Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 15 Offprint

CASE STUDY–ECOLOGY

The Relationship between Fish Morphology and Diet

First published: September 2010
ISBN: 978-84-923846-8-6

Supporting websites:
http://www.fbbva.es
http://www.multivariatestatistics.org

Chapter 15

Case Study 3: The Relationship between Fish Morphology and Diet

The multivariate nature of ecological data is illustrated very well in the morphological data on *Arctic charr* fish, described in Chapters 7 and 12. Apart from the fish morphology, an analysis of the stomach contents of each fish was performed to characterize the fish's diet. Because the diet is measured as a set of percentages of the stomach contents, correspondence analysis is an appropriate way of visualizing the diet variables. Now there are two multivariate observations on each fish: the set of morphological measurements as well as the set of dietary estimates. Our aim will be to decide if there is any non-random relationship between the morphology and the diet, and—if there is—to try to characterize and interpret it. Because we used log-ratio analysis in Chapter 7 to visualize the morphological data, we will maintain this approach while focusing the visualization of the morphology on its relationship to the dietary composition.

Contents

Data set "fishdiet"

The data set "morphology" was introduced in Chapter 7, consisting of 26 measurements on each of 75 *Arctic charr* fish, as well as two dichotomous variables indicating the sex (female/male) of each fish and their habitat (littoral/pelagic). In addition, another set of data is available based on an analysis of the stomach contents of each fish—these are estimated percentages of the contents, by volume, that have been classified into 6 food sources:

PlankCop	plankton – copepods	*InsectLarv*	insects – larvae
PlankClad	plankton – cladocerans	*BenthCrust*	benthos – crustaceans
InsectAir	insects – adults	*BenthMussl*	benthos – mussels

A seventh category *Others* includes small percentages of other food sources. The data for the first 10 fish are given in Exhibit 15.1.This data set, called "fishdiet" constitutes a second matrix of data on the same individuals, and can be considered as explanatory variables that possibly explain the morphological data.

Correspondence analysis of "fishdiet" data

Seeing that the data are compositional, one would immediately think of using logratio analysis (LRA), but the large number of zeros makes this approach impractical. Correspondence analysis (CA) is a good alternative but there are two possible approaches here. The first would be to consider just the seven measured percentages—since CA converts the data to profiles, this would re-express the values relative to the actual stomach contents; for example, the first fish in Exhibit 15.1 has only 40% stomach full, and *PlankClad* is 25/40 of this total and 15/40 *InsectAir*, which would be the profile values in CA. The second is to add a column, called "*Empty*" in Exhibit 15.1, which quantifies the emptiness of the stomach, that is 100 minus the sum of the seven measured values. By including the empty component, the sums of each of the rows is now a constant 100% and CA will treat the data in their original form.

Exhibit 15.2 shows the two alternative CAs together for comparison, where we have excluded one fish with zero stomach contents, which would not be a valid observation for the first analysis (all margins have to be strictly positive for CA).

Exhibit 15.1: *Part of data set "fishdiet", showing the first 10 of the* Arctic charr *fish. Data are percentages of stomach contents of different food sources. A column "Empty" has been added as 100 minus the sum of the percentage values in the first seven columns—for example, fish 28 had the whole stomach full, so "Empty" is 0. The supplementary variables sex (1 = female, 2 = male) and habitat (1 = littoral, 2 = pelagic) are also shown*

Fish no.	*PlankCop*	*PlankClad*	*InsectAir*	*InsectLarv*	*BenthCrust*	*BenthMussl*	*Others*	*Empty*	*Sex*	*Habitat*
19	0	25	15	0	0	0	0	60	1	2
23	0	0	0	20	47	8	0	25	2	1
24	0	0	0	8	32	0	0	60	2	1
25	0	0	0	10	22	18	0	50	2	1
27	0	0	0	2	4	4	0	90	1	1
28	0	0	0	10	55	35	0	0	2	1
30	0	0	0	20	44	6	0	30	2	1
31	0	0	0	15	25	40	0	20	1	1
33	0	65	0	0	0	0	0	35	1	2
34	0	48	0	2	0	0	0	50	1	2
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

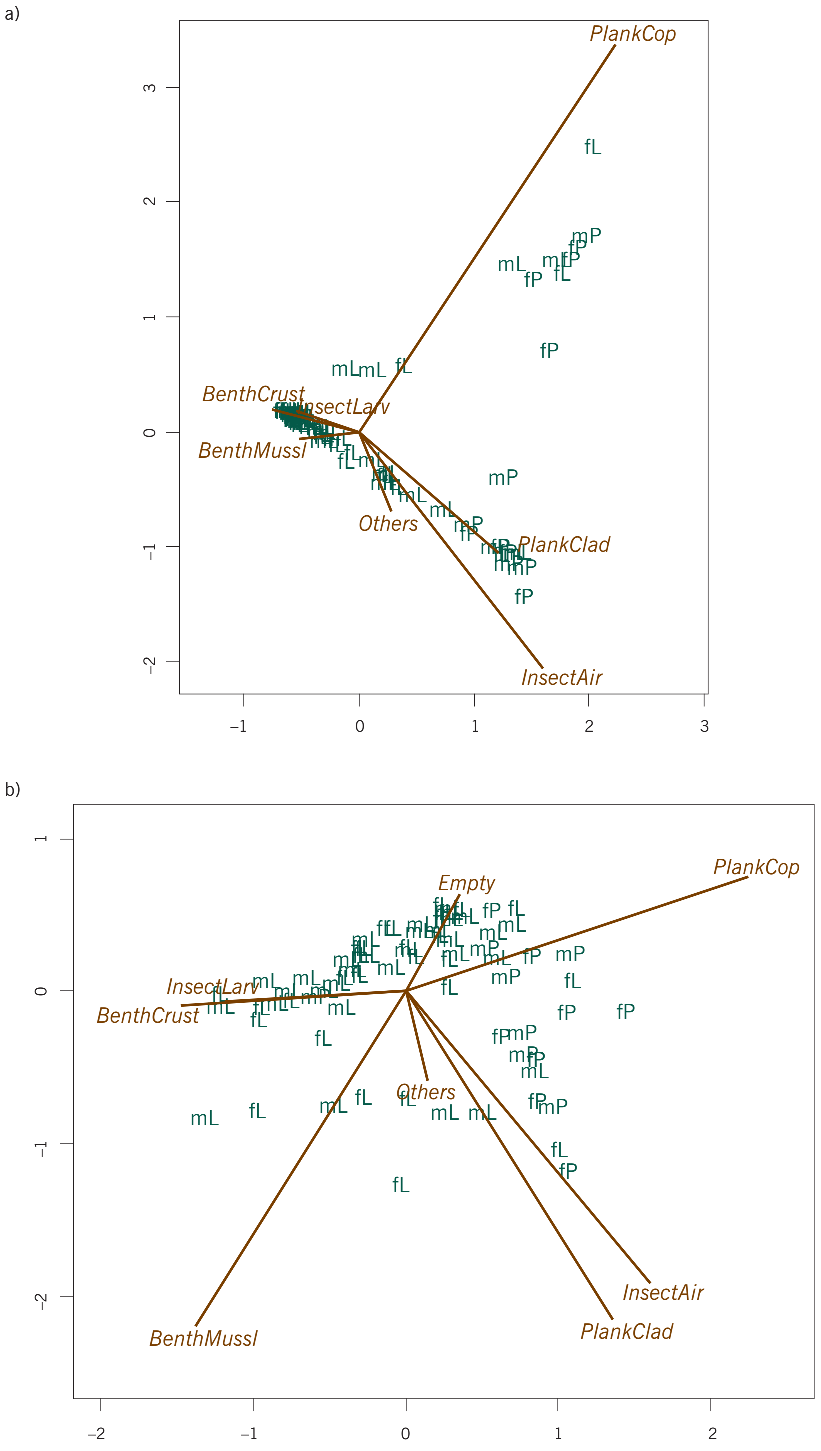
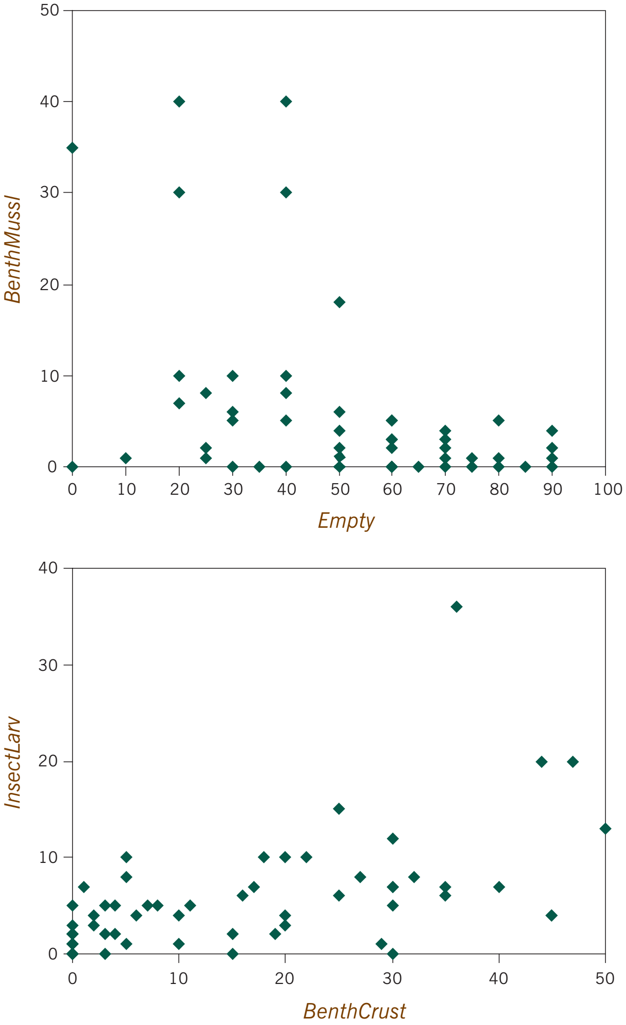


Exhibit 15.2:

*CA biplots of the "fishdiet" data, asymmetric scaling with fish in principal coordinates and food sources in standard coordinates: (a) the biplot is the regular CA of the first seven columns of Exhibit 15.1, while (b) includes column 8 (*Empty*). Fish are labelled by their sex-habitat groups. Total inertias in the two analyses are 1.751 and 1.118 respectively*

Exhibit 15.3:

Scatterplots of two pairs of variables, showing the negative relationship between BenthMussl and Empty and positive relationship between InsectLarv and BenthCrust



There are only a few fish with some *PlankCop*, generally at low percentages, but these tend to be associated with less full stomachs so that in relative terms the presence of *PlankCop* is accentuated in the CA in Exhibit 15.2a. Otherwise, there is an opposition between those with relatively more *PlankClad* and *InsectAir* (bottom right of Exhibit 15.2a) compared to those with relatively more *BenthCrust* and *BenthMussl* (to the left). When *Empty* is included (Exhibit 15.2b), it has a higher mean than the other variables, and the centroid of the display moves close to it. Projecting the fish onto the biplot axis defined by *BenthMussl* and *Empty* implies that there is an inverse relationship between the two columns, shown in the upper scatterplot of Exhibit 15.3. Expressed another way, the proportion of benthic mussels increases with stomach fullness. The coincident directions of *InsectLarv* and *BenthCrust* imply a positive relationship between these two food sources, as shown in the lower scatterplot of Exhibit 15.3.

These two CA biplots display the data in different ways and the biologist needs to decide if either or both are worthwhile. The question is whether the percentages are of interest relative to actual stomach contents, or on their original percentage scale relative to the whole stomach. For example, the separation of the group of fish in the direction of *PlankCop* in Exhibit 15.2a is non-existent in Exhibit 15.2b—relative to what is in the stomach, this group of fish distinguishes itself from the others, but not so much when seen in the context of the stomach as a whole.

How is the diet related to habitat and sex?

As described in Chapter 12, constraining by a categorical variable is equivalent to performing a type of centroid discriminant analysis, illustrated in Exhibit 11.2 for the morphological data. We repeat that analysis on the "fishdiet" data, with the groups defined by the interactively coded sex-habitat variable with four categories: fL, mL, fP and mP. Exhibit 15.4 shows the resulting biplot. As in the morphological analysis of Exhibit 11.2, the habitat differences are more important than the sex differences. Contrary to the morphological analysis, the diet difference between sexes in the pelagic group is bigger than that in the littoral group. The lack of difference between female and male littoral fish (fL and mL) is seen clearly by the single line of individual points to the top left of the biplot, in the direction of *BenthCrust*, *BenthMussl* and *InsectLarv*; while on the right there is a separation into two "streams", mainly female pelagic (fP) to upper right, in the direction of *PlankCop*, and mainly male pelagic (mP) to lower right, in the direction of *PlankClad* and *InsectAir*. There are some exceptions: for example, some fP points are in the lower right group, and there are few male and female littoral fish on the right.

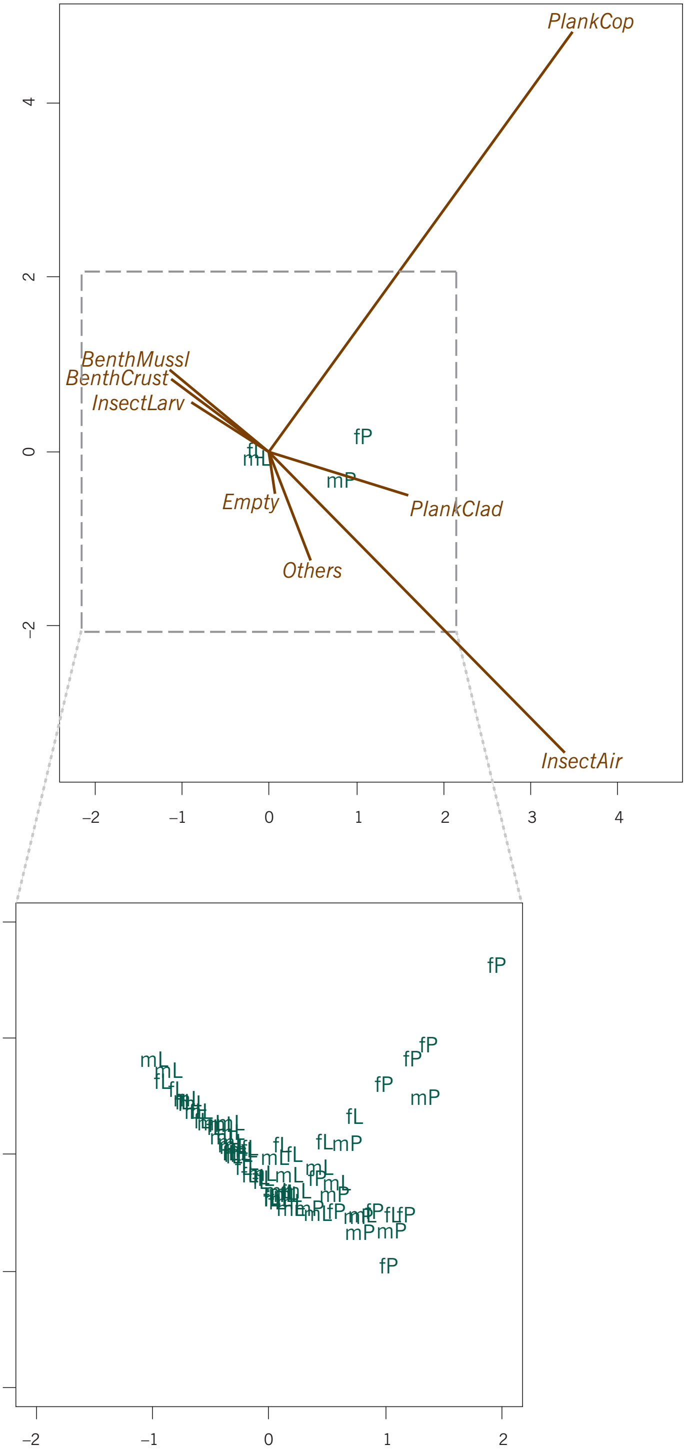
The unconstrained CA in Exhibit 15.2b has total inertia equal to 1.118 and the part of this inertia explained by the sex-habitat variable is equal to 0.213, or 19.0% of the total. A permutation test shows that the relationship between diet and sex-habitat is highly significant: $p < 0.0001$. This part of inertia forms the total inertia in the analysis of Exhibit 15.4, which explains almost all of it (99.6%) in two dimensions (the analysis of the four centroids is three-dimensional, so the 0.4% unexplained is in the third dimension).

Canonical log-ratio analysis

In Chapter 7 we analyzed the morphological data set on its own using the log-ratio approach. We now want to relate the morphological data to the diet data, in other words constrain the dimensions of the log-ratio analysis to be related to the diet variables, which we could call *canonical*, or constrained, *log-ratio analysis* (CLRA). It is useful here to give the equations of the analysis, putting together the theories of Chapters 7 and 12.

Exhibit 15.4:

CA discriminant analysis of the sex-habitat groups (equivalent to CCA with categorical sex-habitat variable as the constraining variable). The centroids of the four groups are shown in the upper plot. The individual fish, which are contained in the box shown in the biplot, have been separated out in the plot, with enlarged scale for sake of legibility. Total inertia of the four centroids is equal to 0.213



In Chapter 7 log-ratio analysis was defined as the weighted SVD: $\mathbf{S} = \mathbf{D}_r^{½}\mathbf{Y}\mathbf{D}_c^{½} = \mathbf{U}\mathbf{D}_\varphi\mathbf{V}^\mathsf{T}$of the double-centred matrix: $\mathbf{Y} = (\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{L}(\mathbf{I} - \mathbf{1}\mathbf{c}^\mathsf{T})^\mathsf{T}$ of logarithms: $\mathbf{L} = \log(\mathbf{N})$ of the data $\mathbf{N}$ (see (7.1)–(7.4)). The dimensionality of this analysis is equal to 25, one less than the number of morphometric measurements. The constraining variables are the 7 diet variables, without the "*Empty*" column (here it makes no difference whether it is included or not as an explanatory variable). The matrix $\mathbf{X}$ consists of the 7 diet variables after they have been standardized. Then (12.2) defines the projection matrix as $\mathbf{Q} = \mathbf{D}_r^{½}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_r^{½}$ and the matrix $\mathbf{S}$ is projected onto the space of the diet variables by $\mathbf{S}^* = \mathbf{QS}$. The unconstrained component of $\mathbf{S}$ in the space uncorrelated with the diet variables is $\mathbf{S}^\perp = (\mathbf{I} - \mathbf{Q})\mathbf{S}$ (see (12.3) and (12.7) respectively). The dimensionality of $\mathbf{S}^*$ is 7 in this case. The SVD of $\mathbf{S}^*$ is performed in the usual way, with subsequent computation of principal and standard coordinates.

Relationship of morphology to diet

The first interesting statistic from this constrained log-ratio analysis is the part of the morphological log-ratio variance that is explained by the diet variables: it turns out to be 14.5%, so that 85.5% is not related—at least, linearly—to the diet. Our interest now turns to just that 14.5% of the explained variance, 0.0002835, compared to the total variance of 0.001961 of the morphological data. This variance is now contained in a 7-dimensional space, and our view of this space is, as always, in terms of the best-fitting plane. The principal axes of this plane account for small percentages of the total (original) morphological variance (5.6% and 4.0% respectively), but for the moment we focus on how the constrained variance (0.0002835) is decomposed, and the axes account for 38.9% and 27.6% of that amount, which is the part of the variance that interests us. Exhibit 15.5 shows the biplot of these first two constrained axes.

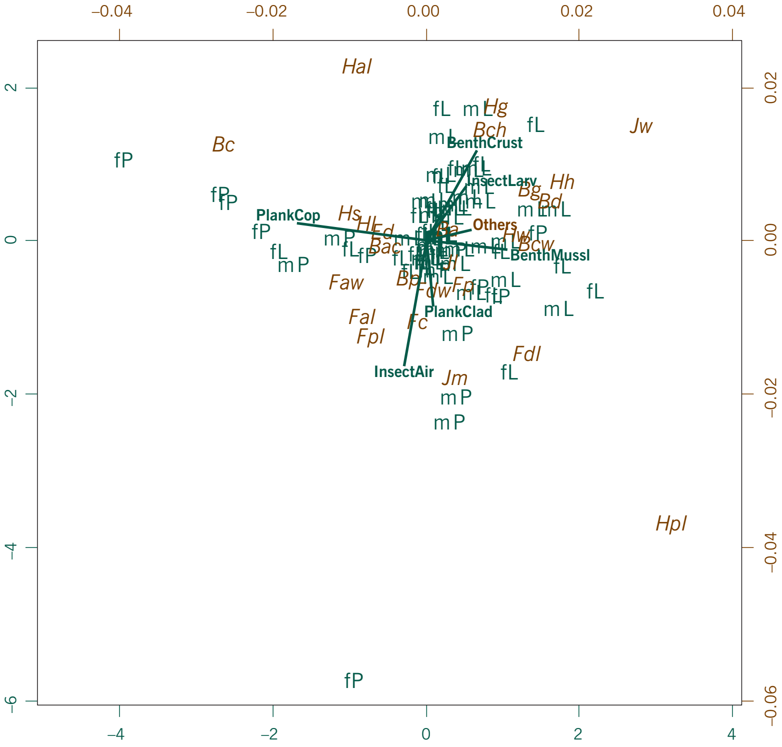
In Exhibit 15.5 the fish points are in standard coordinates and the morphological variables in principal coordinates. As in Exhibit 7.3, the dispersion of the fish points is so low that the coordinates have to be scaled up to appreciate their relative positions. The diet variables are displayed according to their correlation coefficients with the axes, and have also been scaled up (by 2) to facilitate their display. As explained in Chapter 12, there are two ways to show the diet variables: using the correlation coefficients as in Exhibit 15.5, or in terms of the coefficients of the linear combinations of the variables that define the axes. For example, axes 1 and 2 are in fact the linear combinations:

$$\begin{aligned} \mathit{Axis\ 1} = {} & -0.761 \times \mathit{PlankCop} + 0.272 \times \mathit{PlankClad} - 0.159 \times \mathit{InsectAir} + 0.103 \times \\ & \mathit{InsectLarv} + 0.071 \times \mathit{BenthCrust} + 0.388 \times \mathit{BenthMussl} + 0.217 \times \mathit{Others} \end{aligned}$$

$$\begin{aligned} \mathit{Axis\ 2} = {} & 0.280 \times \mathit{PlankCop} - 0.076 \times \mathit{PlankClad} - 0.689 \times \mathit{InsectAir} + 0.140 \times \\ & \mathit{InsectLarv} + 0.505 \times \mathit{BenthCrust} - 0.188 \times \mathit{BenthMussl} + 0.116 \times \mathit{Others} \end{aligned}$$

Exhibit 15.5:

Weighted LRA biplot constrained by the fish diet variables, with rows (fish) in standard coordinates and columns (morphological variables) in principal coordinates. The coordinates of the diet variables have been multiplied by 2 to make them more legible (use the green scale for these points). 66.5% of the constrained variance is accounted for (but only 9.6% of the original total variance)



where the axes (that is, the coordinates of the fish on the axes) as well as the variables are all in standard units, that is with standard deviations equal to 1.

Because the diet variables are correlated, the variable-axis correlations are not the same as the above coefficients, which are regression coefficients if the axes are regressed on the variables.

As explained in Chapter 12, the above equations are exact (that is, $R^2 = 1$ if one were to perform the regression), but thinking of Exhibit 15.5 from the biplot viewpoint, the R^2 of each diet variable can be computed as the sum of squared correlations to measure how accurately each variable is displayed:

$$\begin{array}{lrcrcl}
\textit{PlankCop:} & (-0.860)^2 & + & (0.113)^2 & = & 0.752 \\
\textit{PlankClad:} & (0.055)^2 & + & (-0.447)^2 & = & 0.203 \\
\textit{InsectAir:} & (-0.142)^2 & + & (-0.806)^2 & = & 0.669 \\
\textit{InsectLarv:} & (0.260)^2 & + & (0.370)^2 & = & 0.205 \\
\textit{BenthCrust:} & (0.336)^2 & + & (0.610)^2 & = & 0.485 \\
\textit{BenthMussl:} & (0.496)^2 & + & (-0.052)^2 & = & 0.249 \\
\textit{Others:} & (0.299)^2 & + & (0.083)^2 & = & 0.096
\end{array}$$

PlankCop and *InsectAir* are explained more than 50%—this means that we could recover their values with an error of less than 50% by projecting the fish points onto the biplot axes that they define in Exhibit 15.5. Variables such as *PlankClad*, *InsectLarv* and *BenthMussl* are poorly reconstructed in the biplot. But remember that it was not the intention of this biplot to recover these values—in fact, this was the aim of the correspondence analysis of Exhibit 15.2. The aim here is rather to recover the values of the morphological variables that are directly related to diet, in a linear sense.

Permutation test of morphology–diet relationship

In order to test for significance of the morphology–diet relationships we are detecting, a permutation test can be performed as described previously: use the inertia explained by the diet variables as a test statistic, and then randomly permute the sets of diet values so that many (9999 in this case) additional data sets are constructed under the null hypothesis that there is no morphology–diet correlation. The result is the null permutation distribution in Exhibit 15.6. If there were no (linear) relationship between morphology and diet, we would expect a propor-

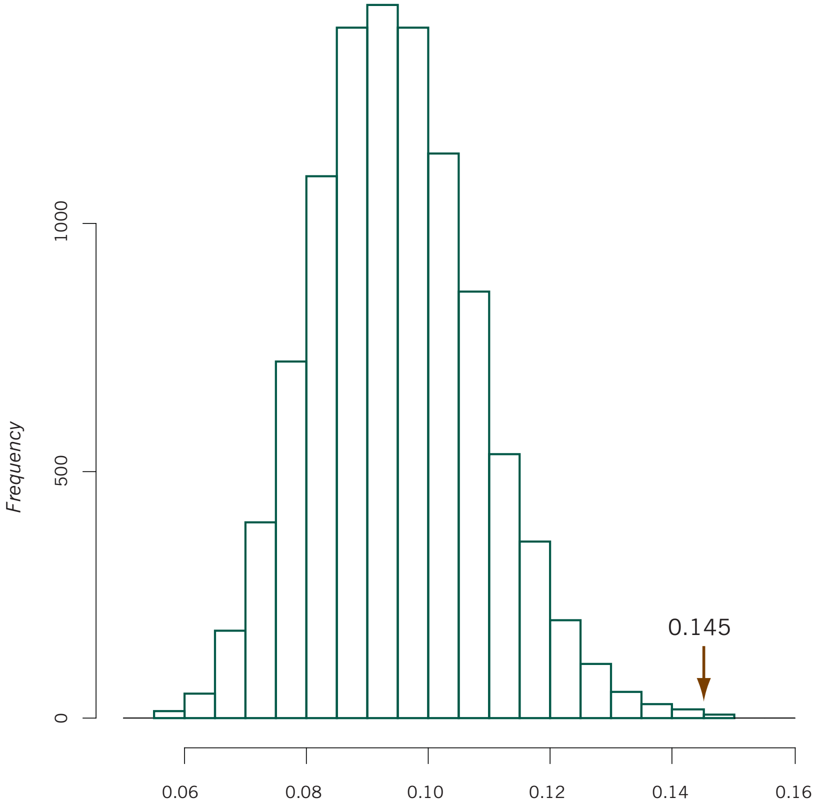


Exhibit 15.6:

Permutation distribution of the proportion of variance explained in the morphological variables by the diet variables, under the null hypothesis of no relationship between these two sets of variables. The p-value associated with the observed proportion of 0.145 is 0.0007 tion of explained variance of 0.093 (9.3%), with the estimated distribution shown. Our observed value of 0.145 (14.5%) is in the far right tail of the distribution, and only 6 of the permuted data sets gives a proportion higher than this value—hence the *p*-value is 7/10,000 = 0.0007 (the observed value is included with the 26 higher ones to make this calculation).

Up to now we included all of the diet variables, but it may be that only a subset of them explain a significant part of the variance. The individual contributions of the variables to this explained variance can not be calculated, but a stepwise search can be conducted similar to that of stepwise regression. First, the single variable that explains the most variance is computed, by trying each one at a time. The amounts of explained variance for each variable are:

PlankCop:	0.0412
PlankClad:	0.0201
InsectAir:	0.0294
InsectLarv:	0.0163
BenthCrust:	0.0241
BenthMussl	0.0285
Others:	0.0139

so that *PlankCop* explains the most. We now perform a permutation test on this explained variance, by permuting the values of *PlankCop* and recomputing the explained variance each time. The *p*-value is estimated at 0.0008, so this is highly significant (see Exhibit 15.7).

The next step is to determine which second variable, when added to *PlankCop*, explains the most variance. The results are:

PlankCop + *PlankClad*:	0.0637
PlankCop + *InsectAir*:	0.0707
PlankCop + *InsectLarv*:	0.0557
PlankCop + *BenthCrust*:	0.0631
PlankCop + *BenthMussl*:	0.0638
PlankCop + *Others*:	0.0535

so that *InsectAir* explains the most additional variance. The permutation test now involves fixing the *PlankCop* variable and permuting the values of *InsectAir*, leading to an estimated *p*-value of 0.0097 (see Exhibit 15.7).

We now continue the stepwise process by looking for a third dietary variable which adds the most explained variance to *PlankCop* and *InsectAir*.

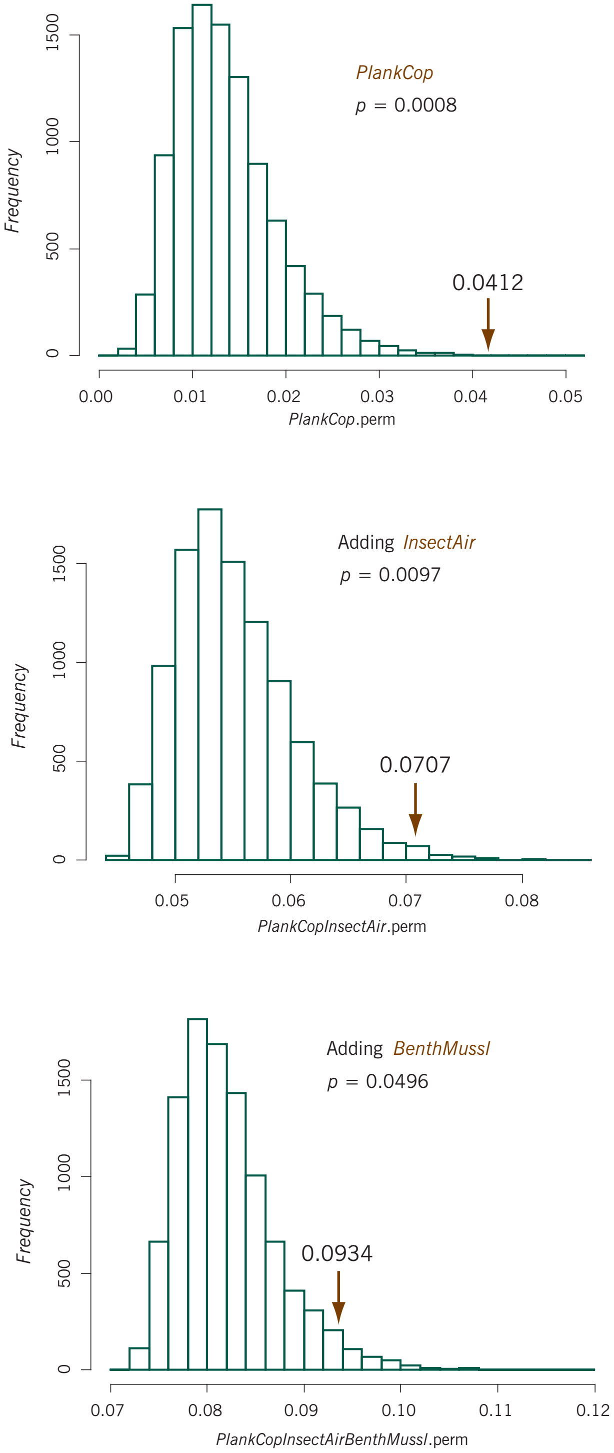
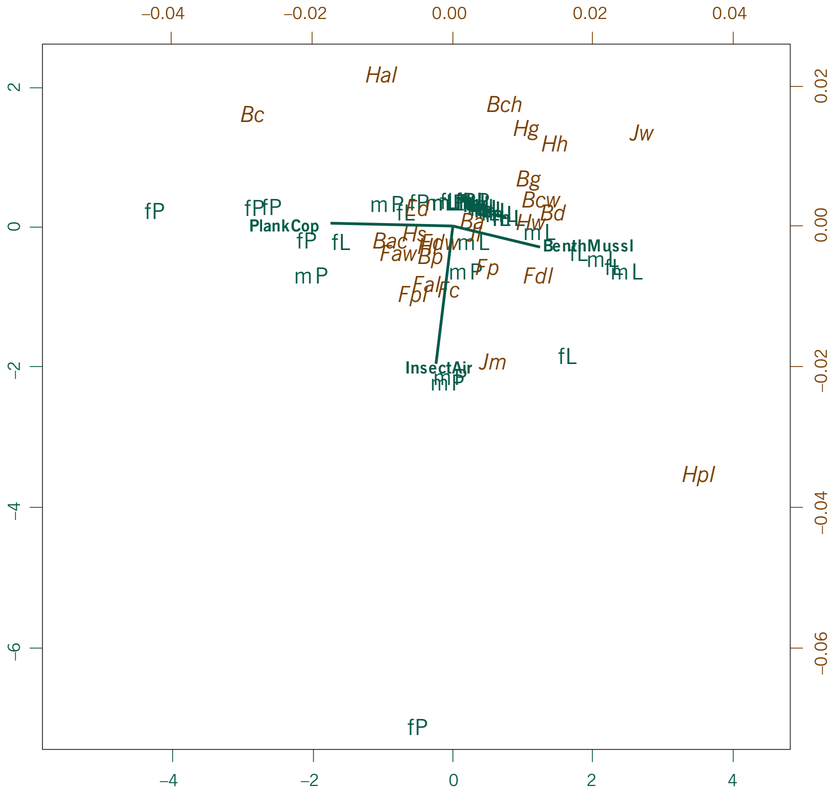


Exhibit 15.7:

Permutation distributions and observed values (explained variances) for the three stages of the stepwise process, introducing successively, from left to right, PlankCop, InsectAir and BenthMussl. The p-values given by the three tests are 0.0008, 0.0097 and 0.0496 respectively

Exhibit 15.8:

Weighted LRA biplot constrained by the three significant fish diet variables, using the same scalings as Exhibit 15.5. 85.8% of the constrained variance is accounted for



PlankCop + *InsectAir* + *PlankClad*:	0.0890
PlankCop + *InsectAir* + *InsectLarv*:	0.0826
PlankCop + *InsectAir* + *BenthCrust*:	0.0862
PlankCop + *InsectAir* + *BenthMussl*:	0.0934
PlankCop + *InsectAir* + *Others*:	0.0827

So the winner is *BenthMussl*. The permutations test fixes *PlankCop* and *InsectAir* and permutes *BenthMussl*, leading to an estimated *p*-value of 0.0496 (see Exhibit 15.7). No other variables enter below the "classical" level of 0.05 and the final canonical LRA, using the three variables *PlankCop*, *InsectAir* and *BenthMussl*, explains a total of 9.15% of the variance of the morphological data. The canonical LRA of the morphological data with just these three significant diet variables is shown in Exhibit 15.8.

Finally, the most highly contributing morphometric variables were identified—there are eight of them, out of the 26—contributing a total of 66% of the iner-

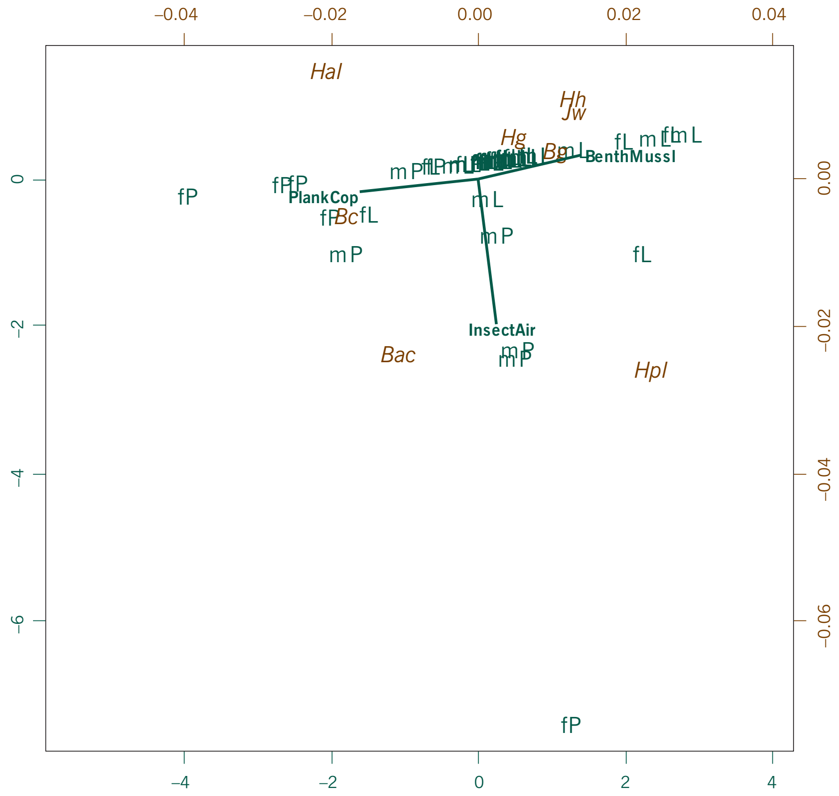


Exhibit 15.9:

Weighted LRA biplot constrained by the three significant fish diet variables, and using only the most highly contributing morphometric variables. The same scalings as Exhibits 15.5 and 15.8 are used for all three sets of points. 94.6% of the constrained variance is accounted for

tia in the constrained biplot. The analysis was repeated from the start, using just these eight variables, and the constrained biplot is shown in Exhibit 15.9. This biplot shows the essential structure in the morphological–diet relationship. The first dimension opposes *PlankCop* against *BenthMussl*, which we already saw in the CA-DA of Exhibit 15.4 was important in the separation of pelagic from littoral groups. The band of fish seen from left to right have zero *InsectAir*, with mostly littoral fish on the right with higher than average benthic mussels in the stomach, also with larger jaw widths and head heights, and mostly pelagic fish on the left feeding on more planktonic cladocerans, and with relatively larger tails (one female littoral fish is also on the left, as in previous biplots, and seems to be an exception in this otherwise pelagic group). *InsectAir* (flying insects) defines a separate perpendicular direction of spread, pulling out a few fish, especially one female pelagic (fP) which was seen to be isolated in previous biplots—this fish has 15% *InsectAir* in its stomach, much higher than any other fish in this data set, and also happens to have one of the highest values of posterior head length (*Hpl*).

SUMMARY

This case study shows how a biplot, specifically the log-ratio biplot in this case, can allow investigation of the patterns in a multivariate data set that are directly related to a set of external variables. The use of permutation tests permits distinguishing the external variables that explain significant variation from the others. Some biological conclusions about the relationship between fish morphology and fish diet are as follows:

1. The fish included in this study are characterized by two distinct forms feeding in different habitats (pelagic *vs* littoral) and on different prey (benthos *vs* plankton).
2. The feeding habits are associated with distinctive morphologies, with fish feeding on benthic crustaceans being more bulky and with greater jaws relative to the more slender plankton eating fish.
3. In the littoral zone males and females display similar diets, whereas in the pelagic males are more oriented towards planktonic cladocerans (waterfleas) and surface insects but females prefer deep dwelling copepods.