. This is achieved by defining a convex linear combination of these two alternative sets of weights: $\mathbf{w} = \gamma\mathbf{r} + (1-\gamma)(1/I)\mathbf{1}$ and using **w** as the weights when γ is set to 1, 0.99, 0.98, …, 0.01, 0. When $\gamma = 1$ the analysis is CA-relative and when $\gamma = 0$ it is CA-raw, and combinations of these two extremes in-between. The box on the right shows the evolution (moving leftwards as γ descends) of the total inertia (top bold curve) and the inertias of axes 1 and 2 (lower two curves). The box on the left shows the evolution (moving rightwards) of the Procrustes statistics that measure differences between the current row and column configurations and the first ones (CA-relative), showing that the site points change more than the species points.


From CA-relative to CA-raw
Starts as CA map with row masses...
Eudo_defo
Myri_ocul
S24
Cyli_cyli
Timo_ovat
Alva_avis
Amph_fili
Tric_sp
Myse_bide
S18
R42
S8
S12
R40
S4
S13
S19
Sthe_limi
S14
S9
Amph_falc
S15
Chae_seto
Procrustes 0 %
γ = 1
Dim 2
Dim 1