Contributions to the Multivariate Analysis of Marine Environmental Monitoring Data: Methodological Aspects and Applications

Doctoral Thesis to be defended by:

Jan Graffelman

under the supervision of:

Tomàs Aluja Banet

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya

Barcelona, September 2000

Voor mijn ouders:

Herminus Graffelman
Hendrika Bussink

Contents

Preface

The idea to write a thesis in the field of multivariate analysis was born during one of the departmental doctorate courses taught by Tomàs Aluja. At about the same time I got to know Michael Greenacre and Reinhold Fieler, who were working on the statistical analysis of a large marine biological database. Joining their project offered me the possibility to combine my background in biology and computing with a growing interest for statistics in general and multivariate analysis in particular.

Several people have contributed to make me enjoy writing this thesis, whether through their direct contributions or by, so to say, improving secondary working conditions. First of all, I'd like to thank Tomàs Aluja for his interest and all his detailed comments on this entire manuscript. His encouragement and enthusiasm have really helped me a lot. Without Reinhold Fieler and Akvaplan-Niva in Trømso this thesis would definitely not have been the same, if it would have been written at all. I'm very grateful to Reinhold for his hospitality, his many emails in which he provided important background information on the data and the sampling, and of course for providing the data on which this thesis is built. I am also indebted to Michael Greenacre for several introductory sessions on correspondence analysis and for his comments on early drafts on some of the chapters of this thesis. I'd also like to thank Robert Gittins for the interest he showed for my work.

During the project some Dutchmen came to visit Barcelona, of who I'd like to mention Emiel Kaper, who I still imagine looking for lost jewelry in the Barcelona drain. I tried to convince him to change to Catalan food, though it seems in vain, as he still seems to stick to Dutch potatoes. Special thanks also goes to ET[1] problem aficionado Michel van de Velden, with whom I jointly solved and published a linear algebra problem (Graffelman and van de Velden, 1999) as described in section 9.2.4. In Catalunya, good second-hand bikes are scarce, and I regret the consequences of this for Michel during a Collserola excursion, especially when we were about to go up from La Rierada to la carretera de Molins de Rei.

Some local colleagues surely deserve to be mentioned for their contributions. Kic Udina occasionally made statements to me about the psychology of thesis writing which were more valuable to me than any automated bandwidth selector

[1] Econometric Theory

could have been, and also turned out to be the best guide for biking excursions around Valldoreix. I also want to express my warm thanks to Albert Satorra for suggesting many interesting references on several occasions.

This thesis does not span my entire research activity over the last few years. For historical reasons, I have been involved with the analysis of human birth data, and some papers got published in this field (Graffelman et al., 1999; Graffelman and Hoekstra, 2000). I was also happy to be invited to talk about these matters at the ESHRE[2] conference in Maastricht in 1997. Indeed, some colleagues occasionally suggested that I might as well have written my thesis in this area, and probably they were right.

A few parts of this thesis have been presented at biometry conferences of which I mention Amsterdam, Cordoba and Mallorca, and all of them bring very good memories. The experiences at the cultural day of the biometry conference in Amsterdam—another biking excursion—did not appear in the conference proceedings, but I hope to see them published, sooner or later, as a novel or a short story by one of the many participants. Finally, I'd like to acknowledge the financial support of DGES grant PB96-0300.

Barcelona, January 2000.

[2]European Society for Human Reproduction and Embryology

Chapter 1

Introduction

It is often maintained that *statistics starts with the data*, and this thesis tries to follow that principle, since it is in large part dedicated to the analysis of (multivariate) data originating from marine environmental monitoring surveys, as well as to aspects of the statistical methodology used in this kind of studies.

Such (expensive) surveys are carried out in order to gain insight in the impact of human industrial activities on biological systems, and their results are, as we hope, to some extent taken into account by authorities as part of their environmental policy.

In Norway, oil companies exploiting platforms in the North Sea are obliged by law to carry out impact studies on a regular basis. Akvaplan-Niva in Trømso is involved in the realization of such studies. Reinhold Fieler of Akvaplan-Niva, involved in the analysis of the data produced by these surveys, kindly provided the data sets used in this thesis.

The data sets obtained in the annually repeated surveys fall broadly into two categories. We have counts of many organisms at various locations (biological data) and measurements of chemicals at the same locations (chemical or environmental data). We are not able to control the level of any of these variables, but merely observe the values they happen to take; data is *observational*, and of *multivariate* nature.

We proceed to give a general outline of this thesis, and at the same time summarize some of the main results. Chapter 2 explains the details of the sampling procedure and provides a univariate analysis of the variables involved. Reliability calculations show that the biological data has in general poor reliability, except for a small group of highly abundant species. The Poisson distribution is the natural candidate for describing the biological abundance, but is seen to be inadequate, except for rare species. After the use of an appropriate transformation, the chemical variables are seen to be approximately normal.

In chapter 3 the problem of finding a particular probability distribution for

species counts is addressed in more detail. The Poisson distribution is often not satisfying, due to many zeros and occasionally high counts. In chapter 3 we try to take the sparseness of the data into account by introducing an extra parameter for the zero outcome. The mathematics of such a zero-inflated Poisson distribution are studied in detail, where we obtain expressions for the expectation and variance of such a distribution, and derive the likelihood equations necessary to estimate the parameters. For most species, the extra parameter for the zero outcome turns out to be statistically significant. A truncated zero-inflated Poisson and mixtures of Poisson distributions are also considered.

The abundance of a species at a certain site is thought to be determined by the physical and chemical characteristics of the environment, though biological factors like competition, cooperation and predator-prey relationships can also play their role. In chapter 4 we start, after some bivariate explorative analysis, to model the survey data with the use of regression models, on a species by species basis, with abundance as the response variable and the chemical data as predictors. Some particular species have been selected for this purpose. Many of the problems that complicate regression analysis are encountered with the survey data: outliers, multicollinearity due to very high correlations between the environmental variables, and violation of the independence assumption due to the fact that repeated observations made at the same site resemble. Though it is hard to generalize, very rare species are probably best modelled by logistic regression, rare species by Poisson regression, and abundant species by random coefficient models. In general, a unimodal response model seems not very apt for the data, as most species display a pattern of decrease with increasing concentrations of the heavy metals.

Treatment of the data on a species by species basis is too elaborate, making it necessary to follow a multivariate approach where all data are used simultaneously. Reciprocal averaging is an algorithm that has been used by ecologists for the analysis of tables of species counts since the seventies, though nowadays the procedure is probably better known under the name of correspondence analysis (CA). Chapter 5 gives a brief review of CA, with attention for some more theoretical details. It provides a new proof for the bounds of the singular values in CA, and also shows that the standard coordinates obtained by CA can be used to construct centring matrices.

Applications of CA to the species data are described in chapter 6. We dedicate some attention to stability issues, and compare different ordinations from different replicates by procrustes rotation. Stacking data matrices from different years into one large matrix allows us to analyze data from different years simultaneously, and gives very well interpretable output. The analysis of the species data is kept separate from the analysis of the chemical data, where for the analysis of the latter we present some results obtained by doing principal component analysis. Chemical changes experienced by the stations are also revealed by an integrated analysis of the combined annual data matrices.

Chapter 7 addresses the problem of the representation of the environmental data as supplementary variables in a biplot obtained by CA. In fact, the representation of a supplementary continuous variable in a CA biplot is a topic of interest

beyond the particular ecological context. Chapter 7 develops some methodology for obtaining optimal directions for supplementary variables in CA. This is done by minimizing projection errors obtained when site coordinates from CA are projected onto supplementary variables. Attention is given to aspects such as the quality of the display of these variables, type of scaling used, relationships with other methods, and the geometrical properties of the solution. It is shown with both real and artificial data that these supplementary variables are of great help in interpreting CA output. The same problem of displaying supplementary variables is also of interest in the context of PCA, and is taken up again in chapter 8, where we develop the same methodology for PCA. If the right type of scaling is used in CA and PCA, the optimal directions for supplementary variables can be obtained by calculating correlation coefficients.

In chapter 7, environmental information is used in an indirect manner, posterior to the analysis of the species data. Canonical Correspondence Analysis (CCA), proposed by Ter Braak (1986), is probably the most popular method for using environmental information in a direct manner. Chapter 9 is a theoretical chapter on CCA, describing how CCA can be obtained by working linear restrictions into the basic CA equations. The chapter also contains many interesting theoretical results, such as bounds obtained for inertias, use of generalized inverses, specification of the trivial dimension, conditional optimality of the representation of the environmental data, and so on. It is shown that CCA can also be performed by doing a principal coordinates analysis of a particular distance matrix. Most important, we find out that CCA does not optimize the representation of species optima, and that the quality statistics in use only resume the quality of the display of the abundance data. Therefore, statistics for the quality of representation of the species optima in CCA are needed and proposed. Quality statistics for the representation of the environmental data are also provided. Biplots in CCA are discussed, and an algorithm for the automated calibration of biplot axes has been developed.

Chapter 10 deals with some applications of CCA to the survey data. CCA reveals the preferences of some of the more abundant species in the survey. A few species are seen to prefer the contaminated conditions. We also do some attempts to reduce the amount of variables, and to partial out spatial effects.

Chapter 11 is an attempt to modify CCA in such a way that it does represent species optima in an optimal way. A weighted principal component analysis of the matrix of weighted averages is seen to be capable to explain more variance of the species optima, and is proposed as an alternative. Samples can be represented in this analysis in a supplementary manner, where one can choose to optimize the representation of the species data or of the environmental data. Artificial data and survey data illustrate this alternative approach, and suggest that the environmental data are also better represented this way.

Some suggestions for further research are commented on in the last chapter, and a selection of the many computer programs used in this thesis are presented in an appendix. Most of the standard types of analysis (regression, anova) were performed with the statistical package STATA, whereas all the multivariate work was done with self-written programs in MATLAB. Finally, this thesis itself was

typeset with the Emtex version of LaTeX on a Pentium PC.

Chapter 2

Sampling & Univariate Aspects

This chapter describes the sampling procedure and the characteristics of the data obtained, and discusses some results of a descriptive univariate analysis of the data.

2.1 The Sampling Procedure

A network of stations has been established in the Norwegian oil field Ekofisk in the North Sea. Geographical maps of the stations are shown in figures 2.1, 2.2 and 2.3. Ekofisk is located west of Stavanger (Norway). All stations are located at a particular distance from a pollution source, an oil platform. The latter is represented by the origin of the three figures. Each station is visited once a year, in May, and eight grab samples are taken at the bottom of the ocean floor of each station (also called "site"). Data from three consecutive years are considered in this thesis: 1990, 1991 and 1992. The station network has undergone some changes from year to year, as the number of stations has been reduced over the years in order to reduce expenses. In 1990 about 40 stations were sampled within a radius of about six kilometers, where the stations form a star-like orientation (see figure 2.1). A more detailed map of the stations close to the platform is shown in figure 2.2. In 1992 most stations visited were within a radius of 2.5 kilometers from the platform (see figure 2.3). A few stations (40,42) are farther away, about 30 kilometers eastward from the platform, and are called "reference" stations, since they are supposed to experience no influence of pollution, and to reflect more "natural" conditions.

A team of specialized biologists analyzes five of the eight grab samples, counting all the animals they find. The animals, more than 200 species, are benthic organisms and consist mainly of worms and molluscs. The other three grab samples are used for chemical analysis, and the concentration of about 13 environmental variables is measured: Total Hydrocarbon Content (THC), Total Organic Content (TOC), Pelite (Pel), heavy metals like Lead (Pb), Zinc (Zn),

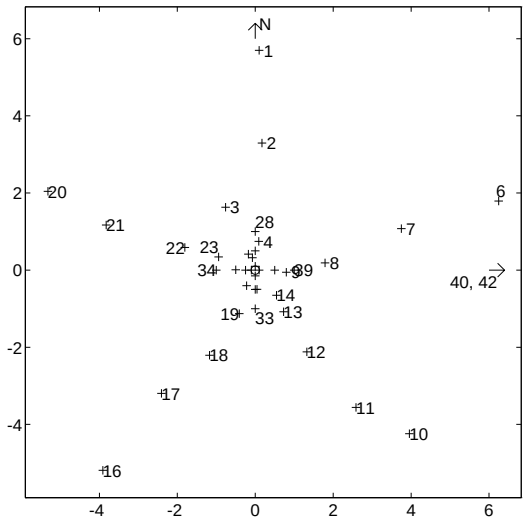

FIGURE 2.1: EKOFISK STATION NETWORK IN 1990

Cadmium (Cd), Copper (Cu), Iron (Fe) and Mercury (Hg), Barium (Ba), Pristane and the ratios n-C17/pristane and n-C18/pristane. Most of these variables were recorded each year. Other variables of potential interest are the distance of each station to the pollution source, temperature and depth. The temperature is not recorded as it is being considered too variable. The depth of all stations in the Ekofisk field is between 67 and 72 meters. The variability in depth is considered irrelevant, as changes in depth of less than 10 meters do not affect the species composition (Reinhold Fieler, personal communication).

We notice here that the chemical sampling is *destructive*; a grab used for chemical analysis cannot be used for biological analysis any more. This is the reason that separate samples are taken for chemical and biological analysis. In later chapters we will want to try to explain species abundance in terms of the chemical variables, for instance by regression. We note here that in such regressions, the chemical measurements of the biological sample are in fact not available, but are estimated from different samples at the same location.

Taking distance apart, we thus have two types of variables, the biological variables and the chemical variables, the latter often also being referred to as environmental variables. A separate section is dedicated to each category.

2.2 The Biological Variables

The biological variables are the species abundances for each year, and consist of counts of species at a series of locations (called stations or sites). Abundance data is known to be bulky, sparse and noisy. (Jongman et al., 1987). Bulky

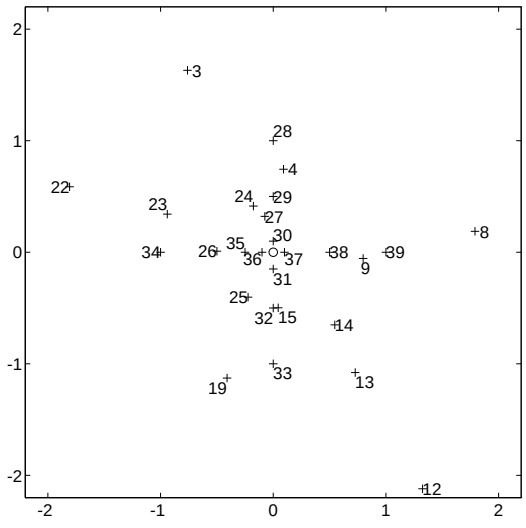

FIGURE 2.2: EKOFISK INNER STATIONS IN 1990

because of the large number of species involved (e.g. 152 in 1990), sparse because of the fact that many species are absent at many locations, and noisy since repeated sampling can produce vastly different values. It is impossible to describe all biological variables one by one, as there are too many. There are species which are highly frequent and others which are absent or rare. A rough indication of this: in 1990 152 species were found; 73 (48%) of these had a total abundance (summing the 5 replicates) in the range 1-10, 52 (34%) of these were in the range 11-100, 23 (15%) in the range 101-1000 and 4 (3%) > 1000.

A few species ranging from highly abundant to rare are selected in order to give an impression of the distribution of the variable abundance. The boxplots of *Amphiura filiformis* (1), *Chaetozone setosa* (2), *Nephtys longosetosa* (3), *Prionospio cirrifera* (4), *Nephtys caeca* (5) and *Jassa marmorata* (6) are shown in figure 2.4 (upper panel). These boxplots illustrate that species abundance tends to be positively skewed, with occasional high outliers, and high probabilities for the lower values (0 in particular). The lower panel of figure 2.4 shows the boxplots of the same species, where the abundance has been transformed by taking the square root. This reduces the positive skew considerably, and symmetrizes the distributions. This transformation will therefore often be applied before any further analysis. To give an impression of the high amount of zero counts, the sparseness of the abundance matrix has been calculated for each year, using only species actually present in at least one of the samples: 1990: 70.7 %, 1991: 59.7 % and 1992: 63.4 % sparse. For individual replicates the degree of sparseness will even be higher.

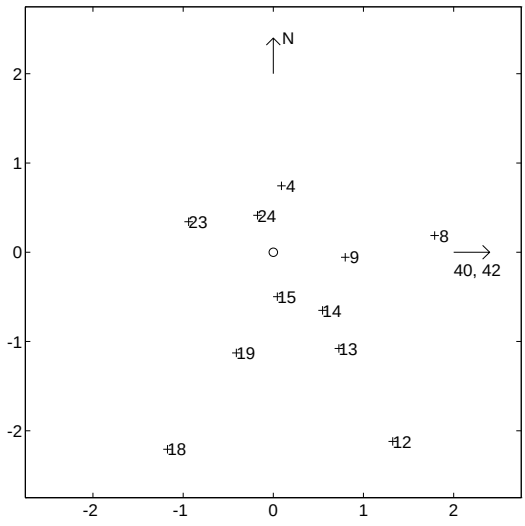

FIGURE 2.3: EKOFISK STATION NETWORK IN 1992

2.2.1 The Reliability of the Biological Data

The fact that we dispose of replicates enables us to calculate reliabilities, also called intraclass correlation coefficients of reliability (Fleiss, 1986, p. 3). An observed value (x) is considered to be the sum of a "signal" plus an error, $x = t + e$, and if the distribution of the errors is independent of signal t, one has that $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$. The intraclass coefficient of reliability (R) is defined as the fraction:

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}. \tag{2.1}$$

Since R is a fraction, theoretically we have that $0 \leq R \leq 1$. In practice however, reliability coefficients are estimated from an analysis of variance. In particular, reliabilities can be calculated as (Fleiss, 1986, p. 11):

$$\hat{R} = \frac{BMS - WMS}{BMS + (k-1)WMS}, \tag{2.2}$$

where BMS and WMS are the "between" and "within" mean sum of squares of the analysis of variance table, and k is the number of replicates. With estimator (2.2) it can occasionally happen that small negative reliabilities are found. In practice, this happens quite frequently with abundance data of rare species (see below). When all replicate measurements coincide with their mean, the WMS term vanishes, and $\hat{R}$ reaches its upper bound of 1. On the other hand, when the means of the replicates at each station coincide with the overall mean of all observations, term BMS in (2.2) vanishes, and $\hat{R}$ achieves a lower bound of $-1/(k-1)$. This in contrast to the ordinary correlation coefficient, which is bounded below by -1. For the data at hand, biological reliabilities are thus

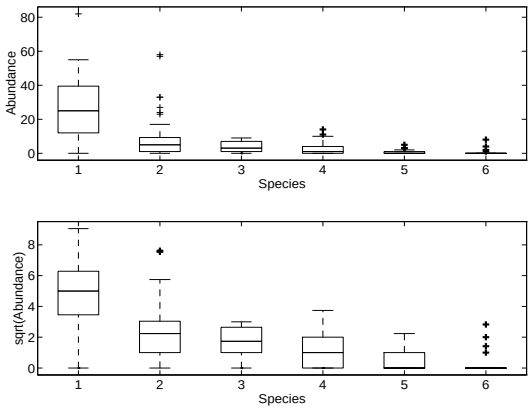

FIGURE 2.4: BOXPLOTS OF ABUNDANCE OF SOME SPECIES

bounded below by -0.25. Some reliability calculations for the species data can also be found in Fieler and Greenacre (1994). Figure 2.5 plots the reliability (indicated by a + sign) of 152 species (abundances transformed by taking square roots) versus the natural logarithm of their total abundance. It is clear that there are many species with a low reliability. 89.5% of the species has a reliability below 0.4, 7.2% of the species has a reliability between 0.4 and 0.75, and 3.3% has a reliability above 0.75. These categories correspond with what Fleiss (1986) calls poor, fair to good and excellent reliability respectively, although, as Fleiss describes, there are no universal standards as to what represents poor or excellent reliability. Figure 2.5 shows that reliability is related to total abundance in the sense that highly frequent species have good to excellent reliability, whereas rare species have poor reliability.

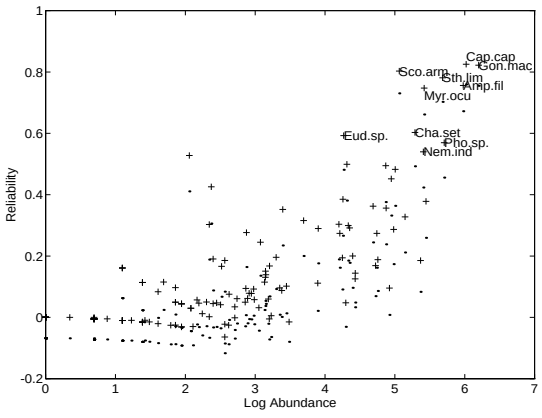

FIGURE 2.5: RELIABILITY OF SPECIES IN 1990

The species with the better reliability are labelled with their abbreviated names

in graph 2.5. These are: *Capitella capitata* (0.83), *Goniada maculata* (0.82), *Scoloplos armiger* (0.80), *Sthenelais limicola* (0.78), *Amphiura filiformis* (0.76), *Myriochele oculata* (0.75), *Chaetozone setosa* (0.60), *Eudorella sp.* (0.59), *Phoronis sp.* (0.57), and *Nemertini indet.* (0.54). These are the species that will be preferentially used in subsequent chapters whenever we try to model species abundance in terms of other variables. As a consequence of the low reliability of the species data, possible correlations between the abundances of different species will be attenuated, and may even be rendered insignificant. Fleiss (1986, p. 12) also gives an expression for an approximate 95% one-sided confidence interval for the reliability. These confidence limits are indicated by a dot for each species in figure 2.5. The reliability of 60% of the species does not differ significantly from zero. This means that for 60% of the species, the differences observed between the stations are due to random measurement error only.

2.2.2 The Distribution of the Biological Variables

What would be an adequate probability distribution to describe species abundance? Counts of phenomena in time or space are often described by a Poisson distribution, and Poisson distributions were fitted to the abundance of some of the species. If the species distributions do follow a Poisson distribution, then the sum of the five replicates should theoretically also follow a Poisson distribution, with a mean that is the sum of the means of the individual replicates. In the first instance, we try to assess whether the sum of five replicates is in agreement with a Poisson distribution. Figure 2.6 shows expected probabilities (open circles) and observed probabilities (plusses) for the six species previously mentioned. This figure shows that if we use the Poisson probability distribution to describe species abundance, we systematically underestimate the amount of zeros, we overestimate the probability of obtaining intermediate values, and we underestimate the outlying higher values. By mere visual inspection, only for the rarer species like *Nephtys caeca* and *Jassa marmorata* the fit of the Poisson distribution seems acceptable.

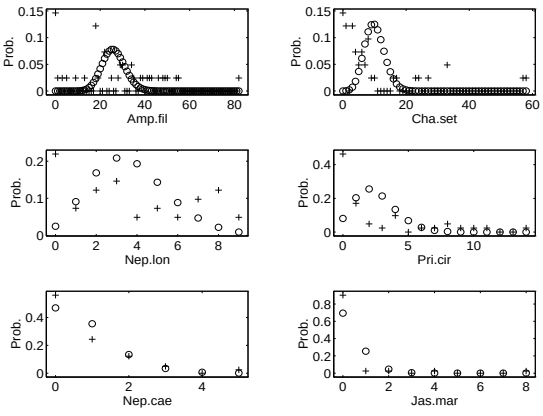

FIGURE 2.6: FIT OF POISSON DISTRIBUTION

We note that summing the five replicates increases the counts. There are however, only 41 samples. Summing the counts then makes that many theoretical outcomes under the Poisson distribution, are in practice never observed in the data. This is especially evident in the graphs of *Amphiura filiformis* and *Chaetozone setosa* above.

The sample means ($\hat{\lambda}$) and sample variances (s^2) of the six selected species are shown in table 2.1. The variance exceeds the mean for all species. For Poisson distributed data the variance theoretically equals the mean, so that the sample variance is higher than expected under a Poisson distribution. Phrased in other words, there is considerable overdispersion.

As a way of assessing whether the abundance of a particular species follows a Poisson distribution, bootstrap resampling was used as described by Noreen (1989, chapter 4), Hamilton (1998, appendix 2) and Manly (1997). The test-statistic considered (T) is the quotient of the sample mean and the sample variance. For data which truly follow a Poisson distribution this statistic is 1. With bootstrapping the theoretical distribution of the test-statistic does not need to be specified, and is in practice often unknown. Using 500 bootstrap samples, a 95% confidence interval for the test statistic was obtained by using the 2.5 and 97.5 percentiles of the bootstrap distribution. If the value of 1 is not included in this interval, the hypothesis that the data follow a Poisson distribution is rejected. In practice, the bootstrap distribution of the test statistic has a mean that does not coincide exactly with the value of T obtained from the original sample. To correct for this bias, the bootstrap distribution can be shifted (Noreen, 1989, chapter 4), so that it is centred on the value of T obtained from the original sample. The test statistic, confidence intervals and the bias for the six species considered are shown in table 2.1, as well as the species' total abundance (N).

Species	N	$\hat{\lambda}$	s^2	T	bias	95% CI
Amp.fil.	1067	26.68	357.15	0.075	0.005	(0.050 - 0.122)
Cha.set.	413	10.33	197.35	0.052	0.009	(0.039 - 0.125)
Nep.lon.	152	3.80	9.14	0.416	0.018	(0.323 - 0.601)
Pri.cir.	103	2.58	13.69	0.188	0.013	(0.149 - 0.287)
Nep.cae.	31	0.78	1.26	0.617	0.071	(0.420 - 1.130)
Jas.mar.	15	0.38	2.04	0.184	0.090	(0.129 - 1.000)

TABLE 2.1: BOOTSTRAP CONFIDENCE INTERVALS OF T

Table 2.1 shows that statistic T is less than one for all species considered. The rarer the species, the wider the confidence interval. The hypothesis that the summed species abundances follow a Poisson distributions must in general be rejected expect for rare species. The bootstrap distributions showed a little bias and positive skew.

Bootstrapping was also applied to a single replicate only, in order to see if individual replicates are in better agreement with a Poisson distribution. The

species	N	$\hat{\lambda}$	s^2	T	bias	95% CI
Amp.fil.	207	5.05	19.30	0.261	0.031	(0.178 - 0.480)
Cha.set.	70	1.71	7.61	0.224	0.079	(0.131 - 0.677)
Nep.lon.	30	0.73	1.40	0.522	0.089	(0.331 - 1.143)
Pri.cir.	29	0.71	3.61	0.196	0.138	(0.123 - 1.116)
Nep.cae.[a]	8	0.20	0.36	0.541	0.086	(0.383 - 1.111)
Jas.mar.[b]	3	0.07	0.22	0.333	0.258	(0.333 - 1.000)

[a] 5 bootstrap samples all zero
[b] 38 % of bootstrap samples all zero

TABLE 2.2: BOOTSTRAP CONFIDENCE INTERVALS OF T, ONE REPLICATE ONLY

results are shown in table 2.2. For the two most abundant species, the Poisson distribution has to be rejected. For species with a total abundance of 30 or lower, the Poisson distribution can, in general, not be rejected. Note that, when we correct the confidence interval of *Jassa marmorata* for bias, the Poisson distribution has to be rejected. For very rare species, bootstrapping becomes problematic, as many bootstrap samples arise that consist only of zeros. For such bootstrap samples the test statistic is not defined. However, a bootstrap sample consisting of zeros only has equal mean and variance, both zero, and this is in perfect agreement with a Poisson distribution. One could therefore argue that these bootstrap samples should be assigned the value $T = 1$, as is done for the two rarest species in table 2.2. The confidence intervals for statistic T are wider when using a single replicate, suggesting that data gets closer to being Poisson distributed as smaller volumes are considered.

From a more formal point of view, one could apply Pearson's χ^2-test for goodness of fit to test the null hypothesis that data are Poisson distributed. However, this requires that the data is grouped into bins with at least 5 observations per bin (Rice, 1995, p. 242). This grouping can be done in many ways, and each grouping will give a different value for the χ^2-statistic. Also, 40 samples is a rather small number to divide over bins with a minimum of 5 counts. A Kolmogorov-Smirnov test for "Poissonness" can neither be applied because the data is discrete. Tests for discrete distributions based on the empirical distribution function (EDF), analogous to the Kolmogorov-Smirnov test, have been described by Stephens (1986) and Pettitt and Stephens (1977), but seem not to be available for the Poisson distribution (Agostino and Stephens, 1986, pp. 176).

The bootstrap test was applied to the whole database of 152 species. For 46% of the species the Poisson distribution had to be rejected, and for 54% it could not be rejected. A separate chapter (3) is dedicated to trying to describe the species distributions more accurately, where we try to take the sparseness of the data into account.

2.3 The Chemical Variables

A total of about 13 chemical variables were measured at each station anually. The variables considered are Total Hydrocarbon Content (THC), Total Organic Content (TOC), the heavy metals Lead (Pb), Zinc (Zn), Cadmium (Cd), Copper (Cu), Mercury (Hg) and Iron (Fe), Barium (Ba), the ratios n-C17/pristane, n-C18/phytane, Pristane and Pelite. Most of these chemicals are related to the drilling process (Reinhold Fieler, personal communication).

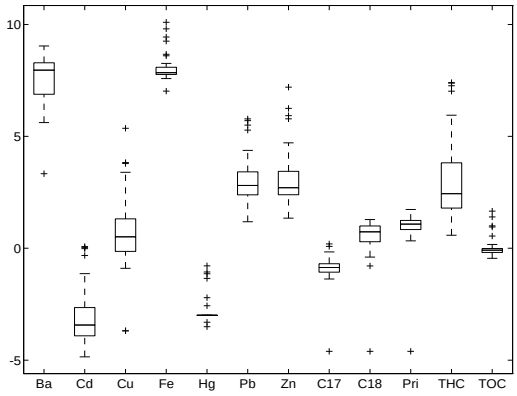

FIGURE 2.7: BOXPLOTS OF LOG-TRANSFORMED CHEMICAL VARIABLES

Barium sulfate is introduced together with other chemicals as a weight component in the drilling fluid that serves to smear the drilling process. The other heavy metals come along with Barium sulfate. Barium is known not to have biological effects but the other heavy metals do. Pristane, a natural component of oil, is an indicator of oil degradation. The ratios n-C17/pristane, n-C18/phytane are used to measure the degree of oil degradation. Pelite is a sedimentological variable, all particles less than 0.063 mm in diameter are called pelite (silt). All chemical variables are measured in milligrams per kilogram (mg/kg) except n-C17/pristane and n-C18/phytane, which are ratios, and TOC which is a percentage. Figure 2.7 shows boxplots of the chemical variables, where the variables have been transformed by taking natural logarithms. Among the heavy metals, Barium and Iron tend to have high concentrations, whereas Cadmium and Mercury have lower concentrations. The logarithmic transformation has considerably symmetrized the distributions of the chemical variables, though some positive skew remains for several variables. It is difficult to find a single transformation that is satisfactory for all the variables simultaneously. Occasionally a zero observation is found among the means of the chemical variables. The natural logarithm of zero is not defined. In order to be able to proceed with the analysis, a small value of 0.01 was assigned to these observations. These recodings pop up as outliers in the boxplots of C17,C18 and Pristane, and correspond to reference station 40, where these components were not detectable. These observations are also outliers in the original scale of measurement, though the arbitrary values assigned will determine how outlying they are in transformed

scale.

Since there are three replicates of each chemical sample, reliability calculations were also performed for the chemical variables. The reliability coefficients ($\hat{R}$) for the variables under study are listed in table 2.3, together with their 95% lower confidence limits.

Chemical	R	95% limit	D	p
Ba	0.93	0.90	0.147	0.300
Cd	0.91	0.87	0.221	0.029
Cu	0.89	0.84	0.169	0.165
Fe	0.96	0.93	0.272	0.004
Hg	0.83	0.75	0.478	0.000
Pb	0.95	0.93	0.217	0.034
Zn	0.96	0.93	0.211	0.042
C17	0.91	0.86	0.225	0.025
C18	0.92	0.88	0.211	0.043
Pri	0.92	0.88	0.277	0.003
THC	0.94	0.91	0.214	0.038
TOC	0.90	0.85	0.324	0.000

TABLE 2.3: RELIABILITIES OF LOG-TRANSFORMED CHEMICAL VARIABLES IN 1990

As is clear from table 2.3, all the chemical variables have excellent reliability. The reliability of Copper and Pristane improved considerably due to the log transformation. It is clear that the chemical data do not suffer so much from measurement error as the biological data. Reliabilities listed in table 2.3 are comparable with values obtained for 1991 and 1992 (Fieler and Greenacre, 1994, p. 13).

What distribution can be used to describe the chemical data, after the logarithmic transformation? The boxplots in figure 2.7 suggest that the transformed chemical variables are probably not far from normality, though the positive skew might make the tails different from those of the normal distribution. Indeed, a formal Kolgomorov-Smirnov test for normality shows that the normality assumption must be rejected for most log transformed variables. Kolgomorov-Smirnov's D-statistics and p-values are listed in the last two columns of table 2.3, and only for Ba and Cu normality can not be rejected. A stronger transformation with a negative power, such as $-x^{-0.25}$ might be employed to further reduce positive skew. The boxplots of the environmental variables transformed by this negative power are shown in figure 2.8.
Normality can now no longer be rejected for Ba, Cd, Pb, Zn and THC (D-statistics not shown). Figures 2.7 and 2.8 illustrate that there is no simple transformation which is satisfactory for all variables. On the other hand, it is not very practical to decide on a different transformation for each variable separately, as the number of variables is quite large. In general the $-x^{-0.25}$ transformation seems more satisfactory than taking natural logarithms, though taking logarithms is the more common statistical practice. Whatever transfor-

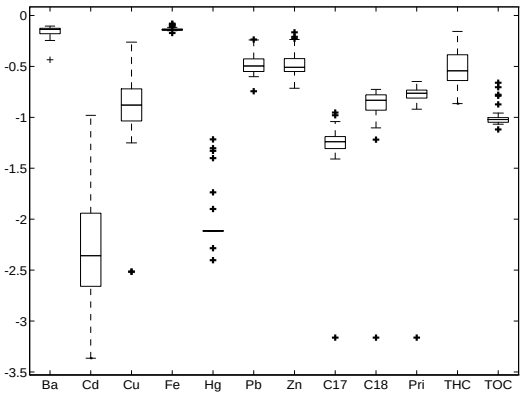

FIGURE 2.8: BOXPLOTS OF $-x^{-1/4}$ TRANSFORMED CHEMICAL VARIABLES

mation we choose, in subsequent chapters on modelling, we can expect a few outliers to cause trouble.

2.4 Total Abundance and Diversity

In order to gain some more basic insight into the database, we study some basic relationships such as the total amount of organisms at each station and the number of different species found at each location in 1990. Figure 2.9, upper left panel, shows the number of species found at each station as a function of their logtransformed distance from the platform. It is clear that the number of species increases as we move away from the platform, and that this increase levels off after a certain distance. The inner ring of stations 30,31,36 and 37, the most close to the platform, have the lowest amount of species. Station 3 is outlying as it is also poor in species content, whereas station 14 has the highest amount of species. Note that many stations in the network have an amount of species that is comparable to the reference station 40.

Graph 2.9, upper right panel, shows the logtransformed total amount of organisms as a function of logtransformed distance. It is striking to see that the same group of stations with few species actually contains the highest amount of organisms. Stations 15 and 24 also stand out for their high total abundance. The high total abundance of the inner ring (30,31,36 and 37) is actually due to one species, *Capitella capitata* which makes up 44% of the total abundance of all organisms 1990. The high abundance of station 24 and 15 is mainly due to *Myriochele oculata*, which makes up more than 9% of the total abundance. Whereas the upper right panel of graph 2.9 suggests that the total amount of organisms decreases and levels off with increasing distance, actually the reverse happens when these two most abundant species are left out of consideration (figure 2.9, lower left panel). Apart from this inner ring, the total amounts of organisms at each station (including the reference station) are roughly of equal

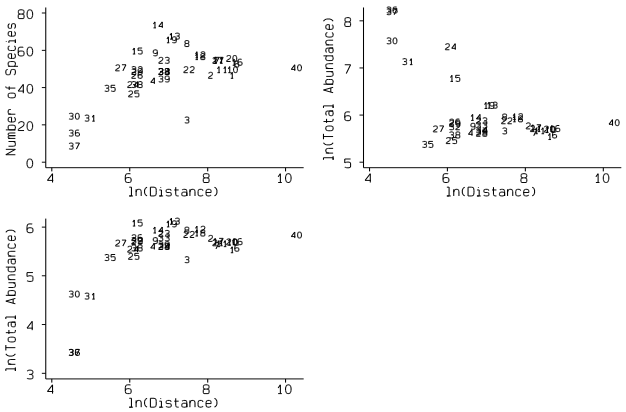

FIGURE 2.9: SPECIES DIVERSITY AND ABUNDANCE IN 1990

order of magnitude, say about 300 organisms on the average.

Chapter 3

The Distribution of Species Abundance

3.1 Introduction

Counts of phenomena in time or space follow, as many elementary textbooks in statistics expose, a Poisson distribution (see Feller (1971) or Rice (1995) for some examples). In ecological applications however, the assumption that counts of species in an area or volume follow a Poisson distribution is often not satisfied (Jongman et al., 1987, pp. 19-20). Count data of organisms in space often consists mainly of zeros, data being extremely sparse. On the other hand, very high counts are sometimes recorded due to clustering of organisms. These two phenomena contribute to overdispersion: the variance of the counts is larger than the mean, whereas for the Poisson distribution sample variance and mean are theoretically equal. Thus, a statistical problem in ecology is to decide upon a particular distribution for species abundance. The data discussed in the previous chapter confirm this picture. In this chapter we continue to adhere to the Poisson probability distribution for describing species abundance, but try to account for the sparse nature of the data in three ways: (i) by using a Poisson distribution and allowing for extra zeros ("zero inflation", (Sørensen, 1999)) (ii) by using a truncated Poisson distribution, without the zero outcome, but with zero inflation, and (iii) by using mixtures of two Poisson distributions. In the next section we derive the maximum likelihood equations for the three different regimes.

3.2 Maximum Likelihood Estimation

The Poisson probability distribution is given by the formula:

$$p(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, \ldots$$

The mean and the variance of a variable with a Poisson distribution are both λ, and it is easily shown that the maximum likelihood (ML) estimator for parameter λ is given by the sample mean. We try to account for overdispersion in three different ways. We derive the likelihood equations for each regime in separate sections below.

3.2.1 A Poisson with Zero Inflation

In this regime a surplus of zeros is accommodated for the in the following way: we imagine that the counts follow in principle a Poisson distribution, but that there is an additional chance θ to obtain a zero. This is why it is called a "zero-inflated" Poisson. A random sample of size N from a zero-inflated Poisson can be generated as follows: take θN zeros and add $N(1-\theta)$ observations from a Poisson distribution with parameter λ. The probability distribution of a zero-inflated Poisson random variable (X) is given by:

$$p(x;\lambda,\theta) = \begin{cases} \theta + (1-\theta)e^{-\lambda} & \text{if } x = 0 \\ (1-\theta)e^{-\lambda}\lambda^x/x! & \text{if } x = 1,2,\ldots \end{cases} \tag{3.1}$$

It is straightforward to show that:

$$E(X) = \sum_{x=1}^{\infty} x(1-\theta)e^{-\lambda}\lambda^x/x! = (1-\theta)\lambda.$$

Naturally, if there are no extra zeros ($\theta = 0$) the expectation $E(X)$ is just that of an ordinary Poisson random variable. The variance of X is found to be:

$$V(X) = E(X^2) - (E(X))^2 = (1-\theta)(1+\theta\lambda)\lambda.$$

Clearly, if $\theta = 0$ the variance is also that of an ordinary Poisson, and indeed, (3.1) reduces to the Poisson frequency function. We notice that for the zero-inflated Poisson the ratio of expectation and variance is no longer a constant, but is given by $1/(1+\theta\lambda)$, this in contrast to the ordinary Poisson. We introduce an indicator variable I_i taking value 0 if the count is 0 ($x_i = 0$) and 1 for nonzero counts ($x_i \neq 0$). The likelihood function is then given by:

$$L(\lambda,\theta) = \prod_{i=1}^{N} \left(I_i(1-\theta)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!} + (1-I_i)(\theta + (1-\theta)e^{-\lambda}) \right). \tag{3.2}$$

Taking natural logarithms, and defining D as the number of zero observations the loglikelihood function becomes:

$$l(\lambda,\theta) = \ln(L(\lambda,\theta)) = \sum_{i=1}^{N-D} \ln\left((1-\theta)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!} \right) + D\ln\left(\theta + (1-\theta)e^{-\lambda}\right), \tag{3.3}$$

where i in the first summation indexes the non-zero counts only. Setting first-order derivatives to zero, we obtain from $\partial l/\partial\theta = 0$:

$$\theta = \frac{D/N - e^{-\lambda}}{1 - e^{-\lambda}}. \tag{3.4}$$

This shows that for large λ, θ approximates the fraction of zero counts in the data. It is clear that parameter θ is a function of parameter λ. From $\partial l/\partial\lambda = 0$ we obtain after some algebra:

$$D - N + (1/\lambda)\sum_{i=1}^{N-D} x_i = \frac{D(1-\theta)}{\theta e^{\lambda} + 1 - \theta}. \tag{3.5}$$

Substituting (3.4) in (3.5), we obtain an expression in one parameter only:

$$\lambda\frac{e^{\lambda}}{e^{\lambda}-1} = \frac{\sum_{i=1}^{N-D} x_i}{N-D}.$$

This shows that if λ is large, λ will approximate the mean of the non-zero counts in the sample. At this point it seems not possible to derive explicit expressions for the ML estimates of λ or θ separately in terms of the data only. As a consequence, the ML estimates for λ and θ need to be obtained numerically, by maximizing (3.3) using for instance a Newton-Raphson algorithm. With the Newton-Raphson method, the maximum of the loglikelihood function can be found iteratively (Dobson, 1991, chapter 4). To do so, we need to obtain from (3.3) the vector of first order derivatives (the score vector $\mathbf{u} = [\partial l/\partial\theta, \partial l/\partial\lambda]$) and the 2 by 2 matrix of second order derivatives:

$$\mathbf{H} = \begin{bmatrix} \partial^2 l/\partial\theta^2 & \partial^2 l/\partial\theta\partial\lambda \\ \partial^2 l/\partial\lambda\partial\theta & \partial^2 l/\partial\lambda^2 \end{bmatrix}, \tag{3.6}$$

where $-E(\mathbf{H})$ is known as the information matrix. The k^{th} approximation of the parameter vector $\mathbf{b} = [\theta, \lambda]$ is then given by:

$$\mathbf{b}^{(k)} = \mathbf{b}^{(k-1)} - \mathbf{H}^{-1}\mathbf{u}^{(k-1)}. \tag{3.7}$$

We need a vector of initial estimates, $\mathbf{b}^{(0)}$. For $\mathbf{b}^{(0)}$ one can take a vector containing for instance the fraction of zeros in the sample and the mean of the (non-zero) counts of the sample. When we define the quantities $S = 1/(e^{\lambda}-1)$ and $Q = \theta/(1-\theta)$, then for the zero-inflated Poisson, the score vector $\mathbf{u}$ and the Hessian $\mathbf{H}$ are given by:

$$\mathbf{u} = \left[\frac{D-N}{1-\theta} + \frac{D}{\theta+S}; D - N + \frac{\sum x_i}{\lambda} - \frac{D}{Qe^{\lambda}+1}\right]$$

and

$$\mathbf{H} = \begin{bmatrix} \frac{D-N}{(1-\theta)^2} - \frac{D}{(\theta+S)^2} & \frac{De^{\lambda}}{(\theta e^{\lambda}+1-\theta)^2} \\ \frac{De^{\lambda}}{(\theta e^{\lambda}+1-\theta)^2} & -\frac{\sum x_i}{\lambda^2} + \frac{DQe^{\lambda}}{(Qe^{\lambda}+1)^2} \end{bmatrix},$$

In section 3.3 we give a numerical example of this algorithm. We note that it is also possible to derive estimators for λ and θ by the method of moments (MOM, (Rice, 1995, section 8.4)). For the zero-inflated Poisson these estimators were derived: $\hat{\lambda} = \sigma^2/(\overline{x}) + \overline{x} - 1$ and $\hat{\theta} = (\overline{x} - \sigma^2)/(\overline{x} - \sigma^2 - \overline{x}^2)$, and are seen to collapse to the sample mean and zero respectively, when data is truly Poisson ($\sigma^2 = \overline{x}$). These estimators have the advantage that they can be calculated straight from the sample mean and the sample variance. These estimators are not considered any further, as maximum likelihood estimators are in general more precise (Rice, 1995, section 8.5). They can however, still be used as initial estimates for the numerical maximization.

3.2.2 A Truncated Poisson with Zero Inflation

We consider a truncated Poisson distribution without the zero outcome. Since the probability of obtaining a zero under the Poisson distribution is $e^{-\lambda}$, the remaining non-zero outcomes sum to $1-e^{-\lambda}$, so that the probability frequency function of a truncated Poisson can be described by:

$$p(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!(1-e^{-\lambda})} \qquad x = 1,\ldots$$

We now consider a regime where we obtain zero counts with probability θ, and non-zero observations come from a truncated Poisson distribution. This can be imagined as drawing balls from two urns. The first urn only contains zeros and ones, obeying a Bernoulli probability distribution, and the second urn contains non-zero integers corresponding to a truncated Poisson. If we draw a zero from the first urn, we write it down. If we draw a one, we take a ball from the second urn and write down its number. Random data from such a scheme can be generated as follows: again we take θN zeros. Next we add $N(1-\theta)$ non-zero observations from a Poisson distribution with parameter λ. From a practical point of view, this means that a sample larger than $N(1-\theta)$ must be drawn from an ordinary Poisson to reach the required number of non-zeros. The probability distribution of the truncated Poisson random variable with an additional probability θ of obtaining a zero is given by:

$$p(x;\lambda,\theta) = \begin{cases} \theta & \text{if } x = 0 \\ (1-\theta)\frac{e^{-\lambda}\lambda^x}{x!(1-e^{-\lambda})} & \text{if } x = 1,2,\ldots \end{cases}$$

It is straightforward to show that the expectation of such a random variable (X) is:

$$E(X) = \frac{\lambda(1-\theta)}{1-e^{-\lambda}},$$

and with some algebra we obtain the variance as:

$$V(X) = \frac{\lambda(1-\theta)\{(\lambda+1)(1-e^{-\lambda}) - \lambda(1-\theta)\}}{(1-e^{-\lambda})^2}.$$

Again using an indicator variable I_i, the likelihood function is given by:

$$L(\lambda,\theta) = \prod_{i=1}^{N}\left(I_i(1-\theta)\frac{\lambda^{x_i}e^{-\lambda}}{x_i!(1-e^{-\lambda})} + (1-I_i)\theta\right). \tag{3.8}$$

and the log-likelihood becomes:

$$l(\lambda,\theta) = \ln L(\lambda,\theta) = \sum_{i=1}^{N-D}(x_i \ln\lambda - \ln x_i!) + (N-D)\ln\frac{(1-\theta)e^{-\lambda}}{1-e^{-\lambda}} + D\ln\theta, \tag{3.9}$$

Setting $\partial l/\partial\theta = \partial l/\partial\lambda = 0$ it can be shown that:

$$\theta = D/N,$$

and

$$\frac{\lambda}{1-e^{-\lambda}} = \frac{1}{N-D}\sum_{i=1}^{N-D} x_i.$$

It turns out that θ is just the fraction of zero counts, and so θ can be obtained directly from the data. There is no explicit expression for λ in terms of the data, λ must be inferred from a graph of function $\frac{\lambda}{1-e^{-\lambda}}$, or be obtained by the numerical optimization of (3.9). In contrast to the zero inflated ordinary Poisson, here parameters θ and λ are not related to each other.

3.2.3 A Mixture of Two Poissons

We consider a mixture of two Poisson distributions with different means λ_1 and λ_2. A mixture coefficient α ($\alpha \ \epsilon \ [0,1]$) indicates the probability that an observation comes from the first distribution with λ_1, and so $(1-\alpha)$ is the probability that an observation comes from the second Poisson distribution with parameter λ_2. A random sample from a mixture can be generated by creating a Bernoulli random variable B, with a probability of success α and generating two Poisson random variables, P_1 and P_2 with parameters λ_1 and λ_2 respectively. The mixture of two Poissons is then calculated as $B * P_1 + (1-B) * P_2$. The probability distribution of a mixture of two Poissons becomes:

$$q(x;\lambda_1,\lambda_2,\alpha) = \alpha p_1(x) + (1-\alpha)p_2(x).$$

It is easily shown that the expectation of a random variable X following a mixture of two Poissons is given by:

$$E(X) = \alpha\lambda_1 + (1-\alpha)\lambda_2,$$

whereas the variance is:

$$V(X) = \alpha\lambda_1(\lambda_1+1) + (1-\alpha)\lambda_2(\lambda_2+1) - \{\alpha\lambda_1 + (1-\alpha)\lambda_2\}^2.$$

The likelihood function is given by:

$$\begin{aligned} L(\lambda_1,\lambda_2,\alpha) &= q(x_1,x_2,\ldots,x_n;\lambda_1,\lambda_2,\alpha) \\ &= \prod_{i=1}^{N} q(x_i;\lambda_1,\lambda_2,\alpha) \\ &= \prod_{i=1}^{N}\left(\frac{\alpha{\lambda_1}^{x_i}e^{-\lambda_1}}{x_i!} + \frac{(1-\alpha){\lambda_2}^{x_i}e^{-\lambda_2}}{x_i!}\right). \end{aligned}$$

Taking natural logarithms, the loglikelihood function becomes:

$$\ln L(\lambda_1,\lambda_2,\alpha) = \sum_{i=1}^{N}\ln\left(\frac{\alpha{\lambda_1}^{x_i}e^{-\lambda_1}}{x_i!} + \frac{(1-\alpha){\lambda_2}^{x_i}e^{-\lambda_2}}{x_i!}\right). \qquad (3.10)$$

Setting first order derivatives to zero, we obtain the set of equations:

$$\begin{aligned}
\partial l/\partial\alpha &= \sum_{i=1}^{N} \frac{\lambda_1^{x_i} e^{-\lambda_1} + \lambda_2^{x_i} e^{-\lambda_2}}{\alpha\lambda_1^{x_i} e^{-\lambda_1} + (1-\alpha)\lambda_2^{x_i} e^{-\lambda_2}} = 0, \\
\partial l/\partial\lambda_1 &= \sum_{i=1}^{N} \frac{\alpha e^{-\lambda_1}(x_i\lambda_1^{x_i-1} - \lambda_1^{x_i})}{\alpha\lambda_1^{x_i} e^{-\lambda_1} + (1-\alpha)\lambda_2^{x_i} e^{-\lambda_2}} = 0, \\
\partial l/\partial\lambda_2 &= \sum_{i=1}^{N} \frac{(1-\alpha) e^{-\lambda_2}(x_i\lambda_2^{x_i-1} - \lambda_2^{x_i})}{\alpha\lambda_1^{x_i} e^{-\lambda_1} + (1-\alpha)\lambda_2^{x_i} e^{-\lambda_2}} = 0.
\end{aligned}$$

Again, it seems not possible to solve these equations in closed form for parameters α, λ_1 and λ_2. We thus proceed by numerically maximizing (3.10). Because $x_i!$ is a constant factor in (3.10), in practice we will try to maximize:

$$\ln L(\lambda_1, \lambda_2, \alpha) = \sum_{i=1}^{N} \ln\left(\alpha {\lambda_1}^{x_i} e^{-\lambda_1} + (1-\alpha){\lambda_2}^{x_i} e^{-\lambda_2}\right),$$

which is a function of three parameters. Some practical issues of the estimation of mixtures are described in a recent review by Haughton (1997).

We note that the previously considered zero-inflated Poisson is in fact a particular case of a mixture of two Poissons. Consider the frequency function of the mixture:

$$q(x; \lambda_1, \lambda_2, \alpha) = \alpha\frac{\lambda_1^x e^{-\lambda_1}}{x!} + (1-\alpha)\frac{\lambda_2^x e^{-\lambda_2}}{x!}.$$

If $\lambda_1 = 0$ this reduces to:

$$p(x; \lambda_1, \alpha) = \begin{cases} \alpha + (1-\alpha)e^{-\lambda_2} & \text{if } x = 0 \\ (1-\alpha)e^{-\lambda_2}\lambda_2^x/x! & \text{if } x = 1, 2, \ldots \end{cases} \tag{3.11}$$

This is precisely the frequency function of a zero-inflated Poisson with parameters (α, λ_2), where mixture-coefficient α represents the additional chance of obtaining a zero (cf. 3.1). In the same way, if $\lambda_2 = 0$ a zero-inflated Poisson arises with parameters $(1-\alpha, \lambda_1)$, and the additional probability of a zero is can be calculated from the mixture coefficient as $1-\alpha$. From a computational point of view, a program that maximizes the likelihood function of a mixture thus provides a general tool if it allows parameters to be fixed. The Poisson ($\alpha = 0$ or $\alpha = 1$) and the zero-inflated Poisson ($\lambda_1 = 0$ or $\lambda_2 = 0$) can then be estimated as special cases of the mixture.

3.3 Application to Species Count Data

In this section we show a detailed example of an application of the three regimes described above, using abundance data of one particular species, *Nephtys caeca*, and we present some results of how these regimes do in general for all species. We use 41 benthic samples of this species taken in 1990. The sample mean is 0.78 and the sample variance is 1.26, indicating that there is overdispersion.

Regime	Parameters			
	λ_1	λ_2	θ	α
Poisson	0.78 (0.14)	-	-	-
Poisson + 0	1.21 (0.31)	-	0.36 (0.14)	-
T. Poisson + 0	1.21 (0.31)	-	0.55 (0.08)	-
Mixture	1.75 (1.16)	0.33 (0.40)	-	0.31 (0.43)

TABLE 3.1: ML-ESTIMATES FOR DIFFERENT REGIMES

Table 3.1 lists estimates for the parameters of the different regimes with their standard errors in parentheses. Notice that the mixture coefficient α is not significantly different from 1, suggesting us that the data is better described by single Poisson distribution. The optimization routine used for maximizing the log-likelihood function (routine ML from STATA version 5.0) frequently had convergence problems when maximizing the objective functions (3.3) or (3.9), giving 'infeasible steps'. These problems were resolved by reparametrizing θ by its logit, $\ln(\theta/(1-\theta))$, and λ by $\ln(\lambda)$. Standard errors for the original parameters can then be obtained using the delta method (Dunn, 1989).

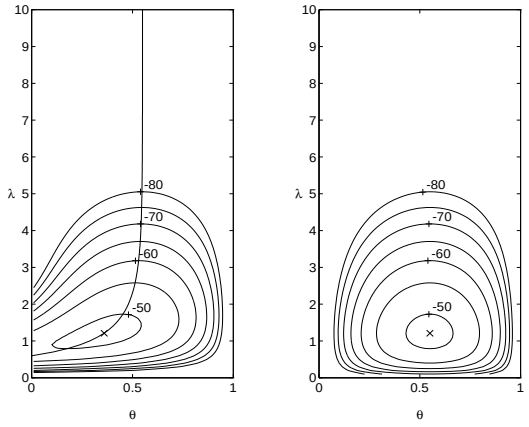

FIGURE 3.1: LOG-LIKELIHOOD LEVEL CURVES FOR *Nephtys caeca*. Left panel shows levelcurves for a zero-inflated Poisson, right panel shows levelcurves for a truncated zero-inflated Poisson.

Figure 3.1 shows the level-curves of the loglikelihood function (3.3) of a zero-inflated Poisson (left graph) for *Nephtys caeca*, and the level-curves of the log-likelihood function (3.9) of a truncated zero-inflated Poisson (right graph). The crosses indicate the optima ((0.36,1.21) and (0.55,1.21) respectively). The likelihood function is seen to be steeper for the smaller values of λ and when θ is close to its boundaries. In the case of the zero-inflated Poisson, the two parameters θ and λ are related (see equation (3.4)), and their relation is indicated by a curve superimposed on the left graph. That the parameters are related is also indicated by the somewhat inclined principal axes of the ellipses in the

left graph, a feature that is absent in the right graph. Whatever data we have, under the zero-inflated Poisson, the optimal pair $(\hat{\theta}, \hat{\lambda})$ should always be on this curve. As is logical, under the truncated regime the estimates of θ and λ are larger then under the non-truncated regime. Parameter θ differs significantly from zero under both zero-inflated regimes, suggesting that it makes sense to include this extra parameter to account for sparseness.

Table 3.2 illustrates the iteration history of the Newton-Raphson algorithm given by (3.7). As initial estimates we use for θ the fraction of zeros in the sample, 0.55, and for λ we take the sample mean, 0.775. The algorithm converges in about 5 iterations. At convergence, the inverse of the information matrix is [0.0199 0.0265; 0.0265 0.0969]. By taking the square root of the diagonal elements, we obtain the standard errors for θ and λ, $\sqrt{0.0199} \approx 0.14$ and $\sqrt{0.0969} \approx 0.31$. This all coincides well with the reported estimates in table 3.1, where we employed optimization routine ML from statistical package STATA. With this ML routine it is sufficient to supply a function that calculates the contribution of one case to the likelihood function. Routine ML has the additional advantage that the parameters can be constrained to be within a certain range. An illustrative program is given in appendix A.1. A simple program using (3.7) can go astray, especially if we are close to the boundaries of the parameter space. The inverse of the information matrix at convergence allows us to calculate the correlation between the parameters as $0.0265/(\sqrt{0.0199}\sqrt{0.0969}) \approx 0.60$. This positive correlation is consistent with the observed positively inclined principal axis of the level curves in figure 3.1 (left panel). At convergence the information matrix turned out to be positive definite, meaning that the Hessian in (3.6) is negative definite, and that the solution corresponds to a maximum.

Iteration	θ	λ	$\ln(L)$
0	0.550	0.775	-54.177
1	0.354	1.040	-49.101
2	0.347	1.169	-48.867
3	0.358	1.206	-48.859
4	0.358	1.207	-48.859
5	0.358	1.207	-48.859

TABLE 3.2: ITERATION HISTORY FOR ML ESTIMATES OF A ZERO-INFLATED POISSON

3.3.1 General Results

The bootstrap test described in chapter 2 was carried out for one particular replicate of all species. By choosing one replicate, several species turned out to contain zero counts only, leaving 112 species with at least one nonzero count. For 70% of these species, bootstrap samples arise which consist only of zeros. Such samples were assigned $T = 1$ (cf. chapter 2 p. 12). For 40% of the species, the Poisson distribution had to be rejected, and for 60% it could not be rejected. In order to see if the zero-inflated regimes and mixtures considered for *Nephtys caeca* in the previous section are useful in general, calculations are repeated for a subset of the species of varying total abundance.

	Poisson		Zero-inflated Poisson		Trunc. Zero-inf. Poisson	
Species	N	$\hat{\lambda}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\lambda}$	$\hat{\theta}$
Amp.fil.	1067	26.68	30.49 (0.93)	0.13 (0.05)	30.49 (0.93)	0.13 (0.05)
Cha.set.	413	10.33	11.80 (0.58)	0.12 (0.05)	11.80 (0.58)	0.13 (0.05)
Nep.lon.	152	3.80	4.71 (0.39)	0.19 (0.06)	4.71 (0.39)	0.20 (0.06)
Pri.cir.	103	2.58	4.64 (0.47)	0.44 (0.08)	4.64 (0.47)	0.45 (0.08)
Nep.cae.	31	0.78	1.21 (0.31)	0.36 (0.14)	1.21 (0.31)	0.55 (0.08)
Jas.mar.	15	0.38	3.65 (0.99)	0.90 (0.05)	3.65 (0.99)	0.90 (0.05)

TABLE 3.3: PARAMETER ESTIMATES FOR THE ZERO-INFLATED POISSON AND THE TRUNCATED ZERO-INFLATED POISSON

Table 3.3 shows that for the six selected species the zero inflated regime has a significant parameter θ. For *Jassa marmorata* the parameter is not significantly different from 1, suggesting data consists of zeros only. When we fit a zero inflated Poisson to all 112 species, about 48% has a parameter θ that differs significantly from zero, though in 16% percent of the species, it is not statistically different from 1. In about 49% of the species, there are problems of convergence. These cases correspond to samples that consist almost entirely of zeros, where we are close to the boundary of the parameter space $\theta = 1$. Only in about 3% of the species, parameter θ does not differ significantly from 0. Table 3.4 shows the estimates of the parameters of a mixture of two Poisson distributions (columns 3 to 5). The mixture coefficient α in table 3.4 applies to the Poisson distribution with λ_1 (the fourth column). For the most abundant species the mixture coefficient is significantly different from zero (and also from 1). In estimating a mixture we typically obtain one λ smaller than the mean of the ordinary Poisson and a second λ larger than the latter. For the *Nephtys caeca* the mixture coefficient is not significant, and the λ's do not differ significantly from zero. For *Nephtys longosetosa* λ_2 does not differ significantly from zero, and therefore one would re-estimate a zero-inflated Poisson. Note that *Jassa marmorata* and *Nephtys caeca*, with insignificant mixture coefficients, were precisely species for which a Poisson distribution could not be discarded (see page 11, table 2.1).

Species	N	λ	λ_1	λ_2	α
Amp.fil.	1067	26.68	33.84 (1.05)	2.00 (0.48)	0.76 (0.07)
Cha.set.	413	10.33	32.04 (1.91)	4.04 (0.37)	0.22 (0.07)
Nep.lon.	152	3.80	5.49 (0.85)	0.79 (0.69)	0.64 (0.16)
Pri.cir.	103	2.58	7.01 (0.82)	0.42 (0.15)	0.33 (0.08)
Nep.cae.	31	0.78	1.75 (1.16)	0.33 (0.40)	0.31 (0.43)
Jas.mar.	15	0.38	4.45 (1.35)	0.02 (0.03)	0.08 (0.05)

TABLE 3.4: PARAMETER ESTIMATES FOR POISSON MIXTURES

In general, we conclude that for frequent to moderately abundant species Poisson mixtures are useful for describing the summed abundances. The zero inflated Poisson distribution does well for many of the species considered, in table 3.3

parameter θ is always significantly different from zero. However, for very rare species θ is close to one, usually not significantly different from 1, suggesting all data are zero. For these cases, the Poisson distribution remains probably more adequate.

Chapter 4

Some Regression Models

4.1 Introduction

The physical, chemical and biological characteristics of the environment determine which species can survive in a certain environment. In many ecological studies, the probability of finding a particular species at a certain location is assumed to depend on environmental variables such as temperature, humidity, pH, etc. It is thus natural to think of regressing species abundance onto environmental variables, in an attempt to explain variations in abundance in terms of environmental variables. In this chapter some results of such regression analyses are presented. Some attention is paid to the particular nature of the data under consideration. First, as explained in chapter 2 when aspects of sampling were considered, the data consists of five repeated measurements of species counts, and three repeated measurements of the chemical variables at each location. When one takes repeated measurements at the same location, it is to be expected that the measurements made at the same location will be more similar to each other, than observations made at different locations. Observations might thus not be entirely independent. Another point to have in mind is that the response variable tends to be sparse and might not be normally distributed. In fact the response variable is discrete and non-negative, because it is a count variable. It is thus natural to consider alternatives to ordinary regression that account for these characteristics. In particular, variance components models are considered when working with repeated measurements, and Poisson regression is used as an alternative when dealing with count data.

An important theoretical reference point in ecological studies of the dependence of abundance on environmental variables is the unimodal response model. Often species are supposed to respond to an environmental variable in a unimodal way: abundance increases over a certain range of the environmental variable, reaches a maximum, and then starts to decrease for higher values of the environmental variable. An example of a unimodal response curve is shown in figure 4.1, where the curve depicted is a Gaussian curve, and the response is symmetric around an optimum (O), with a certain spread around the optimum,

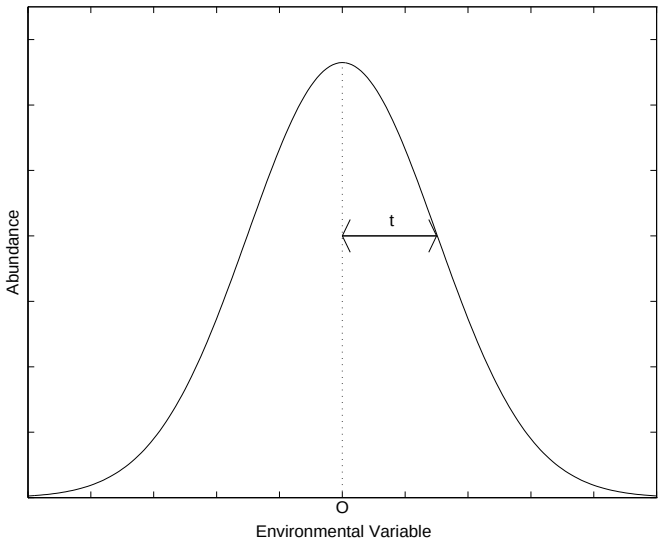

FIGURE 4.1: A UNIMODAL RESPONSE MODEL

called the tolerance (t) of the species. Note that the Gaussian curve in figure 4.1 is conceived as a *response function*, not to be confused with a probability density. Attention is however, not necessarily restricted to symmetric or gaussian curves, parabolas or nonsymmetric curves could do as unimodal response curves as well. Ordinary regression analysis amounts to fitting straight lines to data, not curves like in figure 4.1. However, due to the transformations applied to the data (square root transformation, natural logarithm) and inclusion of quadratic terms, typically curvilinear models are fitted to the original data, though they do not necessarily all correspond to unimodal models. The nature of the fitted models (linear, convex, unimodal, etc) in the applications below will always be indicated.

Naturally, regression theory will not be described in detail here, as there are excellent books on the topic (Draper and Smith, 1981; Hamilton, 1992). However, because of the special relevance of variance components models for repeated measurements and of Poisson regression for rare species (see applications below), brief theoretical accounts of these are interspersed with the applications below.

The outline of the remaining part of this chapter is as follows. First, a separate section is dedicated to an exploratory data analysis of the bivariate relationships in the data from 1990. Next, we treat some applications of ordinary regression models and some of the alternatives to the data of three selected species in particular. The chapter closes with some conclusions and general remarks.

4.2 Descriptive Bivariate Analysis

Before doing any regression, we first explore the bivariate relationships in the data with the aid of some scatterplots matrices in order to get an impression of the associations between the different variables. The data used is from 1990, because this year has the largest number of samples (39). Three different scatterplot matrices are constructed: a scatterplot matrix of the ten species abundances with the best reliabilities (figure 4.2), a scatterplot matrix of all environmental variables (figure 4.3) and a between set scatterplot matrix plotting the ten selected species versus ten selected environmental variables (figure 4.4).

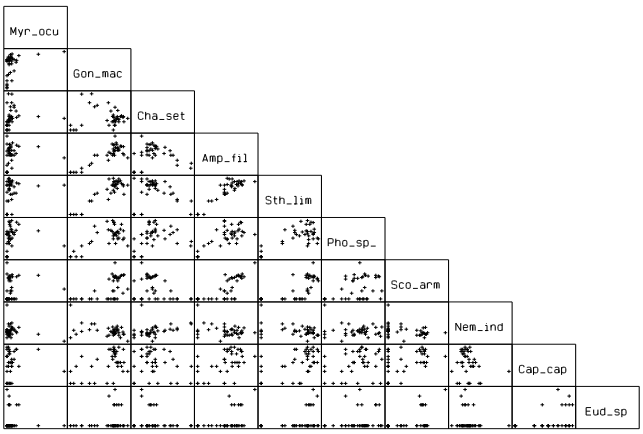

FIGURE 4.2: SCATTERPLOT MATRIX OF 10 SPECIES

Figure 4.2 shows a positive association between the abundances of the species *Amphiura filiformis*, *Sthenelais limicola*, *Goniada maculata* and *Phoronis sp.*. Other species seem to show no association with others at all, like *Capitella capitata* and *Nemertini indet.*. Two very high outliers for *Myriochele oculata* make it difficult to see its relationship with other species. These exceptional outliers correspond to station 24 and 15, both stations relatively close to the platform. Species *Chaetozone setosa* seems to be negatively associated with the group *Amphiura filiformis*, *Sthenelais limicola* and *Goniada maculata*.
The scatterplot matrix of the thirteen log-transformed chemical variables, displayed in figure 4.3 shows clearer patterns. It is evident that many variables are closely associated, in particular the group of heavy metals Cd, Cu, Fe, Pb and Zn. This is also confirmed by an inspection of the correlation matrix of the variables in table 4.2, showing high correlations between all heavy metals, THC, TOC and PEL. An outlier masks much of the relationships of the variables C17, C18 and PRI with the rest, as there is one station, reference station 40, that is very low on these variables. Omission of this outlier reveals a negative associa-

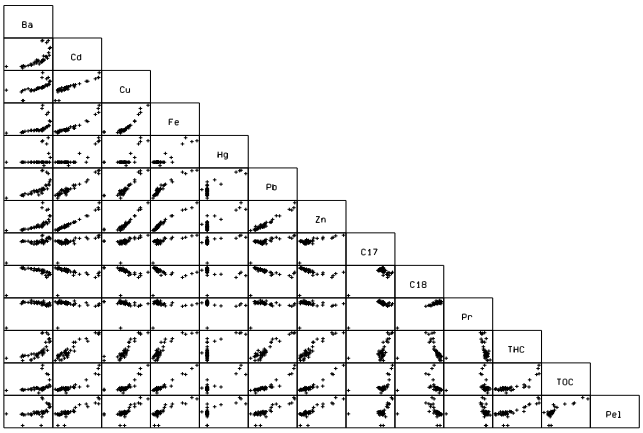

FIGURE 4.3: SCATTERPLOT MATRIX OF ENVIRONMENTAL VARIABLES

tion between C18, Pristane and the heavy metal group. Variable Hg has many coincident measurements.

Figure 4.4 shows scatterplots of the ten selected species against ten selected environmental variables. The ten species in the vertical dimension are, reading down, *Myriochele oculata*, *Goniada maculata*, *Chaetozone setosa*, *Amphiura filiformis*, *Sthenelais limicola*, *Phoronis sp.*, *Scoloplos armiger*, *Nemertini indet.*, *Capitella capitata* and *Eudorella sp.*. The horizontal dimension shows, from left to right, the variables lBa, lCd, lCu, lFe, lPEL, lPb, lZn, lPri, lTHC and lTOC. Some outliers have been removed, as they mask possible relationships (reference station 40, and stations 24 and 15 for *Myriochele oculata*). *Goniada maculata* seems to decrease with increasing concentrations of heavy metals. *Chaetozone setosa* seems to increase with most environmental variables, though at high concentrations its abundance drops.

Exploratory band regression (Hamilton, 1998, p. 187) is used to get an impression of the nature of the responses of the species with respect to the environmental variables, and of the assumed prevalence of the unimodal response pattern. In exploratory band regression, the horizontal axis is divided into a series of vertical bands of equal width. For each band the median is calculated, and the medians so obtained are connected by straight lines or cubic splines. The interest is focused on the character of the species response with respect to the variables in their original scale of measurement. Due to the positive skew of most environmental variables, exploratory band regression with say, eight bands, has the disadvantage that most bands will contain no data. This is improved if the log-transformed data are used, and positive skew is reduced. It

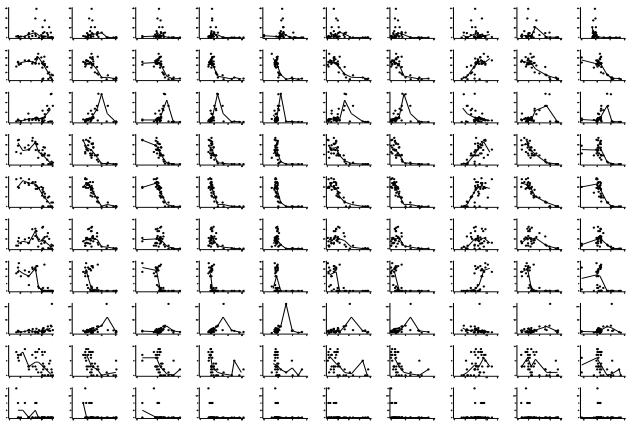

FIGURE 4.4: SCATTERPLOTS OF ABUNDANCE VERSUS ENVIRONMENTAL VARIABLES. Species in the vertical dimension are, reading down, *Myriochele oculata*, *Goniada maculata*, *Chaetozone setosa*, *Amphiura filiformis*, *Sthenelais limicola*, *Phoronis sp.*, *Scoloplos armiger*, *Nemertini indet.*, *Capitella capitata* and *Eudorella sp.*. The horizontal dimension shows, from left to right, the variables lBa, lCd, lCu, lFe, lPEL, lPb, lZn, lPri, lTHC and lTOC.

should be kept in mind that this way, we screen relationships in the logarithmic scale. Figure 4.4 shows exploratory band regressions for the ten selected species with respect to ten environmental variables. The horizontal axes are divided into eight equal-sized bands. If samples are spread uniformly over the horizontal axis, each band should contain about five observations. *Chaetozone setosa* is about the only species that shows a consistent unimodal pattern with respect to nearly all variables. *Nemertini indet.* also shows single-peakedness for several variables. The overall pattern however, is that abundance decays with increasing values for the environmental variables. An exception is the variable PRI, nearly all species seem to increase with higher concentrations of this variable.

In general, the data at hand do not seem to correspond with a unimodal model for species response. A unimodalist might respond that this is due to the fact that only a limited range of the environmental variable has been sampled. If a wider range would have been sampled, the decaying patterns observed in figure 4.4 could turn out to be part of a unimodal response curve. However, these are survey data, and it is not possible to control the range of the environmental variable as in a laboratory experiment.

As a starting point, we choose three species of varying total abundance and try to model them in terms of the environmental variables for 1990. The three

species selected form relatively extreme cases. The first one, *Goniada maculata*, is highly abundant. Considering all individual replicates, this species is only 10% sparse. Another selected species, *Gari sp.*, is rare and 90% sparse. Next, *Chaetozone setosa* is considered because it seems to behave more like a unimodal species. It is hoped that these species form reference points, and that the modelling problems encountered here are representative of what one could encounter, to a lesser extent, in regressions with the other species.

For the three species under consideration, we first regress the total sums of their abundance onto the chemical variables. For the chemical variables we have three replicates, but none of these replicates has a specific link to any of the biological replicates. We use the mean of the 3 chemical replicates as an estimate for the (missing) chemical observation of a biological sample. Next, we go down to the replicate level, and again regress species abundance onto the means of the chemical variables. A categorical variable with about 40 categories is created to indicate to which station each observation belongs. Finally we also take the nonnegative discrete nature of the response variable into account and consider Poisson regressions of the three species.

4.3 *Goniada maculata*

Goniada maculata is one of the most abundant species in the survey, with a mean abundance of 7.6 per replicate and a standard deviation of 4.5. *Goniada maculata* was absent in about only 10% percent of all samples, and has a reliability of 0.60 when data are square root transformed.

4.3.1 Regression of Summed Abundances

First, the sums of the abundances over the five replicates are regressed onto the means of each of the log-transformed environmental variables separately. The results of these regressions are briefly summarized in table 4.1. Table 4.1 shows the amount of explained variance in abundance (R^2) and the regression coefficient (b) with its standard error (se) for each variable. As a convention, a leading "l" in a variable name will be used to stress that we deal with the log-transformed values.

It is clear that nearly all variables explain a substantial part of the variation in abundance of *Goniada maculata*. Only lPRI is not significant, and lC18 is at the borderline. However, for lC17, lC18 and lPRI, there is a very influential outlier, reference station 40, that is very low on these variables. When this outlier is deleted, both lC18 and lPRI are highly significant. Estimates for lC18: $b = 3.21(0.46)$ with $R^2 = 0.5745$, and lPRI: $b = 4.53(0.83)$ with $R^2 = 0.4507$. The estimates for lC17 change drastically, lC17: $b = -4.89(0.45)$ with $R^2 = 0.7648$. These regressions suggest that in principle all variables should enter as candidates in a multiple regression model. The outlying reference station (40) is omitted from the analysis, as it keeps bothering the multiple regressions as well. This station is very different from all the others (see section 6.3.1 on principal components). Table 4.1 shows that all variables except lC18 and lPRI have negative regression coefficients. Increasing concentrations

	R^2	b (se)
lBa	0.229	-0.84 (0.25)
lCd	0.717	-1.36 (0.14)
lCu	0.545	-0.90 (0.14)
lFe	0.692	-2.81 (0.31)
lHg	0.542	-2.51 (0.38)
lPb	0.731	-1.71 (0.17)
lZn	0.707	-1.37 (0.14)
lC17	0.270	-1.52 (0.41)
lC18	0.101	0.68 (0.34)
lPRI	0.021	0.32 (0.36)
lTHC	0.802	-0.95 (0.08)
lTOC	0.671	-3.78 (0.43)
lPEL	0.641	-1.48 (0.18)

TABLE 4.1: REGRESSION COEFFICIENTS FOR *Goniada maculata*

of heavy metals, THC, TOC and PEL would thus reduce the abundance of *Goniada maculata*.

When we construct an initial multiple regression model, regressing abundance on all environmental variables, multicollinearity is present. Many of the explanatory variables are highly correlated, as indicated by their correlation matrix in table 4.2. When we regress each predictor on the others, very high coefficients of determination (R^2) are found, showing that the predictors share large part of their variation. Multicollinearity is further evident from the correlation matrix between the regression coefficients, where some high correlations are present (e.g. a correlation of -0.95 between lFe and the intercept). Under presence of multicollinearity the standard errors are inflated and so the estimated regression coefficients are imprecise (Hamilton, 1992, pp. 133-136).

	Ba	Cd	Cu	Fe	Hg	Pb	Zn	C17	C18	Pri	THC	TOC	PEL
Ba	1.00												
Cd	0.39	1.00											
Cu	0.17	0.77	1.00										
Fe	0.31	0.95	0.89	1.00									
Hg	0.31	0.97	0.70	0.89	1.00								
Pb	0.35	0.97	0.79	0.97	0.91	1.00							
Zn	0.20	0.86	0.98	0.96	0.79	0.89	1.00						
C17	0.55	0.83	0.70	0.85	0.75	0.83	0.77	1.00					
C18	-0.70	-0.49	-0.27	-0.41	-0.43	-0.49	-0.33	-0.36	1.00				
Pri	-0.41	-0.36	-0.18	-0.28	-0.34	-0.36	-0.22	-0.31	0.71	1.00			
THC	0.36	0.96	0.74	0.93	0.91	0.98	0.84	0.81	-0.50	-0.38	1.00		
TOC	0.27	0.94	0.89	0.99	0.86	0.97	0.95	0.83	-0.41	-0.28	0.93	1.00	
PEL	0.30	0.88	0.74	0.93	0.79	0.92	0.84	0.84	-0.43	-0.31	0.88	0.94	1.00

TABLE 4.2: CORRELATIONS BETWEEN ENVIRONMENTAL VARIABLES, 1990

Insignificant terms, often offending because of multicollinearity, are dropped from the full model one at a time. Proceeding in this manner, we arrive at a final regression model:

```
sqrt(N) =  5.72 - 2.36 lC17 - 0.59 lTHC
          (0.84) (0.63)      (0.12)
```

Log transformed THC and C17 together account for 85% of the variance in

the square root transformed total abundance of *Goniada maculata*. The intercept gives the square root of the expected abundance when lTHC and lC17 are both zero in the log scale. This corresponds to an expected abundance of $(5.72)^2 \approx 33$, when there is 1 milligram of Hydrocarbon per kilo and the n-C17/pristane ratio is 1. It remains nevertheless difficult to assess which of the variables do really affect the abundance of *Goniada maculata*. Many insignificant predictors have been dropped from the regression equation, but these share variation with the ones remaining in the equation, and so the importance of the latter might easily be overstated. As an alternative a data reduction technique is employed, as many variables are so highly correlated. Thus the abundance of *Goniada maculata* is regressed onto principal components (regression onto principal components often shortly indicated as PCAR). The principal component analysis of the chemical data is discussed in more detail later in chapter 6. The first component accounts for 85% of the variance of the chemical data. Only the first principal component turns out to be significant, and the corresponding regression model is:

```
sqrt(N) = 5.78 - 1.88 PC1,
          (0.16) (0.16)
```

where 79% of the variance in the transformed abundance of *Goniada maculata* is explained by the first principal component, and the first principal component is highly significant. Figure 4.5 shows the fitted regression line.

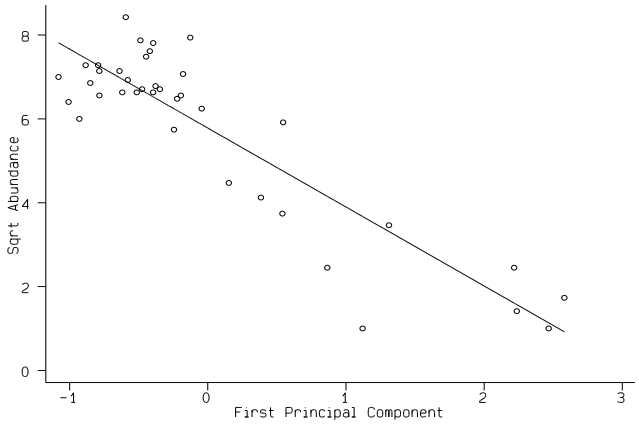

FIGURE 4.5: REGRESSION ON FIRST PRINCIPAL COMPONENT

In order to get a better idea of how the different variables affect the abundance of *Goniada maculata*, some conditional effects plots are constructed. Since both

the response variable and the predictors have been transformed, the relationship between abundance and environmental variables is not linear any more, but curvilinear. A conditional effects plot graphs the fitted values of the response variable against one of the predictors, where all other predictors are held constant (Hamilton, 1992, pp. 158). With a conditional effects plot, it becomes clear what the curvilinear regression on THC and C17 above means in terms of the original variables.

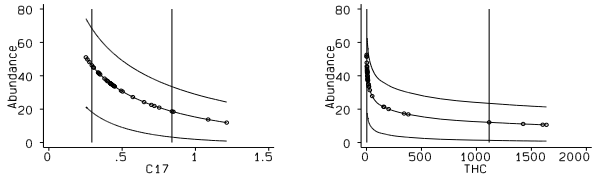

FIGURE 4.6: CONDITIONAL EFFECTS PLOT

The conditional effects plots of C17 and THC are shown in figure 4.6. The middle curve shows the relationship between abundance and the x-variable, when the other covariate is held fixed at its mean. The top curve show the same relationship when the other covariate is kept constant at its minimum, and the bottom curve depicts likewise the relationship when the other covariate is at its maximum. The vertical lines in the plots correspond to the 10^{th} and 90^{th} percentile of the x-variable. As can be seen from the plots, THC exerts its largest effect for values in the range of its 10^{th} percentile, and the drop in abundance tails off pretty quickly as THC increases. On the other hand C17 seems to affect the abundance of *Goniada maculata* at a slowly decaying rate, ranging from its 10^{th} percentile to its 90^{th} percentile. There is some sign of interaction. We see that if THC is at its maximum, the effect of C17 seems less severe, as the corresponding curve tails off slower compared to the curve with THC at its mean.

We notice here that the square root and log transformation were chosen to reduce skew. If we consider one environmental variable (x) only, then the fitted abundance (N) is expressed in terms of x as:

$$N = (a + b \ln x)^2 \tag{4.1}$$

This function achieves a minimum if $x = e^{-ab/b^2}$, and increases without limit for large x. There is an inflection point at $x = e^{(b-a)/b}$. Function (4.1) is convex over $(0, e^{(b-a)/b})$ and concave over $(e^{(b-a)/b}, \infty)$. It runs thus contrary to the idea that the response of a species to an environmental variable is unimodal, that is, single-peaked and with a maximum rather than a minimum. Applying standard statistical transformations to reduce skew thus leads to an empirical model that does not correspond to the theoretical unimodal model. A response that is unimodal, at least over part of the range of the environmental variable, can be obtained if quadratic terms of the log-transformed environmental variable are included in (4.1).

Unimodal response curves can also be obtained by fitting parabolas to log-transformed abundance data (Jongman et al., 1987, p. 41), thus choosing a different transformation for the response variable. In the case of *Goniada maculata*, the log-transformation is not adequate, as it would introduce considerable negative skew, whereas the square root transformation is satisfying. When quadratic terms are introduced in the multiple regression model they turn out to be insignificant and offending as they cause multicollinearity. Another problem of the logarithmic transformation is that the log of zero is not defined, whereas for the square root transformation zeros are not problematic. We thus keep the multiple regression model proposed, even though it does not correspond to a unimodal response.

We also note that the variables in the study are mainly pollutants, likely to have toxic effects on the organisms. Their effect might be very different from a variable like say, temperature. For temperature some interval can probably be discerned for which its increase improves the physiological conditions of the organism under study, and an increase in abundance over that particular interval is expected. For pollutants like heavy metals the situation could be different, e.g. if no such a profitable range exists, a pattern of decay is maybe more adequate than a unimodal response.

4.3.2 Taking Replicates into Account

In the previous section, abundances were summed over the five replicates. Doing so, a possible component of variation within the stations is ignored. In this section the five replicates are kept separate. What are the corresponding chemical measurements of each replicate? In fact, these are not available. There are only three chemical measurements of other samples at the same location. The mean of these three measurements is used as the estimate of the chemical variables of the biological sample. At each mean of a chemical variable, there are thus five biological observations. This situation is illustrated in figure 4.7, where a scatterplot of *Goniada maculata* is shown against Barium.
One can test for significant differences between the stations, by using a categorical variable indicating to which station an observation belongs. A oneway analysis of variance (ANOVA) shows that there are significant differences be-

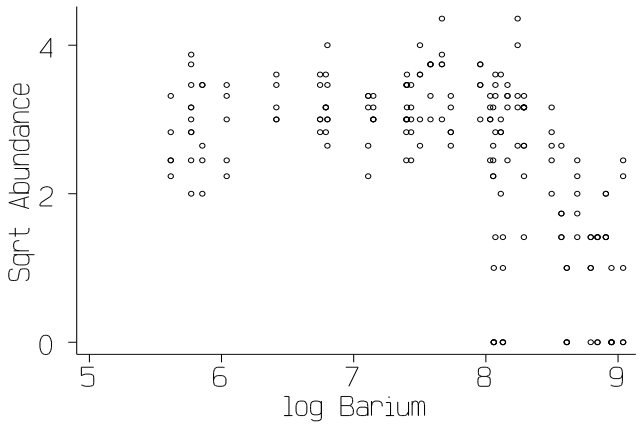

FIGURE 4.7: SCATTERPLOT OF *Goniada maculata* VERSUS BARIUM

tween the stations ($F = 11.70; p < 0.00005$). However, the Bartlett's χ^2-test for equality of the variances is significant, and this casts doubt on the assumption of equality of variances that underlies ANOVA. The square root transformation improves this situation, since ANOVA with the transformed abundances ($F = 18.41; p < 0.00005$) gives a nonsignificant Bartlett χ^2. Anyway, a Kruskal-Wallis test is used as a non-parametric alternative, where the null-hypothesis now states that the median abundance is equal for all stations. The Kruskal-Wallis test gives a χ^2_{38} of 130.5, ($p = 0.0001$), and supports the hypothesis that there are differences between the stations. Which are the stations that differ? By inspecting the boxplots for each station separately it becomes clear that *Goniada maculata* is nearly absent at stations 3 and 37, and considerably lower on 25, 30, 31, 32, 35 and 36. This is a set of stations that is close to the platform (see chapter 2, p. 7).

The regressions of square root transformed abundance are again performed onto each of the variables separately. Doing so for the "unaggregated" data gives results that are qualitatively comparable to table 4.1: all variables have significant negative coefficients, except lC17 and lPRI that have positive regression coefficients. The amounts of variance explained are in general lower than in table 4.1. Dropping insignificant variables from the full model one by one, we again end up with the regression model:

```
sqrt(N) =  2.50  - 1.13 lC17 - 0.29 lTHC,
          (0.28)  (0.21)      (0.04)
```

and the coefficient of determination (R^2) is 71%. Figure 4.8 shows the residuals of this regression, where residuals are labelled with their station number. At first

sight these residuals look okay, there are no signs of curvature or large outliers. However, the sample size is 190, and figure 4.8 seems to contain a much smaller amount of points. This is suspicious as there must be many coinciding residuals. When we take a close look at the residuals within one particular stations, we see that the residuals tend to be similar in order of magnitude. But when we compare residuals across different stations, we see larger differences. Phrased in other words, the residuals show some intraclass correlation. In a standard residual plot like figure 4.8 without station numbers, this intraclass correlation can easily go unnoticed. The intraclass correlation of the residuals is found to be 0.23, and differs significantly from zero. The residuals can thus not be regarded independent, and one of the basic assumptions of the regression model is violated.

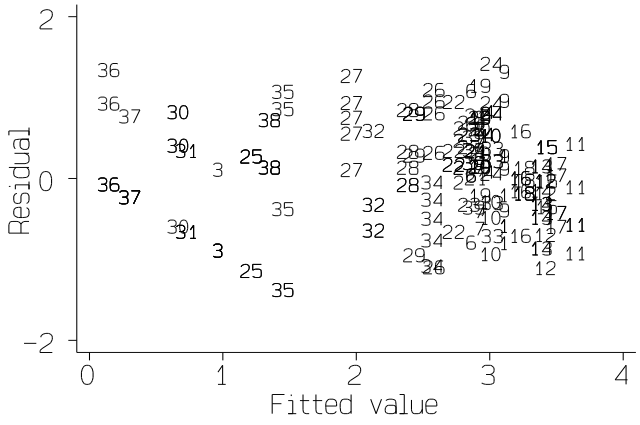

FIGURE 4.8: RESIDUAL PLOT

As a consequence, standard errors might be biased, and T-tests might be invalidated. In order to cope with this problem of intraclass correlation, a random coefficient model is used, where the intercept of the regression is allowed to vary among stations. The model then becomes:

$$N_{ij} = \alpha_j + \beta x_{ij} + \varepsilon_{ij}\,, \tag{4.2}$$

where the first term on the RHS can be written as: $\alpha_j = \alpha + u_j$, α being the mean intercept, and u_j the deviation from the mean intercept of station j. The intercept is supposed to follow a $N(\alpha, \sigma_u^2)$ distribution. Residuals u_j and ε_{ij} are assumed uncorrelated. Model (4.2) is a multilevel model, where the hierarchical structure of the data is recognized: the first level is formed by the actual measurements, the second level is formed by the stations. More details on this type of models, embedded in an educational context, can be found in Goldstein (1987). The important feature of this model in this context is that it allows

for intraclass correlation in the response variable. In particular, the variance in abundance given by model (4.2) is $\sigma_u^2 + \sigma_e^2$, and the covariance between two measurements at the same station is: $cov(u_j + e_{ij}, u_j + e_{ij}) = cov(u_j, u_j) = \sigma_u^2$, so that the intraclass correlation of the response variable is: $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$.

Estimating this model we obtain:

```
sqrt(Abun) = 2.50 - 1.13 lC17 - 0.29 lTHC
             (0.40) (0.30)      (0.06)
```

Notice that the estimated coefficients are virtually the same as in the ordinary regression model previously commented, but that the standard errors obtained are larger, and that the confidence intervals for the coefficients are thus wider. When we calculate again the intraclass correlation, but now for the residuals of the random coefficient model, we find a value of -0.15, not significantly different from 0. The residuals (level 1) are now free of intraclass correlation, and the assumption of independence is no longer being violated. Phrased in another way: an ANOVA of the residuals of the ordinary regression model gives significant differences between the means of stations ($F = 2.51; p = 0.000$). An ANOVA of the residuals of the random coefficients model however, is not significant any more ($F = 0.36; p = 0.999$). The random coefficient model is thus the preferred model. The variance components σ_u^2 and σ_ε^2 are estimated 0.09 (s.e. 0.03) and 0.28 (s.e. 0.03) respectively. The variance within stations is thus about three times larger as the variance between stations. The intraclass correlation in square root abundance, given the model, is then 0.24, that is to say that 24% of the total variance in abundance of *Goniada maculata* is due to variation between stations. A graphical illustration of the random intercept model is given in figure 4.9, using only one predictor, lTHC.

4.3.3 Poisson Regression

Poisson regression is a particular case of a generalized linear model (McCullagh and Nelder, 1989). Three components are distinguished in generalized linear models: a random component, a systematic component and a link function. The random component consists of a response vector $\mathbf{y}$ of n components, containing outcomes of a random variable $\mathbf{Y}$, identically distributed with vector of means $\boldsymbol{\mu}$:

$$E(\mathbf{Y}) = \boldsymbol{\mu}. \tag{4.3}$$

The systematic component consists of a linear combination of the covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ and is called the linear predictor $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \sum_{i=1}^{p} \mathbf{x}_i \beta_i = \mathbf{X}\boldsymbol{\beta}. \tag{4.4}$$

A link function, $g(\cdot)$, links the random and systematic component:

$$\eta_i = g(\mu_i). \tag{4.5}$$

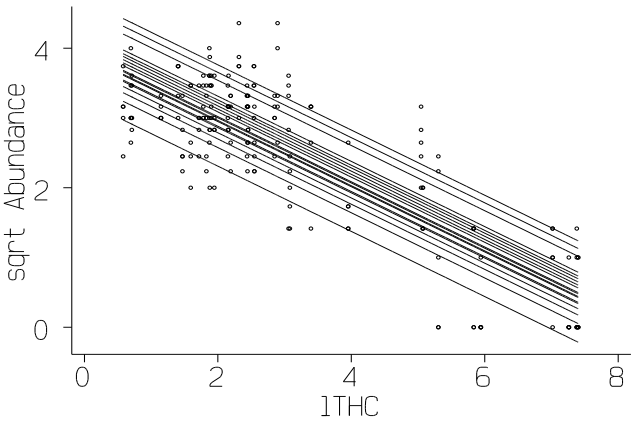

FIGURE 4.9: RANDOM INTERCEPT MODEL FOR *Goniada maculata*

Ordinary regression corresponds to a generalized linear model where the response variable is normally distributed, and where the link function is the identity function. In the case considered here, Poisson regression, the response variable is a count variable, assumed to follow a Poisson distribution, and the link function is chosen to be the natural logarithm, so that one has:

$$\eta_i = \ln \mu_i. \tag{4.6}$$

The model fitted to the data so becomes:

$$\mu_i = e^{\beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}}. \tag{4.7}$$

The parameters of a generalized linear model can be estimated by an iterative weighted least squares algorithm (McCullagh and Nelder, 1989; Dobson, 1991). In this context, μ_i is the expected abundance. When we consider only one predictor, holding the other ones constant, we have:

$$\mu_i = e^{\beta + \beta_1 x_{i1}} = c e^{\beta_1 x_{i1}}, \tag{4.8}$$

with $c = e^{\beta}$, and $\beta = \beta_o + \beta_2 x_{i2} + \cdots \beta_p x_{ip}$. Depending on the sign of coefficient β_1, we thus fit a model of exponential growth or exponential decay, not a unimodal response model. Model (4.8) has the disadvantage that, for $\beta_1 > 0$, abundance increases without limit as x increases, which is not very realistic.

When a Poisson regression is carried out for the total abundances of *Goniada maculata* onto the first principal component, we find the regression model:

```
N = 3.45 - 0.72 pc1,
   (0.03) (0.05)
```

The regression coefficients of this model are significant, but the goodness of fit statistic χ^2 is large, 153.4 ($p = 0.000$). Given the model, one has to reject the hypothesis that these data are Poisson distributed. This also happens when we build regression models using the original (log-transformed) variables rather then principal components, and also when we use all individual replicates rather than their sums. For an abundant species like *Goniada maculata*, Poisson regression thus seems not to make much sense. In fact, the χ^2-statistic suggests that data might be overdispersed. This is effectively the case, as the sample variance is much larger than the mean. We would have to try to take this overdispersion into account, or might consider negative binomial regression as an alternative (Hilbe, 1994). These approaches are not further considered here.

4.4 *Gari sp.*

In this section we consider the same type of models as considered for *Goniada maculata*. Species *Gari sp.* differs from *Goniada maculata* in the sense that it has a very low reliability of 0.08, not significantly different from zero, and that it is a rare species. There is overdispersion, as the quotient of sample mean and variance is 0.6. However, the distribution of the total abundance of *Gari sp.* can hardly be considered different from a Poisson distribution, because the quotient has a 95% bootstrap confidence interval of (0.43,0.99) (see also chapter 2).

4.4.1 Regression of Summed Abundances

Regressing the square root transformed total abundance of *Gari sp.* onto the log-transformed environmental variables, one by one, shows that only the regression on lBa is significant:

```
sqrt(N) = -1.54 + 0.26 lBa
          (0.76) (0.10)
```

lBa explains 16% of the variance in square root abundance. The fitted regression line is shown in the upper left panel of figure 4.10. The upper right panel shows the relationship between abundance and Barium when the variables are transformed back to their original scale.
In the left panel, the size of the points is proportional to their leverage. There are 4 cases at the extremes of the regression line that exert leverage above the critical level of $2K/n$, where K is the number of estimated parameters and n is the sample size (Hamilton, 1998). Another case statistic called dfbeta measures the influence of each case on the regression coefficients. The dfbetas of all cases however, are below the critical level, and the estimated regression coefficients do not change much if these cases are deleted. Upon deletion, other cases in turn appear with high leverage. The station with the highest abundance is an outlier with an exceptionally high residual, but its deletion neither changes the regression estimates very much. Thus, we retain the regression equation above as describing the relationship between the abundance of *Gari sp.* and lBa.

A word of care is in place here, however. If we adopt a significance level of 0.05, we have a chance of 1/20 of finding a significant regression coefficient by chance alone. When we screen 20 variables, we can be almost sure that one of them is

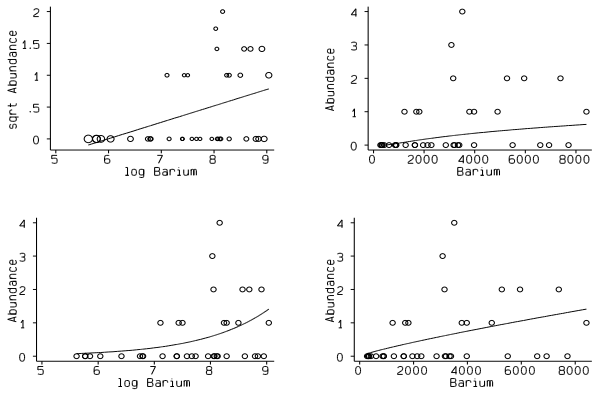

FIGURE 4.10: REGRESSION OF *Gari sp.* ON BARIUM. Upper panels show linear regression, lower panels show Poisson regression, both in original and log scale.

significant. With 13 variables in the survey, it thus might come as no surprise that a species seems to respond at least to one of them.

When the regression of the abundance of *Gari sp.* with respect to principal components is considered, *Gari sp.* turns out to decrease significantly with the second principal component, whereas the first principal component is insignificant.

$$\underset{}{\texttt{sqrt(N) =}}\ \underset{(0.10)}{\texttt{0.01 pc1}}\ \texttt{-}\ \underset{(0.10)}{\texttt{0.28 pc2}},$$

About 14% of the variance is explained by the first two principal components. This regression is consistent with results reported for lBa, as the first principal component correlates positively with lBa, and the second negatively.

Since *Gari sp.* is so sparse, Poisson regression is also considered for its total abundance (see section on Poisson regression below).

4.4.2 Taking Replicates into Account

When we consider all replicates of the abundance, forming five repeated measurements at the means of the chemical variables, only about 10% of the observations is non-zero, and they are all whether 1 or 2. Only a 2% of the counts consist actually of the value 2. With so few counts different from outcomes (0,1) we recode the data as absence-presence data and perform logistic regression, rather than considering ordinary regression or Poisson regression. In

logistic regression, the logit (log odds) is modelled as a linear function of the k predictors:

$$\hat{L} = \ln\left(p/(1-p)\right) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i, \tag{4.9}$$

where p is the probability of occurrence. Logistic regression is also a particular case of a generalized linear model, where the distribution of the response variable is binomial, and the link function is the logit function. Only the logistic regression onto lBa is significant:

```
L = -9.78 +  0.94 lBa
    (3.02)  (0.37)
```

If the concentration of Barium is zero in the log scale (corresponding to 1 milligram of Barium per kilo in the original scale), the estimated logit is -9.78. This means that the odds of finding *Gari sp.* are $e^{-9.78}$ or in other words that the probability of finding one or more specimens of *Gari sp.* at this concentration of Barium is $1/(1+e^{-\hat{L}}) = 1/(1+e^{9.78}) = 5.7 \times 10^{-5}$. An increase of one unit in the log scale of Barium (that is a multiplication of the concentration of Barium by a factor $e \approx 2.718$ in the original scale) multiplies the odds for finding *Gari sp.* by a factor $e^{\beta_1} = e^{0.94} \approx 2.56$. Logistic regression fits a S-shaped curve to the probabilities of finding *Gari sp.* as a function of lBa. Data points and fitted curve are shown in figure 4.11.

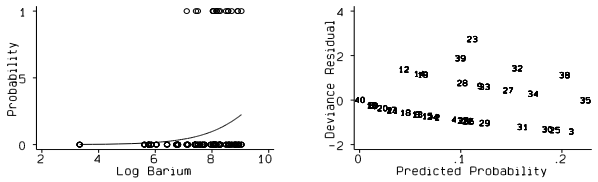

FIGURE 4.11: LOGISTIC REGRESSION OF *Gari sp.* ON BARIUM

It should be kept in mind however, that this result is based on a highly skewed response variable: only about 10% of the cases correspond to $y = 1$. We also

note that because the log transformation was used to reduce skew in the environmental variables, the odds of finding *Gari sp.* are in fact modelled as a power function of the concentration of Barium in its original scale, β_1 being the power:

$$p/(1-p) = e^{\beta_0+\beta_1 \ln x} = e^{\beta_0} x^{\beta_1}. \tag{4.10}$$

The deviance residuals of the logistic regression are shown in the right panel of graph 4.11. Most residuals line up along horizontal curves. The lowest curve corresponds to stations where *Gari sp.* has not been found. The second lowest curve correspond to stations where *Gari sp.* has been found once, the third lowest curve where it has appeared twice. The one station (23) with an exceptional high residual is the only one where is has been detected three times. Deletion of this station does hardly alter the coefficient of lBa ($b_0 : -10.09, b_1 : 0.96$).

Curious things happen when we analyze the deviance residuals or Pearson residuals of the logistic regression. For any one station, the five chemical measurements for lBa are identical (because the biological replicates were assigned the means of the chemical replicates). Whereas the sample size is 195, the number of predictor patterns is thus only 39. Software for logistic regression typically calculates the residuals for the number of predictor patterns (X-patterns), and each observation within such a pattern is assigned the same residual. As a consequence, the residuals of the logistic regression on lBa have an intraclass correlation of 1. This suggest that residuals can not be regarded as independent, even though the response variable has no intraclass correlation at all ! This is an undesirable characteristic of the analysis. If the binary response variable shows variation within groups, it would be natural that the residuals within groups also display variation. The detected intraclass correlation in the residuals is here taken to be an artifact produced by the way residuals are calculated. As the response variable has an intraclass correlation of only 0.08, we do not worry about dependence of observations.

4.4.3 Poisson Regression

Because the distribution of the summed abundances of *Gari sp.* hardly differs from a Poisson, Poisson regression can be considered for the sums:

```
N = -7.30 + 0.85 lBa
    (2.59) (0.31)
```

In this case, the χ^2-statistic for goodness of fit is 43.27, and the null hypothesis that the data, given the model, are Poisson can not be rejected ($p = 0.189$). For a rare (and thus sparse) species like *Gari sp.*, Poisson regression seems to be useful. The fitted regression line is shown in the two bottom panels of figure 4.10, both in the log transformed scale as well as in the original scale of measurement. We notice that when log transformed environmental variables are used, the actual response model that is fitted when Poisson regression is used is also a power of the environmental variable. E.g. in Poisson regression with a log-transformed predictor we have:

$$\hat{y} = e^{\beta_0 + \beta_1 \ln x} = e^{\beta_0} x^{\beta_1}. \tag{4.11}$$

This function is concave if $0 \leq \beta_1 \leq 1$, and convex if $\beta_1 \geq 1$ (we assume $x > 0$, otherwise the log-transformation would not be possible), and without optimum. Such a response function is realistic in the sense that negative outcomes are not possible, but it has the disadvantage that abundance can increase without limit.

The deviance residuals of the Poisson regression are plotted against the linear predictor in figure 4.12. The residuals line up in curves which correspond to total abundances of 0,1,2,3 or 4. There is clearly heteroscedasticity, as the variance of the residuals increases for higher values of the linear predictor.

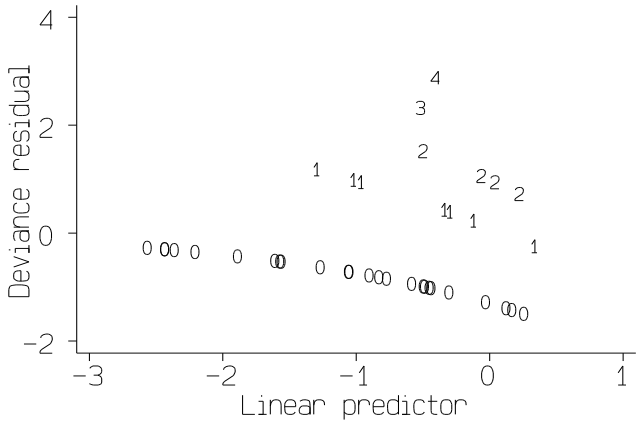

FIGURE 4.12: DEVIANCE RESIDUALS VERSUS LINEAR PREDICTOR FOR *Gari sp.*

Merely for the sake of comparison with logistic regression, we also consider the Poisson regression for the unaggregated data. The response variable is assumed discrete, and we no longer use the square root transformation. Again, only the regression with predictor lBa is significant:

```
N =  -8.96 +  0.85 lBa
     (2.56)  (0.31)
```

The χ^2-statistic for goodness of fit is now 96.4, and the null hypothesis that the data, given the model, are Poisson can certainly not be rejected ($p = 1.000$). It is clear that the Poisson distribution becomes more apt when all replicates are considered rather then their sums. The regression equations for sums and replicates are very similar, the coefficients for lBa are the same, and the intercepts do not differ significantly from each other. The deviance residuals have an intraclass correlation of 0.06, which is not significantly different from 0, and

thus there is no need to specify a particular model to account for intraclass correlation.

4.5 *Chaetozone setosa*

Some regressions for *Chaetozone setosa* are considered, as this is one of the few species that seems to show a unimodal response. *Chaetozone setosa* is one of the most abundant species, and has a reliability of 0.6, and its distribution is not Poisson (cf. table 2.1).

4.5.1 Regression of Summed Abundances

When the total abundance of *Chaetozone setosa* is regressed onto each predictor separately, only lBa and lTOC turn out to be significant. The exploratory band regressions of section 4.2 however, showed evidence for curvature in the relationship of the abundance of *Chaetozone setosa* and several environmental variables. When quadratic terms are included more variables pop up as significant, notably the heavy metals lCd, lFe, lPb, lZn and lTHC. The results of the significant regressions are summarized in table 4.3 (intercepts not shown).

	R^2	b_1	b_2
lBa	0.16	0.78 (0.29)	-
cCd	0.43	1.18 (0.30)	-0.78 (0.14)
cFe	0.32	2.78 (0.97)	-2.78 (0.69)
cPb	0.35	1.30 (0.42)	-1.06 (0.23)
cZn	0.36	1.14 (0.35)	-0.65 (0.14)
cTHC	0.22	0.58 (0.21)	-0.28 (0.08)
lTOC	0.12	-1.43 (0.65)	-

TABLE 4.3: REGRESSION COEFFICIENTS FOR *Chaetozone setosa*. Intercept not shown, b_1 for the linear term, b_2 for the quadratic term.

The introduction of quadratic terms leads to multicollinearity, as the quadratic terms tend to be correlated with their non-quadratic counterparts. This is partly circumvented by centring the log transformed environmental variables on their means. This reduces multicollinearity considerably, and produces more precise standard errors. To stress this centring, a leading 'c' in a variable name indicates that a variable is centred on the mean after the log transformation.

All regressions with quadratic terms (b_2) in table 4.3 have a negative coefficient for the quadratic term. Without backtransforming the variables we fit a parabola to the log-transformed data, which corresponds to a concave response function if the coefficient of the quadratic term is negative. Thus we fit a unimodal model with a maximum in the log-transformed scale. In the original scale, abundance is related to an environmental variable by:

$$N = \left(b_0 + b_1 \ln x + b_2 \ln^2 x\right)^2, \tag{4.12}$$

which is a polynomial in $\ln x$. Depending on the values of the coefficients, a maximum is possible.

But how unimodal is *Chaetozone setosa*? Four of the fitted quadratic regressions shown in table 4.3, in the log-transformed scale, are shown in figure 4.13. In all four graphs we see that most data points scatter around the left half of the parabola. Only about five stations are actually responsible for the curvature in the data. These five stations are the same stations in all four graphs, and correspond to stations 3,31,37,30 and 36. Apart from station 3, these are the stations that are the most close to the platform, and are very high on heavy metals and THC. If this group of stations is left out of consideration, *Chaetozone setosa* shows an overall pattern of linear increase with the environmental variables. More data in the right tail of the distribution of Cd, Fe, Pb and THC would be welcome in order to confirm the unimodal response pattern.

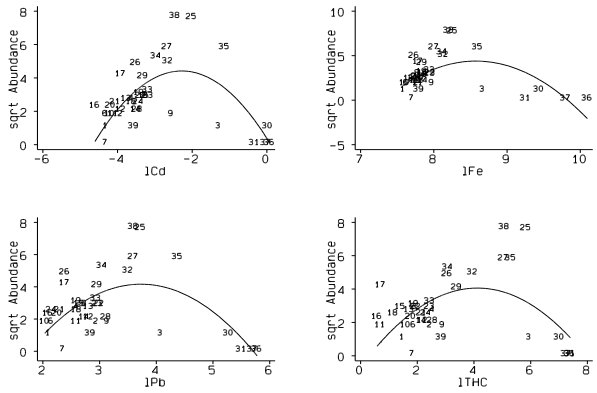

FIGURE 4.13: REGRESSIONS WITH QUADRATIC TERM FOR *Chaetozone setosa*

Another feature of figure 4.13 is that the parabolic response curves fitted in the log scale are strictly symmetric. This means that they rise as steep on the left as they fall on the right. This seems a bit restrictive, as the response pattern with respect to lTHC for instance suggests rather a gradual rise followed by a steep drop. In the original scale of measurement however, this symmetry does not exist. The relationship between abundance and the four variables previously considered in the original scales of measurement is shown in figure 4.14.
In the original scale of measurement, the quadratic regression models imply a unimodal response pattern. Note that the response curves do not follow the steep rise in abundance entirely, but incline much before. With the appearance of so many relevant quadratic terms, it becomes complicated to construct an

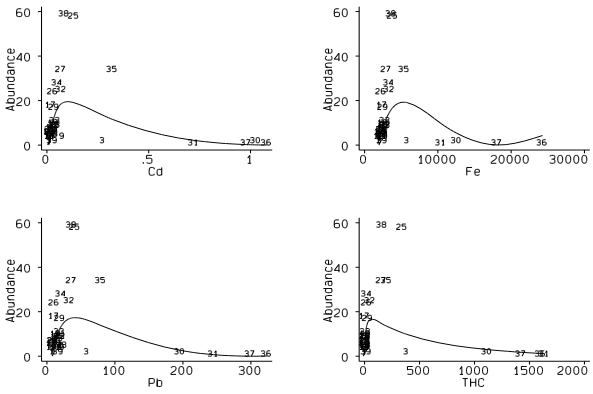

FIGURE 4.14: RESPONSE OF *Chaetozone setosa* IN ORIGINAL SCALE

overall final model for *Chaetozone setosa*. When the amount of variables is reduced by a principal component analysis, the abundance of *Chaetozone setosa* depends in a quadratic fashion on both components. We resume this with the model:

```
sqrt(N) = -0.56 pc1 - 1.63 pc2 + 0.57 pc2^2
          (0.26)     (0.24)     (0.18)
```

This model explains 53% of the variation in abundance of *Chaetozone setosa*. The squared first principal component is left out, as it correlates highly with both the first and the second principal component, causing multicollinearity and inflating the standard errors.

4.5.2 Taking Replicates into Account

For the desaggregated data, we use the variance components model previously described in (4.2), this because the relatively high reliability of *Chaetozone setosa* leads to residuals with intraclass correlation that are not independent. Many significant relationships are detected. These are summarized briefly in table 4.4.

Note that there are some differences in comparison with regressions using sums. Specifically, lPRI is now significant, where it was not, and lTOC is insignificant, where it was before. The amounts of variance explained are lower in comparison with regressions using sums. If we again adopt the strategy to include all terms in table 4.4, and drop insignificant terms one by one, we end up with the regression model:

	R^2	b_1	b_2
lBa	0.12	0.37 (0.13)	-
cCd	0.28	0.56 (0.14)	-0.34 (0.07)
cFe	0.21	1.34 (0.45)	-1.26 (0.32)
cPb	0.23	0.61 (0.20)	-0.46 (0.11)
cZn	0.23	0.53 (0.17)	-0.28 (0.07)
cTHC	0.16	0.28 (0.09)	-0.12 (0.04)
lPri	0.07	-0.88 (0.43)	

TABLE 4.4: REGRESSION COEFFICIENTS FOR INDIVIDUAL REPLICATES OF *Chaetozone setosa*

```
sqrt(N) = 2.76 + 0.33 cCd - 0.32 cCd^2 - 1.17 lPRI
          (0.44) (0.15)     (0.06)       (0.42)
```

Though, as stated in previous sections, such an equation probably overstates the importance of the predictors (notably Cd), because of shared variance with other predictors.

4.5.3 Poisson Regression

Chaetozone setosa is highly abundant, and its distribution is not Poisson. Poisson regression does not work for the total abundance and neither for the unaggregated data. One would have to take overdispersion into account.

4.6 Conclusions

This chapter illustrates that the construction of regression models for the species is elaborate, requiring checks for multicollinearity, careful checking of residuals for outliers, leverage, intraclass correlation and so on. If the interest is focused on the response of a particular species, all this is necessary. However, with a total of 152 species in 1990, regression modelling for each species separately is a prodigious amount of work. Ecological interest might be focused on a particular species, but often the response of the community of species as a whole is of interest. This asks for an approach that treats all the species data simultaneously in a single multivariate approach, and this is the subject of the next chapters.

4.6.1 To Sum or Not to Sum

In the models above, distinction has been made between regressions based on the summed abundance and on separate treatment of all replicates. When should we choose to perform the "aggregated" analysis or the "unaggregated" analysis? Imagine we have a theoretical species with a reliability of 1. The counts of the five replicates of such a species at a particular station are equal, as there is no variation within stations. Say our total sample size is about 200. The regression coefficients obtained when regressing abundance onto an environmental variable are now identical to the regression coefficients obtained would we regress the

means of the five replicates onto the means of the environmental variables. Evidently, the standard errors of the coefficients do change as we use aggregated data, because the sample size is now 40. Do we regress using summed abundance, then the regression coefficient for the environmental variable will still be identical to the one obtained in the unaggregated regression, whereas the intercept will be exactly 5 times as large. With regressing the sums, evidently the standard errors are also larger compared to the "unaggregated" regression. In the "unaggregated" regression however, intraclass correlation violates the independence assumption, and variance components models would be needed to account for this, which in turn will produce larger standard errors.

In short, if the purpose is to study one particular species, there is no need to sum or average the species counts. A larger sample size is preferable, and if intraclass correlation is present, we can use a variance components model like (4.2) to account for this.

In fact the idea to sum is a preliminary step to prepare the stage for a the multivariate approach when we study many species simultaneously. It simplifies the multivariate analysis, as we need to consider only one correspondence analysis or one principal component analysis of the biological data matrix. Not aggregating the abundances implies five multivariate analyses of the biological data, whose results then need to be integrated in some way. If the reliability of the species data is very high, then nearly all variation is between stations, and probably not much is lost by summing or averaging before doing multivariate analysis. Unfortunately however, most species have a low reliability, and we would ignore a important component of variation by summing. This topic is treated in some more detail in the next chapter.

4.6.2 Variation over Time

In this chapter we only considered data of one particular year, and that year was chosen for regression analysis because it had the highest number of samples. But because the sampling is repeated every year, there is also a time dimension. A complication is that the station network has undergone changes from year to year, and in more recent years less samples have been taken. A subset of about 12 stations can be identified that has been sampled every year. Several approaches are possible in order to consider the time dimension. A simple approach would be to concatenate the data sets from several years, and to use binary indicators that tell to which year a particular sample belongs. By testing if slope or intercept dummies made with these indicators differ from zero, one could test for significant differences between the years. Another approach would be to use a random coefficients model with three levels: the replicates being level one, station level two and years level three.

4.6.3 Poisson Regression

Poisson regression was found to useful for rare species, for which a Poisson distribution can often not be rejected as a probability model. If nearly all the counts consists of zero and ones, one might even consider to recode the data

entirely as absence-presence data and to perform logistic regression. For more abundant species, as already detected in section 2.2, species counts are often overdispersed. One would have to take overdispersion into account, or try other alternatives such as negative binomial regression.

4.6.4 Correlations between Species

It should also be stressed that in this chapter a few species have been analysed, but independently from others. In practice, the abundances of the different species can be related, as is also clear from the scatterplot matrix in figure 4.2. In particular, some species might live in symbiosis where they mutually profit from each other, and their abundances might be expected to correlate positively. Others might be predators or preys, and their abundances could correlate negatively. Yet others might be entirely indifferent with respect to each other. Detailed biological knowledge of the relationships between the species themselves and their population dynamics is needed for building models accounting for correlations between species. With two species this could already get pretty complex, not to speak of 152.

4.6.5 Unimodal Models

With *Chaetozone setosa* as an exception, not many species seem to show a unimodal response. This can be because there is no such response, or because a too limited range of the environmental variables was sampled. It would be interesting to measure abundance over a wide range of equally spaced intervals of the environmental variable. In a monitoring survey like this, that is not possible. In order to assess if a species responds in a unimodal manner to an environmental variable, one would need to control for these variables in an experimental setting.

4.6.6 Relationships Detected

We briefly summarize the relationships detected for the species considered in this chapter. The abundance of *Goniada maculata* is seen to decrease significantly with the first principal component. We cannot disentangle the effects of the variables really responsible (C17, THC, heavy metals) for this, as they all correlate very highly. *Gari sp.* increases significantly for higher concentrations of Ba. However, we screened so many variables that this still might be a chance effect. *Chaetozone setosa* shows a unimodal pattern with respect to many of the contaminants, and seems to be a species preferring contaminated conditions. This species seems to respond significantly to heavy metals and THC, though again we can not disentangle their separate effects.

Chapter 5

Theory of Correspondence Analysis

5.1 Introduction

In previous chapters, we have focused on the analysis of abundance data considering one species at a time. In this chapter we move to the multivariate analysis of the species data, where the abundances of all species are considered simultaneously. The techniques that come to mind in this context are principal component analysis (PCA) and correspondence analysis (CA). PCA is usually applied to a data matrix of continuous variables, whereas here we deal with nonnegative count data. Abundance data are however, often treated as quantitative, and PCA has been applied to abundance data (See e.g. Digby and Kempton (1987) for some examples). The more usual approach however, is maybe to use PCA for the matrix of continuous environmental measurements, and CA for the species count data.

Correspondence analysis is a multivariate method used for the analysis of categorical data in the form of contingency tables. It is also used in a graphical, exploratory sense for tables of nonnegative count data, where the data are strictly speaking not given in the form of a contingency table. CA has often been employed as a tool for the analysis of multivariate species count data, e.g. a matrix of species counts at different locations, in order to obtain an ordination diagram of species and sites. Such ordination diagrams were initially made by ecologists using a reciprocal averaging algorithm, which is equivalent to CA (Hill, 1974).

The theory of correspondence analysis has extensively been described by Greenacre (1984) and can also be found in other textbooks on multivariate analysis such as chapter 8 in Gifi (1981), or section 8.5 in Mardia (1979).

We briefly mention the different approaches one could use to introduce CA. In

the Gifi-system of multivariate analysis (Gifi, 1981), CA is considered to be a special case of homogeneity analysis (Michailidis and de Leeuw, 1998, section 2.5.1). In homogeneity analysis a loss function is defined which is minimized with respect to both scores for cases and classes of categorical variables, using particular restrictions. An alternating least squares algorithm is employed to compute the solution.

The French approach to CA is essentially geometrical, and described by Benzécri (1973) and Greenacre (1984; 1993b). Here, the data table (called abundance matrix in this ecological context) is expressed as a set of profile vectors. In geometrical terms, the profile vectors form a cloud of points in high-dimensional space, and the object is to fit an optimal plane that best approximates this cloud in a least squares sense.

Yet another approach to CA is Nishisato's (1980) dual scaling (also called optimal scaling (Nishisato, 1996, p. 563)). In dual scaling the purpose is to assign a real number to the categories of a categorical variable, as if we would transform a categorical variable into a quantitative one (quantification). This is achieved by maximizing the so-called correlation ratio (the ratio of the between sum of squares and total sum of squares of the quantifications), which is algebraically equivalent to the eigenvalue problem of CA (Greenacre, 1984, section 4.3). However, dual scaling seems to be applicable to a wider variety of categorical data than is CA (Nishisato, 1996).

This chapter presents a brief account of the theory, following the notation of Greenacre, and serves mainly for the purpose of reference in the next chapters (7,9). The chapter also exposes some more theoretical details concerning correspondence analysis, e.g. bounds obtained for singular values (inertias), biplots of correspondence analysis, etc. Applications with the marine biological survey data are dealt with in the next chapter.

5.2 Basic Theory

We consider an $I \times J$ matrix $\mathbf{N}$ with all elements non-negative, $n_{ij} \geq 0$. From this matrix we construct the correspondence matrix $\mathbf{P}$ formed by dividing all elements of $\mathbf{N}$ by the sum of all elements of $\mathbf{N}$:

$$\mathbf{P} = \mathbf{N}/\mathbf{1}'\mathbf{N}\mathbf{1}. \tag{5.1}$$

Would we multiply our data in $\mathbf{N}$ by a scalar, then the sum of all elements of $\mathbf{N}$ would also be multiplied by the same scalar, and thus the correspondence matrix would remain the same. Since $\mathbf{P}$ is at the heart of the analysis that is to follow, it is clear that CA is invariant with respect to multiplication of the data by a scalar.

We introduce two column vectors $\mathbf{r}$ and $\mathbf{c}$ containing the row and column sums (also called row and column masses) of $\mathbf{P}$ respectively, and build diagonal matrices $\mathbf{D}_r$ and $\mathbf{D}_c$ from these vectors:

$$\mathbf{r} = \mathbf{P1}, \qquad \mathbf{c} = \mathbf{P}'\mathbf{1}, \qquad \mathbf{D}_r = \mathrm{diag}(\mathbf{r}), \qquad \mathbf{D}_c = \mathrm{diag}(\mathbf{c}). \tag{5.2}$$

If rows and columns would be independent then the elements of the correspondence matrix $\mathbf{P}$ could simply be calculated as the products of the corresponding marginals: $p_{ij} = r_i c_j$. In CA we precisely study the deviations from this independence model, and do the following least squares approximation to these deviations:

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c{}^{-1/2} = \bar{\mathbf{U}}\mathbf{D}\bar{\mathbf{V}}', \tag{5.3}$$

with identification conditions $\bar{\mathbf{U}}'\bar{\mathbf{U}} = \mathbf{I}$ and $\bar{\mathbf{V}}'\bar{\mathbf{V}} = \mathbf{I}$. Principal and standard coordinates for rows ($\mathbf{F}$ and $\mathbf{\Phi}$ respectively) and columns ($\mathbf{G}$ and $\mathbf{\Gamma}$ respectively) are obtained as:

$$\begin{aligned} \mathbf{F} &= \mathbf{D}_r^{-1/2}\bar{\mathbf{U}}\mathbf{D}, \quad & \mathbf{G} &= \mathbf{D}_c^{-1/2}\bar{\mathbf{V}}\mathbf{D}, \\ \mathbf{\Phi} &= \mathbf{D}_r^{-1/2}\bar{\mathbf{U}}, & \mathbf{\Gamma} &= \mathbf{D}_c^{-1/2}\bar{\mathbf{V}}. \end{aligned} \tag{5.4}$$

Expression (5.3) clarifies how the data a weighted prior to the search of an optimal plane of representation. The weighting is maybe best understood in terms of the correspondence matrix $\mathbf{P}$, as this relates directly to our data in $\mathbf{N}$. The correspondence matrix $\mathbf{P}$ gets premultiplied by $\mathbf{D}_r^{-1/2}$ and postmultiplied by $\mathbf{D}_c^{-1/2}$. Column categories with a high associated mass have their elements divided by the square root of that mass, and are so downweighted with respect to categories with a low mass. The same holds for the row categories. In our ecological context it means that rare species and sites with few organisms are upweighted in the analysis.

5.3 Variations on a Computational Theme

Instead of working with a matrix of deviations from independence, one can also do CA by working with a matrix of profiles. From the correspondence matrix we build the matrix of row profiles, by dividing each row by its sum. The matrix of row profiles $\mathbf{R}$ and the matrix of column profiles $\mathbf{C}$ can then be expressed as:

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P}, \qquad \mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}'. \tag{5.5}$$

Notice that the vector of column masses $\mathbf{c}$ is just the vector of the weighted averages of the row profiles, where the weights are the row masses $\mathbf{r}$, since: $\mathbf{r}'\mathbf{D}_r^{-1}\mathbf{P} = \mathbf{1}'\mathbf{P} = \mathbf{c}'$. Similarly, the row masses also equal the weighted averages of the column profiles. The profiles are now centred with respect to the average row profile, $\mathbf{c}$, to obtain the centred row profiles $\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}'$. The centring operation can also be performed by the postmultiplication of the row profiles by an idempotent centring matrix $\mathbf{I} - \mathbf{1}\mathbf{c}'$. In the next step of the analysis, we try to get an optimal display of the centred row profiles in a subspace of low dimensionality, usually a two-dimensional graph. Geometrically we can imagine this as stacking a plane into the multidimensional cloud of profile vectors that is as "close" to the cloud as possible. This is done by a (weighted) least squares approximation to the matrix of centred row profiles, which can be achieved by the generalized singular value decomposition:

$$\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}' = \mathbf{U}\mathbf{D}\mathbf{V}'. \tag{5.6}$$

The left singular vectors satisfy the identification conditions $\mathbf{U}'\mathbf{D}_r\mathbf{U} = \mathbf{I}$ and the right singular vectors $\mathbf{V}'\mathbf{D}_c^{-1}\mathbf{V} = \mathbf{I}$. A similar decomposition can be performed on the column profiles. The singular value decomposition gives us the axes of the optimal subspace. In this context, the columns of $\mathbf{V}$ form the basis of the optimal subspace for the row profiles. The coordinates of the profiles in the optimal subspace, the principal coordinates, are given by equation (5.7). In the same way, using the decomposition of the column profiles, the principal coordinates of the latter ($\mathbf{G}$) can be obtained as well:

$$\mathbf{F} = \mathbf{U}\mathbf{D}, \qquad \mathbf{G} = \mathbf{D}_c^{-1}\mathbf{V}\mathbf{D}. \tag{5.7}$$

We have so arrived at an optimal representation of the row profiles. The principal coordinates have a weighted mean 0, and the weighted variance of these coordinates are precisely the elements of $\mathbf{D}^2$, since $\mathbf{r}'\mathbf{F} = \mathbf{r}'(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{D}_c^{-1}\mathbf{V} = (\mathbf{1}'\mathbf{P} - \mathbf{c}')\mathbf{D}_c^{-1}\mathbf{V} = (\mathbf{c}' - \mathbf{c}')\mathbf{D}_c^{-1}\mathbf{V} = \mathbf{0}$, and $\mathbf{F}'\mathbf{D}_r\mathbf{F} = \mathbf{D}\mathbf{U}'\mathbf{D}_r\mathbf{U}\mathbf{D} = \mathbf{D}^2$. In the same manner, we derive for the weighted mean and weighted variances of the principal column coordinates that $\mathbf{c}'\mathbf{G} = \mathbf{0}$ and $\mathbf{G}'\mathbf{D}_c\mathbf{G} = \mathbf{D}^2$.

It would be interesting to have, in the same plane, representations for the columns of the data table (the sites). We can think of a column category as an extreme profile, a unit vector with all subjects concentrated in one category only. In our ecological context, this means that we represent a site as a theoretical species, a species that only occupies that one particular site. The representation of the column categories can then be obtained by projecting these theoretical profiles ("vertices") onto the optimal plane. We thus form a $J \times J$ identity matrix, and can obtain the vertices or standard column coordinates, with respect to the basis of the plane as:

$$\mathbf{\Gamma} = \mathbf{D}_c^{-1}\mathbf{V}, \tag{5.8}$$

where the identification conditions of (5.6) now imply $\mathbf{\Gamma}'\mathbf{D}_c\mathbf{\Gamma} = \mathbf{I}$. When the decomposition of the column profiles is considered, we find in an analogous manner the standard row coordinates $\mathbf{\Phi} = \mathbf{U}$, where $\mathbf{\Phi}'\mathbf{D}_r\mathbf{\Phi} = \mathbf{I}$. The standard coordinates have a weighted mean of zero because $\mathbf{c}'\mathbf{\Gamma} = \mathbf{c}'\mathbf{G}\mathbf{D}^{-1} = \mathbf{0}$ and $\mathbf{r}'\mathbf{\Phi} = \mathbf{r}'\mathbf{F}\mathbf{D}^{-1} = \mathbf{0}$.

Decomposition (5.6) can easily be obtained from (5.3); the singular vectors of decomposition (5.3) are related to the singular vectors of (5.6) by $\bar{\mathbf{U}} = \mathbf{D}_r^{1/2}\mathbf{U}$ and $\bar{\mathbf{V}} = \mathbf{D}_c^{-1/2}\mathbf{V}$.

The analysis based on the profiles has maybe some intuitive appeal, as we can imagine our data as profile vectors which we want to represent optimally. Equation (5.6) can, however, be rewritten in many ways. We can for instance, also work with the profiles without centring them. This has the consequence that we will find an extra (trivial) dimension in the solution with an associated singular value of 1, a left singular vector $\mathbf{1}$ and a right singular vector $\mathbf{c}$. The centring operation is usually done just with the purpose of omitting this trivial dimension. CA has the particular property that dimension k of the solution of the

centred data precisely equals dimension $k+1$ of the solution of the uncentred data. For the sake of comparison, we note that in principal components analysis this is not the case. In principal component analysis one usually does a singular value decomposition of the centred data matrix of continuous variables, but the so obtained singular values and singular vectors will usually not appear in the singular value decomposition of the raw uncentred data.

Note that (5.6) can also be reexpressed in such a way that we do a decomposition of $\mathbf{P}$, or if one wants, of the raw data $\mathbf{N}$. Through algebraic manipulation the corresponding identification conditions can be easily derived as well as the modified expressions for the principal coordinates. The point is that the final output of the analysis, the numerical values of $\mathbf{F}$, $\mathbf{G}$, $\mathbf{\Gamma}$ and $\mathbf{\Phi}$ will always be the same.

A singular value decomposition can always be rephrased as an eigenvector-eigenvalue decomposition (also called the spectral decomposition). If we call $\mathbf{T} = \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$, using the non-centred version of (5.3), then the characteristic equations of correspondence analysis become:

$$\begin{aligned} \mathbf{T}'\mathbf{T} = &\quad \mathbf{D}_c^{-1/2}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1/2} = \quad \bar{\mathbf{V}}\mathbf{D}^2\bar{\mathbf{V}}', \\ \mathbf{T}\mathbf{T}' = &\quad \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1/2} = \quad \bar{\mathbf{U}}\mathbf{D}^2\bar{\mathbf{U}}'. \end{aligned} \tag{5.9}$$

By postmultiplying equation (5.6) by $\mathbf{D}_c^{-1}\mathbf{V}'$ we obtain the equation:

$$\mathbf{F} = (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{D}_c^{-1}\mathbf{V} = (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{\Gamma} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{\Gamma}. \tag{5.10}$$

This well-known result expresses that the principal coordinates are weighted averages of the standard column coordinates, the weights being given by the elements of the row profiles. These relationships are known as the *transition formulae* or barycentric relationships. It also makes clear that the principal coordinates are always "interior" with respect to the standard coordinates, and that the principal coordinates will coincide with the vertex points if the profiles are all elementary vectors. Similarly, from the decomposition of the column profiles (equation (5.15) below) we derive that the principal coordinates of the columns are weighted averages of the standard row coordinates:

$$\mathbf{G} = (\mathbf{D}_c^{-1}\mathbf{P}' - \mathbf{1}\mathbf{r}')\mathbf{\Phi} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{\Phi}. \tag{5.11}$$

We still need to specify precisely what we mean by "closeness" of the profiles to the optimal plane. The distance measure used in correspondence analysis is not the ordinary Euclidean distance, but the so-called χ^2-distance. The squared χ^2-distance between two particular row profiles $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ is given by:

$$d^2(\mathbf{x}_i, \mathbf{x}_{i'}) = (\mathbf{x}_i - \mathbf{x}_{i'})'\mathbf{D}_c^{-1}(\mathbf{x}_i - \mathbf{x}_{i'}) = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2. \tag{5.12}$$

The difference between the Euclidean distance and the χ^2-distance lies in the factor $1/c_j$, which gives the rarer column categories a relatively larger contribution to the χ^2-distance. The total dispersion in the multidimensional cloud of profiles, called the total *inertia*, is measured by a weighted average of squared

χ^2-distances between profile vectors and the average profile, and is expressed as:

$$\sum_{i=1}^{I} r_i(\mathbf{x}_i - \mathbf{c})'\mathbf{D}_c^{-1}(\mathbf{x}_i - \mathbf{c}) = \text{tr}(\mathbf{D}_r(\mathbf{R} - \mathbf{1c}')\mathbf{D}_c^{-1}(\mathbf{R} - \mathbf{1c}')') = \text{tr}(\mathbf{D}^2), \quad (5.13)$$

where we substituted (5.6). Thus, the total dispersion or inertia is precisely the sum of the squared singular values. Each dimension in the solution gives a contribution to the total inertia given by one element on the diagonal of $\mathbf{D}^2$, and these contributions are called the principal inertias. The plane "closest" to the cloud of profiles is the plane for which the weighted sum of χ^2-distances to the plane is minimal. The actual minimization problem of CA is maybe best understood in terms of decomposition (5.3). If we do a rank 2 approximation to the data, we want to minimize the sum of the squared errors, that is we minimize the squared Euclidean matrix norm:

$$\| \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} - \bar{\mathbf{U}}_{(2)}\mathbf{D}_{(2)}\bar{\mathbf{V}}'_{(2)} \|_E^2 = \| \bar{\mathbf{U}}_{(r)}\mathbf{D}_{(r)}\bar{\mathbf{V}}'_{(r)} \|_E^2 =$$

$$\text{tr}(\bar{\mathbf{V}}_{(r)}\mathbf{D}_{(r)}\bar{\mathbf{U}}'_{(r)}\bar{\mathbf{U}}_{(r)}\mathbf{D}_{(r)}\bar{\mathbf{V}}'_{(r)}) = \text{tr}(\mathbf{D}^2_{(r)}),$$

where we use $\bar{\mathbf{U}}_{(2)}$ to indicate the first two columns of $\bar{\mathbf{U}}$ and $\bar{\mathbf{U}}_{(r)}$ to indicate the remaining columns. Minimizing squared errors thus corresponds to minimizing the inertia in the remaining dimension. Because the total inertia is a constant, it means that we maximize inertia in the 2-dimensional plane. The equation above can be seen as the loss function of correspondence analysis.

We note that the inertia of a data table is related to the well-known χ^2-statistic used for testing for independence of rows and columns of the table since:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - nr_ic_j)^2}{nr_ic_j} = n\sum_i \sum_j \frac{(p_{ij} - r_ic_j)^2}{r_ic_j} =$$

$$n\text{tr}(\mathbf{D}_r^{-1}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{rc}')') = n\text{tr}(\mathbf{D}^2),$$

where $n = \mathbf{1}'\mathbf{N}\mathbf{1}$. In this thesis however, we will hardly ever calculate this χ^2-statistic. Using the χ^2-statistic for inference presupposes that the data in the contingency table constitutes a random sample of observations for which two categorical variables have been recorded. With our matrix of species counts the sample does not consist of the sum of all elements of the table. Here, a sample is one column in the abundance matrix (one grab). The abundance matrix is thus a compilation of information from different samples, and it would be inappropriate to calculate a χ^2-statistic.

5.4 Biplots in Correspondence Analysis

Decomposition (5.6) shows that the profiles can be factored as the product of the matrices $\mathbf{F}$ and $\mathbf{\Gamma}$:

$$\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c}' = \mathbf{UDV}' = \mathbf{F}(\mathbf{D}_c\mathbf{\Gamma})'. \quad (5.14)$$

This is a form of the biplot factorization described by Gabriel (1971) and Gabriel and Odoroff (1990). The joint plot of the first two columns of $\mathbf{F}$ and the first two rescaled columns of $\mathbf{\Gamma}$ so gives an optimal 2-D representation of the data, where the individual data points are represented by the scalar products between the rows of $\mathbf{F}$ and $\mathbf{\Gamma}$. This particular biplot is called the asymmetric map of the row profiles. Alternatively, the decomposition of the centred column profiles gives:

$$\mathbf{D}_c^{-1}\mathbf{P}' - \mathbf{1}\mathbf{r}' = \mathbf{G}(\mathbf{D}_r\mathbf{\Phi})', \tag{5.15}$$

which offers another biplot, and is called the asymmetric map of the column profiles. Both factorizations described by (5.14) and (5.15), are approximations to the matrices of profiles that are optimal in the least squares sense. As described in detail by Greenacre (1993a), vectors from the origin of the biplot to the vertices can be calibrated, that is, tick marks could be drawn on those vectors, allowing one to approximately read off the original data in the biplot like one would do in a scatterplot.

However, the graphical output of a correspondence analysis is often reported in the form of a symmetric map, that is by jointly plotting the columns of $\mathbf{F}$ and $\mathbf{G}$. Such a map must be interpreted with care, and is not a biplot (Greenacre, 1993b, chapter 13). We pursue this point here in some detail. If we consider the scalar products between the rows of $\mathbf{F}$ and $\mathbf{G}$ we have:

$$\mathbf{F}\mathbf{G}' = \mathbf{U}\mathbf{D}^2\mathbf{V}'\mathbf{D}_c^{-1} = (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{\Gamma}\mathbf{D}\mathbf{\Gamma}'. \tag{5.16}$$

This shows that the scalar products in $\mathbf{F}\mathbf{G}'$ are not approximations to the profiles, but to a linear transformation of the profiles. The transformation seems not to correspond to a simple rotation or stretching. We do not recover our original data, but elements of matrix that is of no particular interest, and that seems neither to be optimally approximated in the least squares sense.

We want to draw some attention to the interpretation of the distances between the vertex points (standard column coordinates) and the origin of the display. In general, frequent column categories will tend to lie in the centre of the biplot, whereas rare categories will often appear towards to border of the display. This can be inferred from the standardization $\mathbf{\Gamma}'\mathbf{D}_c\mathbf{\Gamma} = \mathbf{I}$. Imagine we consider all dimensions in the CA solution, including the trivial one, a first column of ones in matrix $\mathbf{\Gamma}$. Matrices $\mathbf{\Gamma}$ and $\mathbf{\Gamma}\mathbf{\Gamma}'$ are then square and of full rank and we can write:

$$\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}_c\mathbf{\Gamma}\mathbf{\Gamma}' = \mathbf{\Gamma}\mathbf{\Gamma}', \tag{5.17}$$

from which we obtain that $\mathbf{\Gamma}\mathbf{\Gamma}' = \mathbf{D}_c^{-1}$. If we define $\hat{\mathbf{\Gamma}}$ to be the solution without the trivial dimension, that is $\mathbf{\Gamma} = \left[\mathbf{1} \mid \hat{\mathbf{\Gamma}}\right]$ then:

$$\mathbf{\Gamma}\mathbf{\Gamma}' = \begin{bmatrix} \mathbf{1} & \hat{\mathbf{\Gamma}} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \hat{\mathbf{\Gamma}}' \end{bmatrix} = \mathbf{1}\mathbf{1}' + \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}' = \mathbf{D}_c^{-1}, \tag{5.18}$$

and thus $\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}' = \mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}'$. When we use all columns of the CA solution, except the trivial column of ones, we can build an idempotent centring matrix that,

when applied to a vector, centres it on the weighted mean. This centring matrix is given by:

$$\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}'\mathbf{D}_c = (\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}')\mathbf{D}_c = \mathbf{I} - \mathbf{1}\mathbf{c}', \tag{5.19}$$

which is seen to be idempotent because $(\mathbf{I}-\mathbf{1}\mathbf{c}')'(\mathbf{I}-\mathbf{1}\mathbf{c}') = \mathbf{I}-\mathbf{1}\mathbf{c}'-\mathbf{1}\mathbf{c}'+\mathbf{1}\mathbf{c}'\mathbf{1}\mathbf{c}' = \mathbf{I}-\mathbf{1}\mathbf{c}'$. This matrix will be used extensively in chapter 7.

A CA has $\min(I-1, J-1)$ dimensions in the solution. In the above, we tacitly assumed $J < I$, which is usually the case in ecological research, as there are normally more species than samples. If however, the number of species determines the number of dimensions in the solution, because there are fewer species than sites, then evidently $\mathbf{\Gamma}\mathbf{\Gamma}'$ is singular and (5.18) does not hold any more.

Analogous results can be derived for the standard row coordinates, giving the idempotent centring matrix $\mathbf{\Phi}\mathbf{\Phi}'\mathbf{D}_r$ when $I < J$.

The diagonal elements of $\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}'$ are the squared euclidean distances of the vertex points from the origin, so that the distance of a vertex point to the origin in biplot (5.14) is given by $1/\sqrt{c_j} - 1$. Thus, if a column category is rare, c_j will be small, and its distance to the origin will be large. Conversely, a large column weight gives a small distance to the origin. Note that this is a "full space" result, meaning that it will be exact when we have a data matrix of three columns only that is perfectly represented in 2-dimensional space. For larger data matrices with more columns, the distances of the vertex points to the centre of the map will only be approximately $1/\sqrt{c_j} - 1$, where at the moment we ignore whether this approximation is optimal in any sense.

5.5 Bounds for Principal Inertias

The singular values in decompositions (5.6) and (5.3) turn out to be always in the [0,1] interval. This is explained by Greenacre (1984, pp. 108-116) by showing that principal inertias correspond to squared canonical correlations obtained in a canonical correlation analysis of the indicator matrices corresponding to a contingency table. Neudecker, Satorra and van den Velden (1997) formulated a fundamental matrix result on scaling in multivariate analysis, stating that any matrix $\mathbf{T}$ of the from $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$, matrix $\mathbf{P}$ being non-negative, has singular values in the interval [0,1]. Notice that $\mathbf{T}$ is precisely the noncentred version of equation (5.3). A algebraical proof of this result based on Gergshgorin's theorem was given by Graffelman (1998), and is included below.

Consider the singular value decomposition of $\mathbf{T}$ as $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}'$ with $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$. The singular values of matrix $\mathbf{T}$ are the square roots of the eigenvalues of matrix $\mathbf{T}'\mathbf{T}$ since $\mathbf{T}'\mathbf{T} = \mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\mathbf{D}^2\mathbf{V}'$. Bounds on the eigenvalues therefore imply bounds on the singular values.

First, since $\mathbf{T}'\mathbf{T}$ is a nonnegative definite matrix, all its eigenvalues are nonnegative, and so the singular values of $\mathbf{T}$ are also nonnegative.

Bounds for the largest eigenvalue λ_F of a nonnegative matrix $\mathbf{A} = \mathbf{T}'\mathbf{T}$ can be obtained by applying Gershgorin's theorem. Barbolla and Sanz (1998, pp. 336-338) give a detailed derivation of these bounds, and we use their result:

$$\min_i \{\sum_{j=1}^{p} a_{ij}\} \leq \lambda_F \leq \max_i \{\sum_{j=1}^{p} a_{ij}\}. \tag{5.20}$$

Note that $\mathbf{A} = \mathbf{T}'\mathbf{T} = \mathbf{D}_c^{-1/2}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1/2}$ has the same eigenvalues as $\mathbf{B} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}$ since $\mathbf{D}_c^{-1/2}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1/2}\mathbf{v} = \lambda\mathbf{v}$ implies $\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1/2}\mathbf{v} = \lambda\mathbf{D}_c^{-1/2}\mathbf{v}$ and so $\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{u} = \lambda\mathbf{u}$, where $\mathbf{u} = \mathbf{D}_c^{-1/2}\mathbf{v}$.

The rows of $\mathbf{B}$ sum to 1 since $\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{1} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{r} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{1} = \mathbf{D}_c^{-1}\mathbf{c} = \mathbf{1}$. Applying Gershgorin's theorem we obtain bounds for the eigenvalues of $\mathbf{B}$:

$$\min_i \{\sum_{j=1}^{p} b_{ij}\} = \max_i \{\sum_{j=1}^{p} b_{ij}\} = 1.$$

So $\lambda_F = 1$, which shows that the matrix $\mathbf{B}$, and so $\mathbf{A}$, always has an eigenvalue of 1, and that this is the largest one. Consequently, the singular values of $\mathbf{T}$ lie all in the closed interval [0,1], and there will always be a singular value of 1.

Alternative proofs were given by Puntanen and Styan (1998). Yet another elegant proof exists (van de Velden, personal communication; van de Velden et al. (1999)), based on the fact that eigenvalues are bounded by the matrix norm (Graybill, 1983, p. 98).

5.6 Some Extreme Cases

In this section we consider the correspondence analysis of a few extreme data matrices, which are very unlikely to occur in practice, but which can serve as a reference pictures. In the first place, consider a square matrix of profiles where all subjects (species or whatever) are concentrated into one particular column category (e.g. a site), and that each column category also only contains items of one particular subject. Such a data matrix $\mathbf{N}$ has the strongest possible association between its column and row categories. If there are as many rows as columns, then they could be reordered in such a way as to obtain a diagonal data matrix, and we have that $\mathbf{P} = \mathbf{N}/\mathbf{1}'\mathbf{N}\mathbf{1} = \mathbf{D}_r = \mathbf{D}_c$. The matrix of profiles will then be the identity matrix, $\mathbf{D}_r^{-1}\mathbf{P} = \mathbf{D}_r^{-1}\mathbf{D}_r = \mathbf{I}$. By the transition formulae (5.10), vertices and row profiles will coincide ($\mathbf{F} = \mathbf{\Gamma}$), and since the singular values of an identity matrix are all 1, all principal inertias are 1, and the profiles attain their maximal dispersion. The total inertia for such data has the maximum possible value of $J - 1$.

On the other end of the extreme is a data matrix with no association between the rows and the columns whatsoever. Such a data matrix is in perfect agreement with the independence model, and each element of the correspondence matrix p_{ij} can be obtained as the product of the elements of the masses $r_i c_j$. In that case the matrix of row profiles is of rank one, since $\mathbf{D}_r^{-1}\mathbf{P} = \mathbf{D}_r^{-1}\mathbf{r}\mathbf{c}' = \mathbf{1}\mathbf{c}'$,

and there will exist no solutions beyond the trivial one, and the total inertia is 0.

All data sets that can arise in practice, will be somewhere in between these two extremes. A MATLAB program for performing CA is given in appendix A.4.

Chapter 6

Applications of Correspondence Analysis and Principal Component Analysis

6.1 Introduction

In this chapter we keep in first instance the biological and chemical data separate. We discuss applications of correspondence analysis to the species data in some detail, and with attention for some particular topics: resampling techniques to study the stability of the obtained ordination diagrams, the comparison of CA maps obtained from replicates and aggregated data and the use of the LONG matrix to study all replicates simultaneously

From an ecological point of view, it is interesting to see whether ordinations obtained over the different years are similar, or very different. In particular, ecologists are interested to see how and why the species composition of a community changes over time. We therefore select a subset of stations that has been sampled every year, and compare the respective ordinations obtained. We will make occasional use of procrustes analysis to compare ordinations. For analyzing the environmental data matrix we use principal component analysis, and compare the biplots produced for the three successive years. We also do some attempts to perform an integrated analysis of data from all years simultaneously.

A particular analysis is often based on data from one particular year. The sampling is however, annually repeated, so that the full data set does not consist of a single matrix, but is better considered a three-way matrix (or data "cube") in which time is the third dimension. One could easily be led into thinking that the data at hand are *longitudinal*, though this is strictly speaking not

the case. Longitudinal data tables often arise in social surveys where a group of respondents is classified on two categorical variables at several time points. Often the *same set* of respondents is being interviewed each time, making that the total of the data table is in principle fixed. In our context, the organisms are disposed of after sampling, and each year a different group of organisms is collected. As a consequence, the grand total of the data table is not fixed, but is a random variable. Data of this type has been called "trend data" (van der Heijden, 1987, pp. 89).

6.2 Biological Data

As a starting point, we perform CA for the species data from 1990. The abundance matrix (sum of five replicates) consists of 152 species at 39 locations with a total of 22.280 organisms. 1990 is the year with the largest number of samples. For 1991 we study a subset of 36 species at 12 locations that sum to a total of 3791 organisms. CA is applied at the replicate level and procrustes rotation is employed to compare ordinations. Next, the 1992 data will be treated (totalling 9445 organisms, 166 species and 12 locations) with some consideration for stability issues. The section closes with a study with integrated use of the data from three successive years.

6.2.1 The CA of 1990

Figure 6.1 shows the asymmetric map of the row profiles for 1990. The two dimensional map captures 59% of the total inertia of the data matrix. More details of the inertia decomposition are given in table 6.1. The first axis contrasts the stations 30,31,36 and 37 with the rest, whereas the second axis opposes stations 24 and 15 with the rest, with 40 on the other extreme. Species with a quality in 2-D larger than 0.5 (that is to say with more that 50% of their inertia accounted for by the display) are labelled in the map.

Dim.	Inertia	%	% Cum.
1	0.9572	35.89	35.89
2	0.6175	23.15	59.04
3	0.2460	9.22	68.27
4	0.1087	4.07	72.34
5	0.0831	3.12	75.46
⋮	⋮	⋮	⋮

TABLE 6.1: INERTIA DECOMPOSITION OF SPECIES DATA, 1990

Of all 152 species only two have determined the orientation of the principal axes. The relative contributions of all species to the first and the second axis are very small, except for *Capitella capitata* and *Myriochele oculata*. *Capitella capitata* contributes 55% to the inertia of the first axis, whereas *Myriochele oculata* contributes 80% to the second axis. The most salient features of this analysis are thus the very high abundance of *Capitella capitata* at stations 30,31,36

and 37, and the very high abundance of *Myriochele oculata* at stations 24 and 15. Indeed, when we check the data matrix we find that *Capitella capitata* is extremely abundant at the four stations mentioned, is also present at station 3, but practically absent everywhere else.

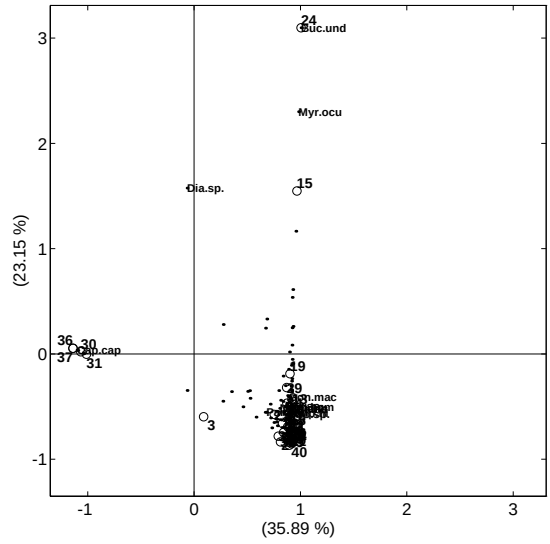

FIGURE 6.1: ASYMMETRIC MAP OF ROW PROFILES FOR 1990

Some species like *Buccinum undatum* and *Diastylis* have a good quality of representation, as they are accidentally close to the optimal plane, though they do not contribute to the axes, as these are very rare species. We see that, although CA is known to downweight frequent species, the highly frequent ones still dominate the analysis. Most stations and species cluster very close together at the bottom of the diagram, their interpretation being obscured by the outliers. We can consider deletion of one or more outliers. For instance, when we delete *Capitella capitata*, the newly obtained map captures about 44% of the inertia of the remaining 151 species, and what was the second dimension in the previous analysis, the contrast 24,15 versus 40 then essentially becomes the first dimension in the analysis, with *Myriochele oculata* contributing 81% to the inertia of this dimension. The second dimension then captures a distinction between 40 and the group 30,36,37 and 3, the latter group sorted out as being more high on *Nemertini indet.* which is the main contributor of this new second axis (22%). The process of deleting influential species can be repeated ad infinitum, and in fact, it is surprising to see that often only one or two species determine the orientation of a principal ax.

6.2.2 The CA of 1991; Replicates and Sums

There are five biological replicates, and the CA shown in the previous section is based on the sum of these five replicates. But how consistent is the informa-

tion provided by these five replicates? This has been considered for individual species with reliability calculations (cf. chapter 2, page 9). Here, we judge this consistency in the multivariate sense by performing CA for each of the five replicates separately, and comparing the ordinations of the sites so obtained. We choose that subset of species that is present in all the replicates. Figure 6.2 shows the ordination diagrams (asymmetric maps of the column profiles) for the sum and the five replicates in 1991. The first dimension in the CA of the sum shows that this dimension opposes reference station 40 with station 15. The same contrast is also detected clearly in the replicates A and B, though in B it pops up as the second dimension in the analysis.

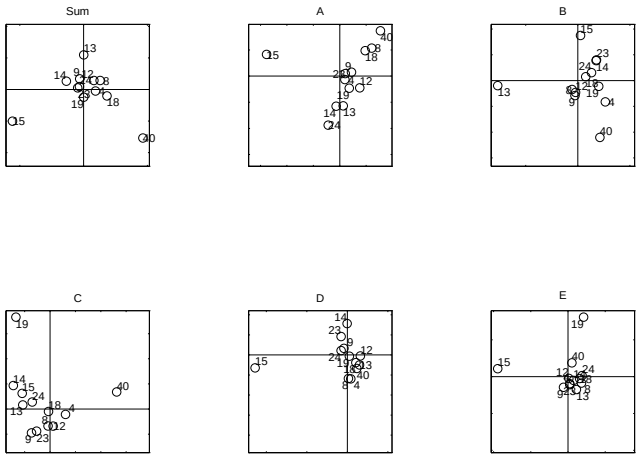

FIGURE 6.2: CA OF 5 REPLICATES AND THEIR SUM, 1991

Replicates C,D and E are consistent with this in the sense that they also show a large distance between stations 40 and 15 along the first axis, though 40 is close to many other stations in replicates D and E. The second dimension of the CA of the sum reveals a contrast between station 13 on one side, and 40 and 15 on the other side. The "triangle" 13,15,40 is also present in replicates A and B. It is clear that replicate B has its first two dimensions interchanged with respect to the CA of the sums. In replicates C and D, the "triangle" has one edge collapsed, as in C 13 and 15 are close and in D 40 and 13 are close. As just outlined, some of the main contrasts found in an analysis of the sums can also be identified in the replicates, though the central cluster of points shows considerable variation. It is difficult to judge the similarity by comparing the ordinations by visual inspection, and we proceed to do this in a more formal way in the next section by procrustes rotation.

6.2.3 Procrustes Rotation of Replicates

Procrustes rotation is a multivariate method for comparing ordinations, designed by Gower (1971; 1975). The mathematics of the method can also be found in Mardia (1979, section 14.7) and Digby and Kempton (1987, chapter 4). There are two types of procrustes rotation. One, called classical procrustes rotation, tries to optimally match one ordination to a second one which is kept fixed, by the operations of translation, rotation, reflection and stretching. The other type of analysis is called generalized procrustes rotation, and considers the problem of finding one consensus ordination based on a series of ordinations, e.g. replicate samples or sampling repeated at different time points. On one

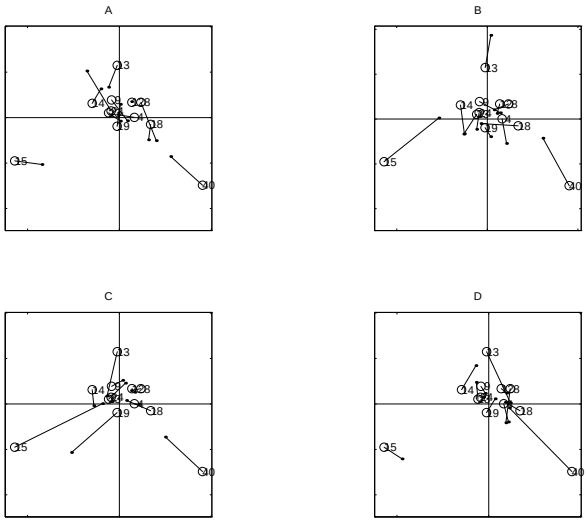

FIGURE 6.3: PROCRUSTES ROTATION OF 4 REPLICATES, 1991

hand, we want to know if there is a large difference between analyzing the sums of the replicates, or individual replicates. We therefore try to match each of the five replicate ordinations to the ordination diagram of the sums. The graphs of four of the five procrustes rotations are shown in figure 6.3, and the corresponding statistics (scale factor and residual sum of squares (RSS)) are given in table 6.2. Lines in these diagrams connect the rescaled rotated ordinations of the replicates to the ordination points based on the sums (open circles), the latter being considered fixed. Long lines indicate large residuals (bad fit). Only two dimensions of the CA solution are considered in the procrustes rotation. Replicate C is the most aberrant from an analysis based on sums, and we see that the largest residuals repeatedly correspond to the extreme points of the ordination (40,15,13). It is clear that there is considerable variation in the ordinations obtained from different replicates, though the contrast 15-40 is always present.

Replicate	p	RSS
A	0.6378	1.1072
B	0.5821	1.6820
C	0.4174	2.3559
D	0.5160	1.7875
E	0.4628	2.1399

TABLE 6.2: STATISTICS OF PROCRUSTES ROTATION

6.2.4 Analysis of the LONG Matrix

In this section we consider another way to compare the analysis of the replicates with an analysis based on their sum. Imagine we stack the five replicates vertically on top of each other, in one large matrix, the LONG matrix. The LONG matrix has a certain total amount of inertia reflecting the variability of the profiles. If we take averages (or sums) of the five samples, and replace the original measurements by these averages (or sums), we obtain a LONG matrix that will in general have a lower inertia, as part of the variation is averaged out. Due to the principle of distributional equivalence (Greenacre, 1984, pp. 65-66), an analysis based on the LONG matrix of (repeated) means will essentially be the same as an analysis based on just one single matrix of means (or sums). This is because the species profiles of the LONG matrix of sums will have five identical profiles for each species, and consequently five coinciding points for each species in the biplot. We would have found the same coordinates as with an analysis based on a single copy of the matrix of sums. The total amount of inertia and its decomposition are the same for the LONG matrix of sums and a single copy of the matrix of sums.

The LONG matrix of the five replicates can be seen as a partitioning of the species. The total inertia (I_t) of the LONG matrix can then be decomposed into a between-sites component (I_b) and a within-sites component (I_w): $I_t = I_b + I_w$. By calculating the percentages $(I_b/I_t) \times 100$ and $(I_w/I_t) \times 100$, we can have an impression of how much inertia is due to variation between stations and how much is due to variation within stations. The latter quantity indicates how much variability we ignore by using sums instead of replicates.

We note here that the inertia of the matrix of the sums of the five replicates is identical to the I_b component of the analysis of the LONG matrix. Column coordinates and the centres of gravity (weighted means) of the replicated species points obtained in the CA of the LONG matrix are usually numerically different from column coordinates and species coordinates obtained in an analysis of the matrix of sums. They differ however, only by a simple rotation. A procrustes analysis of both configurations of points shows that one can be perfectly matched to the other.
Table 6.3 shows the total, between-sites and within-sites inertia for the LONG matrix of each of the three years. For each year, a subset of species has been determined that was present in all five replicates. It is clear that for all the three years a substantial part of over 40% of the total inertia is due to variation

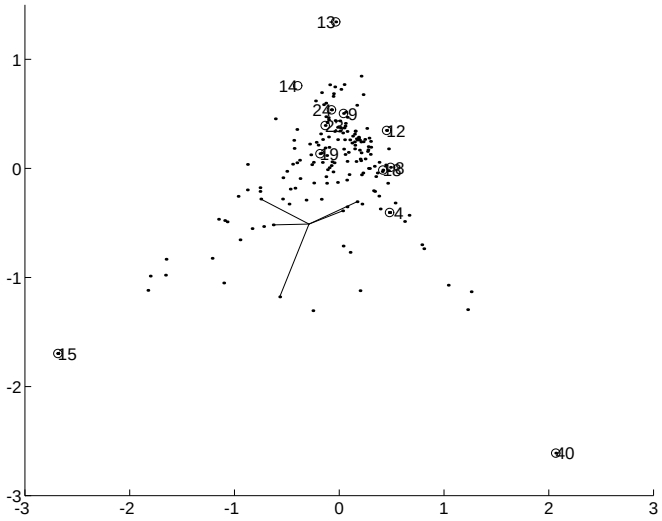

FIGURE 6.4: CA OF THE LONG MATRIX FOR 1991

within stations. Figure 6.4 shows the asymmetric map of the row profiles for the LONG matrix from 1991. For one particular species, *Nemertini indet*, the points of the replicates are connected by lines to their centre of gravity (their weighted mean). The configuration of the station points and the centres of gravity of the species points is very similar to the panel for the sum in figure 6.2.

Year	# spec.	I_t	I_b	I_w	$(I_w/I_t) \times 100$
1990	71	4.1725	2.4299	1.7425	41.7
1991	36	1.2933	0.5419	0.7514	58.1
1992	58	1.7136	0.9709	0.7427	43.3

TABLE 6.3: BETWEEN AND WITHIN INERTIAS FOR THE THREE YEARS

It is also possible to perform an analysis of the LONG matrix where the data from the three years are used simultaneously. In such analysis there is a double partitioning of the species: according to year and replicate. A subset of species needs to be chosen that is present every year and in every replicate, otherwise singularity problems arise. Only 17 species fullfilled this criterion, and these are precisely the most abundant species. The total inertia, I_t, can be decomposed in a between-years and within-years component: $I_t = I_{by} + I_{wy}$. The within-years component I_{wy} can be decomposed in a between-sites and within-sites component: $I_{wy} = I_{bs} + I_{ws}$, so that we get a decomposition of the total inertia as $I_t = I_{by} + I_{bs} + I_{ws}$. For the data at hand, these quantities were $I_t = 0.9428$, $I_{by} = 0.0312$, $I_{wy} = 0.9116$, $I_{bs} = 0.1640$, $I_{ws} = 0.7476$. From these, we compute that only 3.3% percent of the variability is between years, that 96.7%

of the total variability is within years, and that within-sites components makes up the largest part of the total inertia: 79.3%.

6.2.5 The CA of 1992; Stability Issues

It is possible to investigate the stability of the map obtained by CA (Greenacre, 1993b, chapter 20). For instance, if we would have obtained another abundance matrix with the same total amount of organisms, would the ordination remain the same and still separate out station 40? We can get some idea of the variability of the points in the CA map by using resampling techniques. The abundance matrix does not correspond to a particular sampling design where the row or column totals are fixed prior to analysis. We resample the data with the only constraint that the overall sum (n) of the abundance matrix is constant.

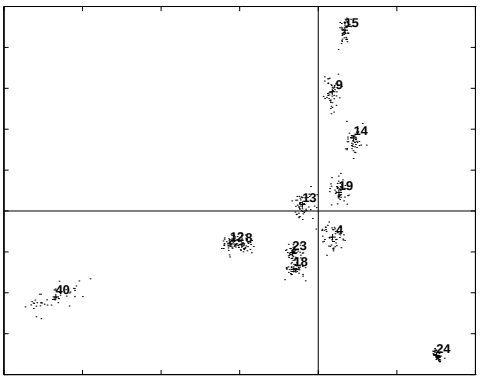

FIGURE 6.5: CA WITH BOOTSTRAP RESAMPLING, 1992

Computationally this can be done as follows. Each cell in the data matrix gets a number. We write out the matrix as a string of length n, and fill this string with the cell numbers, where each cell number occupies as many positions in the string as it had counts in the original data matrix. From this string we sample n cells with replacement, and reassemble this sample into a new data matrix. This procedure can be repeated many times, and so many artificial data sets are generated. By using the transition formulae (equation (5.10), p. 57), we can project each bootstrap sample into the CA map. For the large data matrices considered here, the resampling is a rather computer-intensive exercise. Figure 6.5 shows the result obtained for 100 bootstrap samples of the data of 1992. Figure 6.5 shows that the station points in the map show relatively little variability. Only the pairs of stations 12,8 and 18,23 show some overlap in their positions, the rest of the stations all being well separated from each other. Little modifications of the data thus do not change the ordination, and in this sense the map can be called stable. The maps are not stable though, with respect to deletion of highly abundant species, as was discussed in section 6.2.1.

6.2.6 The Time Dimension

As already stated in the introduction, the survey data consist in fact of a series of abundance matrices pertaining to different time points. We can study each year separately, and compare the annual ordinations obtained, as is illustrated in figure 6.6. These three ordinations are based on a subset of 62 species that was present over all the years and 12 stations that have been sampled every year. The ordination diagrams capture respectively 73%, 55% and 63% of the total inertia for each year. Only the two species with the highest quality (≥ 0.84) are labelled in the display, in order to keep them readable. The ordination from 1991 can be matched to some extend to the one from 1992 if we interchange the first and the second principal axis. Indeed, there are considerable correlations between the principal axes of different years, though we have to be aware that the sample size is small. Most notably, the first axis of 1990 has a correlation of -0.69 with the first axis of 1992, and the first axis of 1991 has correlation -0.75 with the second axis of 1992. The main contributors to the axes are *Myriochele oculata* (axis 1) and *Eudorella sp.*, *Scoloplos armiger* (axis 2) in 1990, *Amphiura filiformis*, *Chaetozone setosa* (axis 1) and *Chaetozone setosa* (axis 2) in 1991, *Myriochele oculata* (axis 1) and *Chaetozone setosa* (axis 2) in 1992. Briefly, we see that in 1990, station 24 and 15 separate out, with *Myriochele oculata* being high on these stations. In 1991 *Chaetozone setosa* is high on 15, in 1992, *Myriochele oculata* is high on 24, and *Chaetozone setosa* on 15. In general, there is a small subset of a few abundant species that tend to dominate in the ordinations, and stations 15,24 and 40 are in general singled out on the extremes of at least one of the principal axes. However, the ordinations vary a lot, and it is very hard to trace the changes that have take place from year to year. We note that it is possible to calibrate the vertex vectors in, for instance, the ordination of 1990. The species points projecting onto the vector between origin and vertex point have abundance higher than average at that site, whereas species points projecting on the other side of the origin have abundance below the average. But each year has a different origin, and because the vector lengths vary from one year to another, the vector for any particular station will have a different calibration each year. Indeed, if we would want to infer from these graphs that species so-and-so is higher on station so-and-so in 1990 than in 1991, it would be necessary to calibrate the vertex vectors in both graphs. And still the procedure would be prone to error, since the projections recover the data approximately, and species so-and-so might have good quality in 1990, but bad quality in 1991. So it remains difficult to trace the changes that have take place from year to year.

We could also consider to treat the successive years in one integrated analysis. For instance, data from 1991 can be mapped into the CA-1990 map as supplementary points. This has the disadvantage that the optimal plane is determined only by the 1990 data, and such a procedure is neither symmetric since projection of the 1990 data into a 1991 optimal plane would give a different result.

A different integrated approach is possible by stacking the annual matrices into one large data matrix. The different ways to do this (columnwise, rowwise and others) are discussed in more detail by van der Heijden (1987) in the context of contingency tables. By concatenating the columns of the annual matrices in the

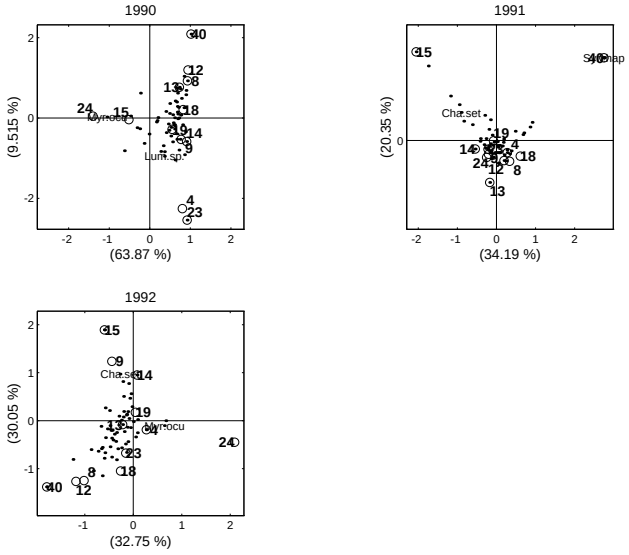

FIGURE 6.6: SEPARATE ANALYSIS OF 3 SUCCESSIVE YEARS.

horizontal sense, the so-called BROAD matrix can be constructed, by concatenating the rows of the annual matrices in the vertical sense we obtain what is known as the LONG matrix. With these data we have, for convenience, stacked the three matrices columnwise, constructing the BROAD matrix. This means that we create one column for each possible combination of site and time. Such a scheme is called "interactive coding" (van der Heijden, 1987, p. 17). The result of the analysis is shown in figure 6.7. Due to the stacking operation, the dimensionality of the problem has increased from 12 to 36, though the two-dimensional plane still captures more than 51% of the total inertia of the stacked matrix.

Stations marked with 'a', 'b' and 'c' correspond to 1990, 1991 and 1992 respectively. In general, stations sampled in 1992 are on the left of the display, stations from 1991 towards the upper right, and stations from 1990 towards the bottom right. In order to aid interpretation, each station has its annual points connect by lines. The map is again determined by only two species, *Myriochele oculata* contributing 65% to the first principal inertia, and *Chaetozone setosa* contributing 59% to the second principal inertia. Contributions from other species are very small. The graph shows the species that have a quality in 2-D that is larger than 50% (*Myriochele oculata* 100%, *Chaetozone setosa* 87%), and whose positions can be interpreted with more confidence. In fact, figure 6.7 basically shows the evolution over the years of the abundance of these species. We see that *Myriochele oculata* is high on most of the stations from 1992, low on the stations from 1991, and higher again on the stations from 1990. The graph shows this most markedly for station 24, where the 1991-point suddenly jumps to the other side of the origin. For some reason or another,

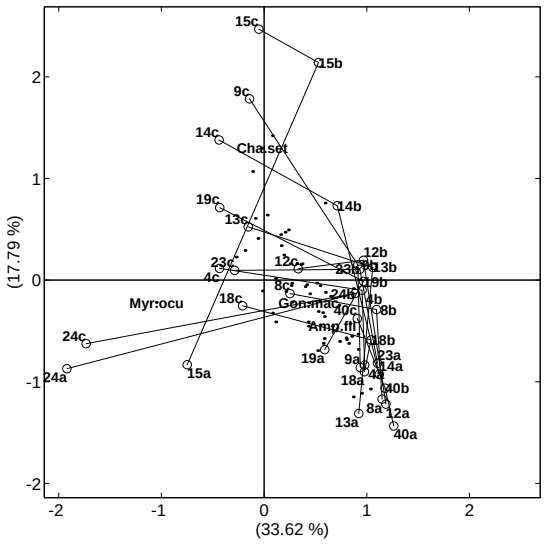

FIGURE 6.7: JOINT ANALYSIS OF 3 SUCCESSIVE YEARS.

Myriochele oculata has suffered a decrease in 1991, and recovered from that in 1992. Species *Chaetozone setosa* also shows a marked pattern. It is low on the 1990 stations, not so low on the 1991 stations, and high on the 1992 stations. Thus, *Chaetozone setosa* displays a pattern of growth over the three successive years. In all these interpretations it is crucial to be aware of the origin of the map, since for valid biplot interpretation we need to project the species points onto vectors from the origin to the station points. Another point of interest is that the three points corresponding to the reference stations are relatively close to each other. In particular, the angles between the vectors pointing from the origin to these three points are small, meaning that if a particular species is high or low on these stations in 1990, the same holds for 1991 and 1992. The graph thus also indicates a more stable species composition for the reference station 40. Last, the group of species close to *Goniada maculata* were relatively abundant in 1990 and 1991 at most stations, and decreased in 1992.

6.3 Chemical Data

Since all environmental variables are continuous, principal component analysis (PCA) is an appropriate technique to obtain a graphical display of the data (a biplot) and to see whether the set of 13 variables can be reduced to a few components that account for most of the variation in the original variables. The theory of principal component analysis can be found in any introductory textbook on multivariate analysis, e.g. chapter 2 in Dillon (1984) or the book by Manly (1989). We will discuss the PCA of the environmental data from 1990 in some detail, and compare the PCA biplots for the three years.

6.3.1 The PCA of 1990

The biplot obtained from a principal component analysis of the 1990 data is shown in figure 6.8. The analysis is successful in the sense that 88% percent

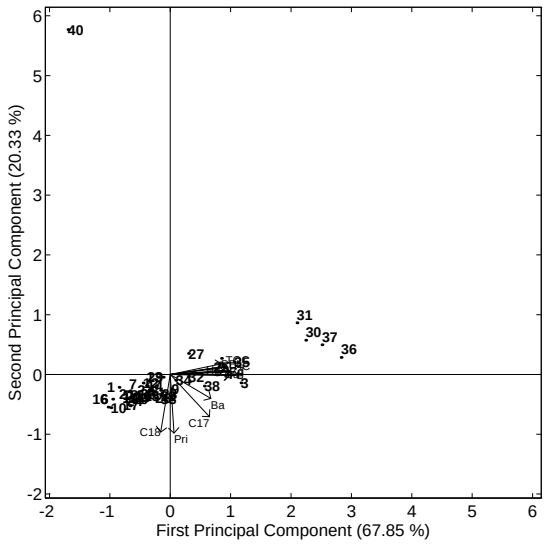

FIGURE 6.8: PCA FOR 1990

of the variation in the log-transformed variables is explained by the first two principal components. The biplot shows that station 40 is an extreme outlier in the analysis, being extremely low on all the measured variables. The chemical composition of this reference station is completely different from the contaminated stations. A set of stations (30,31,36,37) separates from the rest on the first axis, being particularly high on the first principal component, that is, on heavy metals, TOC and THC. This set of four stations is precisely the set of stations that is closest to the platform (cf. figure 2.2, p. 7). These stations turn out to be the most contaminated with respect to heavy metals, TOC and THC. The single outlier (40) dominates the analysis, and makes it difficult to discern more subtle differences between the contaminated stations. The PCA is therefore repeated for the same data set, where station 40 has been omitted from the analysis. The graph of the new analysis is shown in figure 6.9. The eigenvalues and percentages of variance explained are shown for each component in table 6.4. With only two components we account for nearly 88% of the variance in the original variables.

The first principal component is highly correlated with all heavy metals, THC, TOC, Pelite and C17, moderately correlated with Barium, and negatively correlated with Pri and C18. The first principal component can thus be interpreted as a contamination index, though it also seems to capture a *distance-effect*. On the right hand side of the plot we find the group of stations close to the platform, whereas on the left hand side we find the more remote stations like

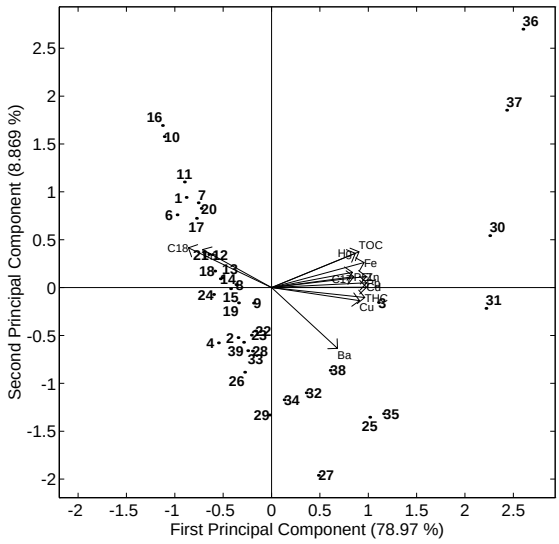

FIGURE 6.9: PCA FOR 1990 WITHOUT STATION 40

16,10,6,1 and 11 (cf. figure 2.1 page 6). This is confirmed by a correlation of -0.62 between distance and PC1, and if we consider log-transformed distance this correlation is as high as -0.84. The outer ring of stations (16,10,6,1,11) is high on the variables Pri and C18. The second component has it highest correlation with barium (-0.63). We note that variable distance is not included in the PCA, as this is a different (non-chemical) variable.

In graph 6.9 the configuration of the station points takes the form of a horseshoe. Horseshoe-effects pop up often in CA or PCA. Not surprising maybe, if we remember that the principal axes are uncorrelated by construction. The absence of linear correlation implies that the cloud of points should have a random spread, or display some pronounced curvature, as is the case here. Indeed, if we regress the second component on the first plus the square of the first, we get an excellent fit and find that 84% of the variance of the second component can be explained by this regression.

We note that the PCA's considered here are based on the standardized log-transformed chemical variables, which implies that we analyze the correlation matrix between the transformed variables.

6.3.2 The Time Dimension

A subset of stations has been chosen that has been sampled every year for the nine variables PEL, THC, TOC, Ba, Cd, Cu, Fe, Pb and Zn, with the objective of comparing the PCA's for the three successive year. The three PCA biplots

Dim.	λ	% Var. Expl.	% Cum.
1	10.27	78.97	78.97
2	1.15	8.87	87.84
3	0.60	4.63	92.48
4	0.31	2.41	94.88
5	0.27	2.08	96.96
6	0.15	1.15	98.12
⋮	⋮	⋮	⋮

TABLE 6.4: EIGENVALUES OF PCA

are shown in figure 6.10.

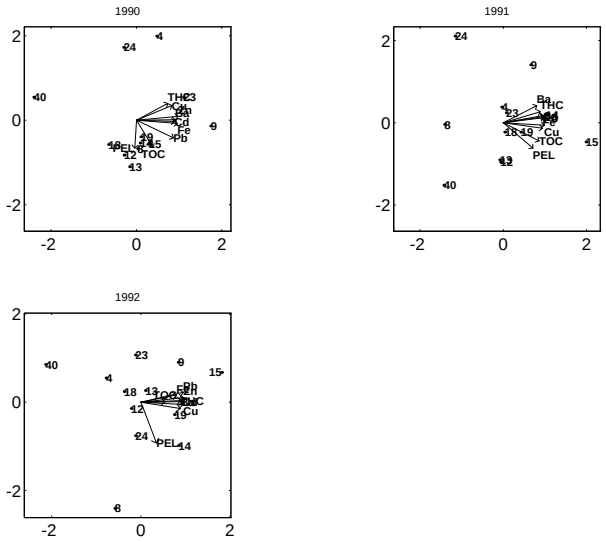

FIGURE 6.10: PCA BIPLOTS FOR THREE SUCCESSIVE YEARS

We note that we mirrored the ordination of 1990 in the vertical axis to obtain a better correspondence with the biplots from 1991 and 1992. This is perfectly acceptable, as the sign of the eigenvectors of the variance-covariance matrix is indeterminate, and does not affect the interpretation (the relative position of the stations with respect to the variable vectors remains the same). A few things can be observed in the graphs in figure 6.10. First, PEL is a variable that over the three year remains more or less perpendicular towards the heavy metal group. In 1990 and 1992, PEL is negatively correlated with the second principal component. For the remaining set of variables it looks like their correlations have increased over the years, as angles between these variables get smaller in the successive diagrams. Station 15 seems to have experienced an increase in contamination from 1990 to 1991. Station 24 seems to have increased with re-

spect to Pelite from 1991 to 1992. Reference station 40 was relatively high on PEL and TOC in 1991. These interpretations were verified by looking at the data matrix. However, the same problems mentioned for comparing CA-biplots also apply here (cf. section 6.2.6 p. 71)

In figure 6.11 we present Gabriel's biplot for the BROAD matrix of the environmental data, constructed by interactive coding of year and environmental variables. Vector labels ending with an 'a' pertain to 1990, with a 'b' to 1991 and with a 'c' to 1992. Such an analysis can reveal the change in correlation structure between the variables (van der Heijden, 1987, pp. 189-192).

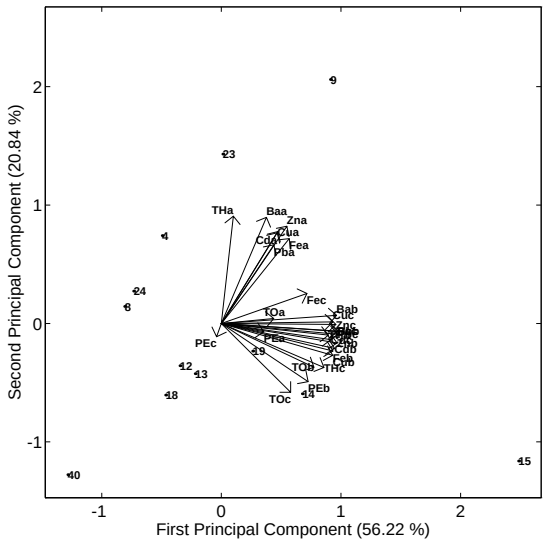

FIGURE 6.11: PCA BIPLOT OF THE BROAD MATRIX

The biplot in figure 6.11 explains 77% of the variance of the variables in the BROAD matrix. The first principal component is seen to be highly correlated with all variables from 1991 and 1992, whereas the second principal component is correlated with a set of variables from 1990 only, most notably THC and Barium. Stations 4, 23 and 9 are particularly high on this second component, whereas they are much lower on the first. These stations experienced a decrease in contamination. Geographically, this is the set of more northern stations in the field (cf. figure 2.3 page 8). Station 40 is singled out as being low on all contaminants during all three successive years. On the other hand, stations 14, 19 and notably 15 are high on all contaminants in 1991 and 1992, but low on the 1990 variables. These stations experienced an increase in contamination. Geographically these stations are on the southern side of the platform, relatively close. The analysis thus indicates that from 1990 to 1991 contamination diminished in the north and increased in the immediate south, what we could imagine to be a consequence of a change of current or a storm or whatever.

We also present the result of the analysis of the LONG matrix in figure 6.12. Here a PCA has been performed on the data matrix consisting of three annual chemical data matrices placed at the bottom of each other, so here we apply interactive coding of year and site. Changes in the means of the variables for the different years can influence the correlation structure between the variables. To avoid this, the stacking operation was done with the centred data matrices for each year, so that the effect of a changing mean has been eliminated (van der Heijden, 1987). The PCA biplot of the LONG matrix in figure 6.12 explains 83% of the variation of the LONG MATRIX, and shows that some stations experience large changes with respect to the environmental data. Notably, station 15 increases in THC, TOC and all heavy metals in 1991, followed by a decrease in 1992. Station 8 is characterized by a decrease on all chemicals in 1991, followed by a sharp increase on Pel in 1992. Station 24 also decreases on all variables in 1991, followed by a general increase in 1992. Station 40 shows a more stable chemical composition, with a slight decrease on all components in 1991.

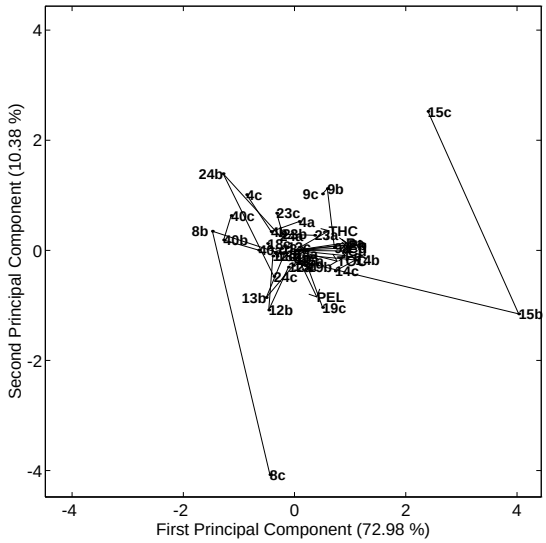

FIGURE 6.12: PCA BIPLOT OF THE LONG MATRIX

6.4 Conclusions

In this chapter we have seen some applications of CA to the biological data. The method has shown to be very useful for detecting some prominent features of the data set, notably the detection of highly abundant species at particular locations.

Another noteworthy point is that the stations of the inner ring (30,31,36,37) and station 40 are separated out in both the analysis of the biological data and in the analysis of the chemical data. This strongly suggests both data sets are

related. In particular, the analysis suggests that a species like *Capitella capitata* has a preference for contaminated conditions. Many biological scenarios can underlie such an observed preference; *Capitella capitata* could be abundant just because the environmental variables take optimum values for this organism, or because *Capitella capitata* feeds on other organisms that do well under these conditions, or because his predators or competitors are absent under these conditions. Without further specific knowledge about the biological relationships between the species and their population dynamics it is impossible to explain the detected high abundance of *Capitella capitata* in more detail.

To get an idea of the degree of association between the two data sets, we compute some correlations between principal axes from CA and the first principal component, and find some high values that draw attention. In particular, the first principal components of the three successive years have correlations 0.72, -0.75 and 0.85 with CA axes 2, 1 and 2 of the succesive years respectively.

The changes in the biological ordinations from year to year are difficult to detect when data from each year is analyzed separately. Though stacking the different matrices increases the dimensionality of the problem, it yields very interpretable output that depicts how abundance of some species has changed with time. Analogously, by stacking the chemical data matrices we get a picture of which stations experienced changes with respect to which variables.

Chapter 7

Optimal Directions for Supplementary Variables in Correspondence Analysis

7.1 Introduction

In the previous chapter we used CA as an ordination method for the species data, and obtained maps of sites and species that allow us to appreciate differences between the sites and between the species. With certain choices of scaling for the rows and the columns, these maps form biplots (Gabriel, 1971; Greenacre, 1993a), and species abundances can approximately be recovered by projecting the species points onto appropriately calibrated site vectors.

One of the basic purposes of ecological studies is to explain the variation in species composition of the sites in terms of environmental conditions (temperature, pH, pollutants, etc). This is called gradient analysis, and it can be performed in several ways.

If environmental information about the sites is not explicitly measured, but present as circumstantial knowledge, for instance if some sites are known to be polluted or extremely dry, this knowledge can be used in the interpretation of the CA biplot, where extreme sites then happen to separate out. For example, station 40 and the inner station ring sorted out in many of the CA's discussed in the previous chapter. The same stations also separated out in the PCA of the chemical data. This is probably not accidental; the chemical composition of the environment will affect the species composition. If the first CA axis separates contaminated stations from non-contaminated ones, then the first principal axis can be interpreted as "contamination", and if it correlates strongly with some explicitly measured environmental variable, the axis might be identified as being that variable. When environmental information is not collected, CA can be thought to uncover the hidden or latent environmental variables that do affect

the species composition (Ter Braak and Prentice, 1988). The identification of the principal axes with environmental variables greatly helps the interpretation of the display. Rather than detecting which species are high (or low) at which station, we can now infer the chemical constitution of the sites and the chemical preferences of the species. Seen this way, depicting environmental information in the CA biplot becomes a topic of keen interest. We are neither restricted to only interpret principal axes. Any direction in the biplot that strongly correlates with an environmental variable can be labelled with the name of that variable.

The process of relating principal axes after an analysis of the species data with environmental information is called indirect gradient analysis. In this chapter we consider a particular approach for indirect gradient analysis, where we first perform CA of the abundance data, and then try to represent environmental information in the CA map as well as possible. This is done by minimizing errors in the projections of the site coordinates onto the environmental vectors. Several particularities of this method will be pointed out. Computationally, the analysis can be performed by any software capable of doing correspondence analysis and regression. It can also be done with the CANOCO program (Ter Braak, 1988). We will use artificial data as an illustration of the method, and apply the method to the survey data described in previous chapters. Part of this chapter was presented at the Spanish Biometry Conference in 1997 (Graffelman, 1997).

7.2 Representing Supplementary Variables

We first perform correspondence analysis of the abundance matrix $\mathbf{N}$ (I species by J sites), by doing a singular value decomposition of the matrix containing deviations from independence (Gifi, 1981, section 8.3), or (5.3) in chapter 5:

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}'. \tag{7.1}$$

$\mathbf{P}$ is the correspondence matrix ($\mathbf{N}$ divided by its grand total), $\mathbf{r}$ and $\mathbf{c}$ are column vectors containing the row and column sums of $\mathbf{P}$ respectively, and $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices built from these vectors. The identification conditions of this singular value decomposition are $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$. The principal coordinates, following the notation of (Greenacre, 1984), are given by $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}$ and $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}$; the standard coordinates by $\mathbf{\Phi} = \mathbf{D}_r^{-1/2}\mathbf{U}$ and $\mathbf{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}$, both for rows and columns respectively. The joint plot of the rows of $\mathbf{F}$ and $\mathbf{\Gamma}$ is called the asymmetric map of the row profiles (Greenacre, 1993b), and this plot forms a biplot (Gabriel and Odoroff, 1990) since we can rewrite (7.1) as:

$$(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{D}_c^{-1} = \mathbf{F}\mathbf{\Gamma}'. \tag{7.2}$$

In Figure 7.1 we consider the asymmetric map of the row profiles of a small fictitious data set. Closed circles ($\bullet$) represent principal coordinates of the rows (species), and open circles ($\circ$) standard coordinates of the columns (sites). In this biplot we want to represent in the first instance, just one variable (column) of the matrix of supplementary environmental variables $\mathbf{Z}$, say $\mathbf{z}_j$. We do not assume any centring or standardization of $\mathbf{z}_j$. However, the algebraical results that follow can be somewhat simplified if we assume $\mathbf{z}_j$ to be standardized

by first centring by subtracting the weighted mean, and then dividing by the weighted standard deviation:

$$\mathbf{z}_j \leftarrow (\mathbf{z}_j - \mathbf{1}\mathbf{c}'\mathbf{z}_j)/\sqrt{(\mathbf{z}_j - \mathbf{1}\mathbf{c}'\mathbf{z}_j)'\mathbf{D}_c(\mathbf{z}_j - \mathbf{1}\mathbf{c}'\mathbf{z}_j)}. \tag{7.3}$$

Variable $\mathbf{z}_j$ will then have a weighted mean of zero and a weighted variance of 1. A hypothetical vector $\boldsymbol{\nu}$ representing this variable is drawn in the biplot (see fig 7.1). We assume that it is possible to calibrate this vector like the axis of a scatterplot. Projecting the column points onto this vector, one should then be able to recover the original supplementary data. Since there are usually many column points, one will hardly ever be able to recover supplementary quantitative measurements exactly. We can at best approximate our variable by a vector of estimates $\hat{\mathbf{z}}_j$, and so there will be errors $\mathbf{e}_j = \hat{\mathbf{z}}_j - \mathbf{z}_j$ in the projections.

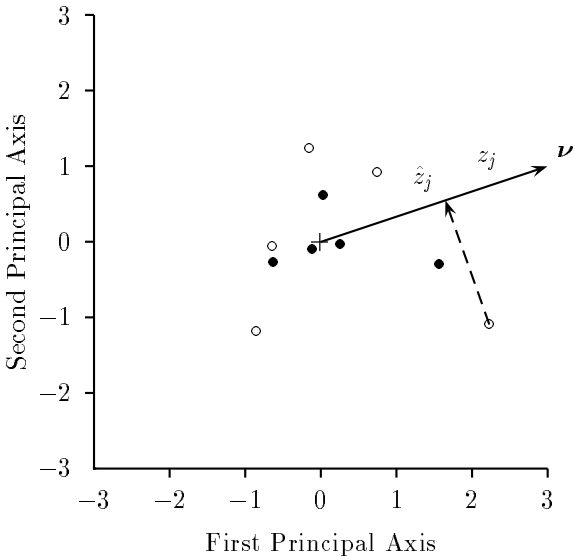

FIGURE 7.1: A CA BIPLOT WITH ADDED VARIABLE VECTOR

The problem then becomes to find an environmental vector $\boldsymbol{\nu}$ in such a way that the overall error is minimized, for example by minimizing the sum of the squared errors. Geometrically one can imagine this as rotating the vector $\boldsymbol{\nu}$ in figure 7.1, until a direction is found where the sum of the squared errors is minimal. So we minimize:

$$\mathbf{e}'\mathbf{e} = (\mathbf{z} - \hat{\mathbf{z}})'(\mathbf{z} - \hat{\mathbf{z}}) = (\mathbf{z} - \alpha\boldsymbol{\Gamma}\boldsymbol{\nu})'(\mathbf{z} - \alpha\boldsymbol{\Gamma}\boldsymbol{\nu}), \tag{7.4}$$

where $\hat{\mathbf{z}}$ are the environmental measurements as estimated in the biplot, $\boldsymbol{\Gamma}$ contains the standard site coordinates of the CA solution (7.1) and α is a scalar that serves as a normalization factor for vector $\boldsymbol{\nu}$. We need to minimize:

$$L(\alpha, \boldsymbol{\nu}) = \mathbf{z}'\mathbf{z} - 2\alpha\mathbf{z}'\boldsymbol{\Gamma}\boldsymbol{\nu} + \alpha^2\boldsymbol{\nu}'\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\nu}. \tag{7.5}$$

Setting first order derivatives equal to zero, the solution of the minimization problem is found to be:

$$\boldsymbol{\nu} = \frac{1}{\alpha}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{z}, \tag{7.6}$$

with $\alpha = \| (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{z} \|$. The solution vector is given by the normalized regression coefficients obtained in the regression of the environmental variable on the site coordinates. Note that, if one wants to explain the ordination in terms of the environmental variables, one would be tempted to regress $\boldsymbol{\Gamma}$ on $\mathbf{z}$. But if one searches for an optimal graphical display of both matrices $\mathbf{N}$ and $\mathbf{Z}$, then precisely the reverse is required. CA uses a particular way of weighting the data, where column j is weighted by mass c_j. It therefore seems logical to weight the projection errors in the same manner, and to minimize $\mathbf{e}'\mathbf{D}_c\mathbf{e}$ rather than $\mathbf{e}'\mathbf{e}$. Introducing this weighting, one obtains a simplified solution vector:

$$\boldsymbol{\nu} = \frac{1}{\alpha}(\boldsymbol{\Gamma}'\mathbf{D}_c\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{D}_c\mathbf{z} = \frac{1}{\alpha}\boldsymbol{\Gamma}'\mathbf{D}_c\mathbf{z}, \tag{7.7}$$

since $\boldsymbol{\Gamma}$ has normalization $\boldsymbol{\Gamma}'\mathbf{D}_c\boldsymbol{\Gamma} = \mathbf{I}$, and where now $\alpha = \| \boldsymbol{\Gamma}'\mathbf{D}_c\mathbf{z} \|$. The solution vector is now a vector of normalized regression coefficients of the weighted regression of $\mathbf{z}$ on $\boldsymbol{\Gamma}$. However, it can be shown that $\boldsymbol{\nu}$ is also a vector of weighted correlation coefficients. We have $\alpha = \sqrt{\mathbf{z}'\mathbf{D}_c\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c\mathbf{z}} = \sqrt{(\mathbf{z} - \mathbf{1}\mathbf{c}'\mathbf{z})'\mathbf{D}_c(\mathbf{z} - \mathbf{1}\mathbf{c}'\mathbf{z})}$, because $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c$ is an idempotent centring matrix (cf. (5.19)). So the scalar α is actually the square root of the weighted variance of $\mathbf{z}$, and any element ν_k of $\boldsymbol{\nu}$ can be written as:

$$\nu_k = \frac{\sum_{j=1}^{J} \gamma_{jk} c_j z_j}{\sqrt{\sum_{j=1}^{J} {z_j}^2 c_j}\sqrt{\sum_{j=1}^{J} {\gamma_{jk}}^2 c_j}}. \tag{7.8}$$

Equation (7.8) is the weighted correlation coefficient between $\mathbf{z}$ and dimension k of the standard column coordinates γ_{ik} of the CA solution; the weighted variance of the latter is 1 by construction. An environmental variable that has a perfect correlation (in the weighted sense) with a dimension in the CA solution, will coincide with the principal axis in the diagram. This is of great help in assigning meaning to the theoretical CA axes. Note that the solution vector $\boldsymbol{\nu}$ has as many elements as there are dimensions in the CA solution. In a two-dimensional biplot, we represent $\boldsymbol{\nu}$ by just plotting the first two elements of the vector. The representation of the correlations in this biplot is not approximate, but is exact in the sense that they can be read off the principal axes. The interpretation of $\boldsymbol{\nu}$ as a vector of weighted correlations is independent of any standardization or centring of $\mathbf{z}$. We note that the solution of (7.6) is identical to the solution of (7.7) if $\mathbf{z}$ is centred by subtracting weighted means. The projections of the site points can now be described as:

$$\alpha\boldsymbol{\Gamma}\boldsymbol{\nu} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c\mathbf{z} = \mathbf{z} - \mathbf{1}\mathbf{c}'\mathbf{z}. \tag{7.9}$$

So environmental scores are recovered as deviations from the weighted mean of the variable. Note that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c$ is an idempotent $\mathbf{D}_c$-symmetric centring matrix (Searle, 1982, chapter 3; Saporta, 1990, p. 480) , that applied to a vector, brings it into deviate form about its weighted mean. This is easily proved from equation (5.18) on page 59, considering $\boldsymbol{\Gamma}$ without the trivial column of ones so

that: $\mathbf{\Gamma\Gamma'D}_c\mathbf{z} = (\mathbf{D}_c^{-1} - \mathbf{11'})\mathbf{D}_c\mathbf{z} = (\mathbf{I} - \mathbf{1c'})\mathbf{z} = \mathbf{z} - \mathbf{1c'z}$.

The next natural step is to look if projections of the species points onto the vector $\boldsymbol{\nu}$ also have some interpretation:

$$\alpha\mathbf{F}\boldsymbol{\nu} = \mathbf{F\Gamma'D}_c\mathbf{z} = (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c'})\mathbf{z}. \tag{7.10}$$

This equation shows that we are recovering the (centred) weighted averages of the species with respect to the environmental variables, where the species abundance at each site is used as a weight. Equation (7.10) shows that the weighted averages of the species are perfectly recovered when we consider the full space, that is, all k dimensions of the CA solution, and all corresponding k dimensions of vector $\boldsymbol{\nu}$. In the case of a two-dimensional biplot, these weighted averages are not represented exactly, the weighted average of species i is approximated by $\alpha(f_{i1}v_1 + f_{i2}v_2)$, normalization factor α being one if $\mathbf{z}$ is standardized. This approximation is optimal in a weighted least squares sense (see below). Equation (7.10) is a nice result since the weighted average of a species is an estimate of the optimum of the species for that particular variable (Ter Braak, 1985), and so the map gives us an indication of species preferences as well. The same interpretation is usually made in canonical correspondence analysis (Ter Braak, 1986) and is also treated in chapter 9. The average row profile, $\mathbf{c}$, is represented by the origin of the CA biplot. This implies that, on the scale of the supplementary vector, the origin represents the weighted average $\mathbf{c'z}$ of the supplementary variable. This will be zero if the variable is centred by subtracting its weighted mean.

If we apply the transition equations of CA (cf. equation (5.10), p. 57) to the projected site points, we get the projected species points as a result:

$$(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c'})\alpha\mathbf{\Gamma}\boldsymbol{\nu} = \alpha(\mathbf{UDV'})\mathbf{\Gamma}\boldsymbol{\nu} = \alpha\mathbf{F}\boldsymbol{\nu}. \tag{7.11}$$

This illustrates that the transition equations hold for projections onto a vector in any possible direction of the biplot. It also suggests that we might minimize errors in weighted averages obtained when we project species points onto the environmental variable vector. The minimization problem, here using the weights $\mathbf{r}$, then becomes:

$$\mathbf{e'D}_r\mathbf{e} = (\alpha\mathbf{F}\boldsymbol{\nu} - (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c'})\mathbf{z})'\mathbf{D}_r(\alpha\mathbf{F}\boldsymbol{\nu} - (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c'})\mathbf{z}). \tag{7.12}$$

If one develops the Lagrangian for this expression, then one obtains the same solution as given by (7.7). This shows that the two minimization problems are equivalent.

Instead of minimizing the errors in the projections, one could also maximize the weighted correlation between the real environmental scores $\mathbf{z}$ and their estimates from the biplot $\hat{\mathbf{z}}$. This gives again the same solution vector described by (7.7).

One can apply the solution given by (7.7) repeatedly to different environmental variables, and in this way, represent several environmental variables in the CA biplot. Each vector is added independently from any other. We can rephrase

(7.7) in matrix terms to obtain the whole set $\mathbf{V}_e$ of environmental vectors in one step as:

$$\mathbf{V}_e = \mathbf{\Gamma}'\mathbf{D}_c\mathbf{Z}\mathbf{D}_\alpha{}^{-1}, \tag{7.13}$$

with $\mathbf{Z}$ a $J \times Q$ matrix of supplementary variables, and $\mathbf{D}_\alpha = \mathbf{diag}(\mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}_c\mathbf{Z})^{\frac{1}{2}}$. The columns of $\mathbf{V}_e$ represent the variables and are (normalized) regression coefficients in the simultaneous multiple regressions of $\mathbf{Z}$ on $\mathbf{\Gamma}$.

There is yet another approach that gives the solution described by (7.7) and (7.13). In fact, when we look for a vector $\boldsymbol{\nu}$ to represent a variable, we try to construct a conditional biplot for matrix $\mathbf{Z}$, where the representation of the rows of that matrix (the sites) are fixed, as they are given by CA. A biplot of $\mathbf{Z}$ is then given by the factorization:

$$\mathbf{Z} = \mathbf{\Gamma}\mathbf{H}' \Longrightarrow \mathbf{H}' = \mathbf{\Gamma}'\mathbf{D}_c\mathbf{Z}, \tag{7.14}$$

where $\mathbf{H}$ is a matrix whose rows represent the variables. If we normalize the rows of $\mathbf{H}$ we obtain (7.13).

7.3 Quality of Representation

The above results apply to the full space of the correspondence analysis solution, and we now consider how well the data are represented in a graphical display of low dimensionality (usually 2). We evaluate the representation of 3 matrices of interest, $\mathbf{N}, \mathbf{Z}$ and the matrix of weighted averages of the species, $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}$. For matrix $\mathbf{N}$, the quality of the display is given by the percentage of inertia captured by the low-dimensional map. For any of the environmental variables in the matrix $\mathbf{Z}$, we take as a quality measure the fraction of weighted variance of the variable represented in the map. Substitution of (7.7) in the weighted version of (7.4) gives us the following expression for the errors:

$$\mathbf{e}'\mathbf{D}_c\mathbf{e} = \mathbf{z}'\mathbf{D}_c\mathbf{z} - \hat{\mathbf{z}}'\mathbf{D}_c\hat{\mathbf{z}} = \mathbf{z}'\mathbf{D}_c\mathbf{z} - \mathbf{z}'\mathbf{D}_c\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}_c\mathbf{z}, \tag{7.15}$$

and we obtain as a measure of quality:

$$\frac{\hat{\mathbf{z}}'\mathbf{D}_c\hat{\mathbf{z}}}{\mathbf{z}'\mathbf{D}_c\mathbf{z}} = \sum_{k=1}^{K} r_k^2(\mathbf{z}, \mathbf{\Gamma}_k) = R^2, \tag{7.16}$$

where r_k denotes the weighted correlation coefficient between $\mathbf{z}$ and dimension k of the standard column coordinates. Equation (7.16) shows that the quality of representation of a variable depends on the number of dimensions of the CA solution considered and is given by the coefficient of determination (R^2). It rephrases the well-known result that regression on orthogonal variables gives an R^2 that is the sum of squared correlations ($R^2 = \sum_{k=1}^{K} r_k^2$), but here in a weighted sense. It is evident from (7.16) that the sum of squared weighted correlations cannot exceed 1. This means that if a variable is highly correlated with a particular dimension in the CA map, it must be nearly uncorrelated with the other dimensions, and this is just another formulation for the fact that the

principal axis in CA are uncorrelated.

The quality of representation of a particular vector in two dimensions is also indicated by its length. The length of the vector in two dimensions ($\boldsymbol{\nu}_{(2)}$) is:

$$\| \boldsymbol{\nu}_{(2)} \| = \sqrt{\boldsymbol{\nu}_{(2)}'\boldsymbol{\nu}_{(2)}} = \sqrt{\frac{\mathbf{z}'\mathbf{D}_c\boldsymbol{\Gamma}_{(2)}\boldsymbol{\Gamma}_{(2)}'\mathbf{D}_c\mathbf{z}}{\mathbf{z}'\mathbf{D}_c\mathbf{z}}} = \sqrt{\frac{\hat{\mathbf{z}}'\mathbf{D}_c\hat{\mathbf{z}}}{\mathbf{z}'\mathbf{D}_c\mathbf{z}}} = \sqrt{R^2}. \qquad (7.17)$$

Note that we use $\boldsymbol{\Gamma}_{(2)}$ to indicate the first two columns of matrix $\boldsymbol{\Gamma}$ and $\boldsymbol{\nu}_{(2)}$ for the first two elements of vector $\boldsymbol{\nu}$. Equation (7.17) shows that the length is just the square root of the amount of weighted variance explained by the biplot. One can therefore, as in principal component analysis, draw a unit circle in the CA biplot. Vectors with their head on the unit circle have a perfect representation in the biplot, since (7.15) and (7.16) show that if $R^2 = 1$ the errors vanish and projection of the site points on the variable vectors recovers the data matrix $\mathbf{Z}$ exactly. This will only happen when the environmental variables are exact linear combinations of the standard CA site scores. For instance, if $\mathbf{N}$ has three columns, the CA solution will have two dimensions, and the norm of $\boldsymbol{\nu}_{(2)}$ will always be one, and R^2 will also be one. In this case, one can recover supplementary data from the biplot without error, regardless of the number of variables.

For multiple variables, we can take the mean of the coefficients of determination as a measure of the overall quality of representation of the supplementary variables, that is $(1/Q)\sum_{q=1}^{Q} R_q^2$.

Next, how well are the weighted averages of the species with respect to the supplementary variables represented? As a criterion for the quality of representation we consider the amount of the weighted variance in the weighted averages explained by a low dimensional map. Using first just one supplementary variable $\mathbf{z}$, we use (7.12) to develop this criterion:

$$\begin{aligned}
\mathbf{e}'\mathbf{D}_r\mathbf{e} &= ((\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{z} - \alpha\mathbf{F}_{(2)}\boldsymbol{\nu}_{(2)})'\mathbf{D}_r((\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{z} - \alpha\mathbf{F}_{(2)}\boldsymbol{\nu}_{(2)}) \\
&= (\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{z} - \alpha\mathbf{F}_{(2)}\boldsymbol{\nu}_{(2)})'\mathbf{D}_r(\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{z} - \alpha\mathbf{F}_{(2)}\boldsymbol{\nu}_{(2)}) \\
&= \alpha^2\boldsymbol{\nu}'\mathbf{D}^2\boldsymbol{\nu} - \alpha^2\boldsymbol{\nu}_{(2)}'\mathbf{D}_{(2)}^2\boldsymbol{\nu}_{(2)}. \qquad (7.18)
\end{aligned}$$

Using the fact that $\boldsymbol{\nu}$ is a vector of correlation coefficients, a measure for the quality of the m-dimensional representation of the weighted averages is:

$$\frac{\boldsymbol{\nu}_{(m)}'\mathbf{D}_{(m)}^2\boldsymbol{\nu}_{(m)}}{\boldsymbol{\nu}'\mathbf{D}^2\boldsymbol{\nu}} = \frac{\sum_{k=1}^{m} r_k^2 d_k^2}{\sum_{k=1}^{K} r_k^2 d_k^2}. \qquad (7.19)$$

This shows that the quality of a two-dimensional representation of the weighted averages depends on two factors: the weighted correlations of the supplementary variable with the principal axes and the amount of inertia explained by the two-dimensional display. Note that if the CA solution has only 2 dimensions, the quality of representation of the weighted averages is 1, and there will be no errors when we try to recover the weighted averages of the species with respect to the supplementary variables from the map. If the CA solution has more than

two dimensions, then the higher the percentage of inertia explained by the two-dimensional map, and the higher the correlations between the supplementary variable and the first two principal axes, the better the representation of the weighted averages in two dimensions.

For multiple supplementary variables the overall criterion of representation of the weighted averages becomes:

$$\frac{\mathrm{tr}(\mathbf{V}_{e(m)}{}'\mathbf{D}_{(m)}^2\mathbf{V}_{e(m)})}{\mathrm{tr}(\mathbf{V}_e{}'\mathbf{D}^2\mathbf{V}_e)} = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{m} r_{qk}{}^2 d_k^2}{\sum_{q=1}^{Q}\sum_{k=1}^{K} r_{qk}{}^2 d_k^2}, \tag{7.20}$$

where r_{qk} denotes the weighted correlation between variable q and principal axis k.

One might wonder about the angles between different supplementary variables. Those angles do turn out to be approximations to the weighted correlations between the supplementary variables:

$$\mathbf{V}_e{}'\mathbf{V}_e = \mathbf{D}_\alpha^{-1}\mathbf{Z}'\mathbf{D}_c\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c\mathbf{Z}\mathbf{D}_\alpha^{-1} = \mathbf{D}_\alpha^{-1}\mathbf{Z}'\mathbf{D}_c\mathbf{Z}.\mathbf{D}_\alpha^{-1} \tag{7.21}$$

Considering the full space of the solution, term $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c$ is the centring matrix that transforms $\mathbf{Z}$ into a matrix of deviations from the weighted means (cf. (7.9) p. 84). Then (7.21) implies that the scalar product between two supplementary variable vectors is their weighted correlation:

$$\boldsymbol{\nu}_i{}'\boldsymbol{\nu}_j = \frac{\mathbf{z}_i{}'\mathbf{D}_c\mathbf{z}_j}{\sqrt{\mathbf{z}_i{}'\mathbf{D}_c\mathbf{z}_i}\sqrt{\mathbf{z}_j{}'\mathbf{D}_c\mathbf{z}_j}} = r(\mathbf{z}_i, \mathbf{z}_j). \tag{7.22}$$

When only a few dimensions of the solution are considered, $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D}_c$ is no longer a centring matrix, and the cosine of the angle between two variable vectors does not represent their weighted correlation exactly, but only approximately, and we ignore whether this approximation is optimal in some sense.

7.4 Supplementary Vectors in the CA Symmetric Map

Though the symmetric map in CA is not an interesting biplot (cf. section 5.4 p. 58), this type of scaling is often used in practice. We therefore also consider the representation of supplementary variables in the symmetric map in some more detail. The minimization problem previously described is basically the same as in equation (7.5), but now $\boldsymbol{\Gamma}$ is replaced by $\mathbf{G}$. Doing the same algebra, the solution is then given by:

$$\boldsymbol{\nu} = \frac{1}{\alpha}(\mathbf{G}'\mathbf{D}_c\mathbf{G})^{-1}\mathbf{G}'\mathbf{D}_c\mathbf{z} = \frac{1}{\alpha}\mathbf{D}^{-2}\mathbf{G}'\mathbf{D}_c\mathbf{z}, \tag{7.23}$$

with $\alpha = \| (\mathbf{G}'\mathbf{D}_c\mathbf{G})^{-1}\mathbf{G}'\mathbf{D}_c\mathbf{z} \|$. $\boldsymbol{\nu}$ is again a vector of normalized regression coefficients. It is tempting to think that $\boldsymbol{\nu}$ then will again be a vector of weighted correlation coefficients, this time between $\mathbf{G}$ and $\mathbf{z}$. This is however, not the case. First we notice that α is again the square root of the weighted variance of

$\mathbf{z}$ because $\sqrt{\mathbf{z}'\mathbf{D}_c\mathbf{G}(\mathbf{G}'\mathbf{D}_c\mathbf{G})^{-2}\mathbf{G}'\mathbf{D}_c\mathbf{z}} = \sqrt{\mathbf{z}'\mathbf{D}_c\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}_c\mathbf{z}} = \sqrt{\mathbf{z}'\mathbf{D}_c\mathbf{z}}$, and thus we can write the k^{th} element of $\boldsymbol{\nu}$ as:

$$\nu_k = \frac{\sum_{j=1}^{J} g_{jk} c_j z_j}{\sqrt{\sum_{j=1}^{J} c_j {z_j}^2 d_j^2}} = \frac{\sum_{j=1}^{J} g_{jk} c_j z_j}{\sqrt{\sum_{j=1}^{J} c_j {z_j}^2}\sqrt{d_j^2}} \cdot \frac{1}{\sqrt{d_j^2}}, \tag{7.24}$$

where we use the property that the principal inertias (d_j^2) are the weighted variances of the principal coordinates in $\mathbf{G}$. Equation (7.24) shows that $\boldsymbol{\nu}$ is a vector of weighted correlation coefficients, but that each correlation is divided by the square root of the principal inertia. Thus, it would be wrong to plot weighted correlations in the symmetric map in order to obtain the optimal direction. Calculating weighted correlations is however, a sensible computational step to arrive at the solution. Result (7.24) can also be understood in a more intuitive way. Principal coordinates are a rescaling of the standard coordinates. If weighted correlations provide the optimal direction when standard coordinates are used, the same rescaling should be applied to the weighted correlations in order to obtain the optimal direction when using principal coordinates. This is also clear if we substitute $\mathbf{G} = \mathbf{\Gamma}\mathbf{D}$ in (7.23):

$$\boldsymbol{\nu} = \frac{1}{\alpha}\mathbf{D}^{-2}\mathbf{G}'\mathbf{D}_c\mathbf{z} = \frac{1}{\alpha}\mathbf{D}^{-2}\mathbf{D}\mathbf{\Gamma}'\mathbf{D}_c\mathbf{z} = \frac{1}{\alpha}\mathbf{D}^{-1}\mathbf{\Gamma}'\mathbf{D}_c\mathbf{z}. \tag{7.25}$$

How wrong is it to plot weighted correlations? Because inertias are positive numbers, there will be no sign reversals, and plotting weighted correlations does yield a vector that always lies in the same quadrant as the correct direction. When the first and the second inertia are approximately equal, the rescaling amounts to multiplying the first two elements of $\boldsymbol{\nu}$ by a constant. As a consequence, considering 2 dimensions, the length of the vector is mistaken by a constant factor, though the direction found is close to the correct one. However, when the first and second inertia are very different, that is, the first principal inertia captures a large part of the dispersion, and the second one a small part, then plotting the weighted correlations gives a very different direction.

7.5 A Different View on Supplementary Points in CA

When we consider supplementary cases instead of supplementary continuous variables, then the position of such supplementary points in the CA map is usually calculated using transition formulae (5.10). However, we could try to find the position of a supplementary point by using the same methodology exposed in this chapter. We could search for a supplementary vector, representing the supplementary point (a case), in such a way that its projections onto the site vectors are as best as possible. As illustrated in section 7.7 below, site vectors can be calibrated in such a way that the profiles of the species can be approximately recovered when projecting species points (rows) onto the site vectors (columns). For a supplementary point, we can apply the same argument: we try to find a vector $\mathbf{x}$ in the biplot that has the property that its projections onto all site vectors, the rows of matrix $\mathbf{\Gamma}$, have minimal error. If there is a set of supplementary points, then they need to have the same centre of gravity as

the profiles that were used in the CA. That is to say, we first need to centre any supplementary point $\mathbf{p}$ onto the average row profile, $\mathbf{c}$, so that $\mathbf{p} \leftarrow \mathbf{p} - \mathbf{c}$. Next, when we project the species points onto the site vectors, and want to recover data in profile form, then we need to use the rescaled site vectors $\mathbf{D}_c\mathbf{\Gamma}$, rather than $\mathbf{\Gamma}$. This because the profiles can be written as $\mathbf{D}_r^{-1}\mathbf{P} = \mathbf{F}(\mathbf{D}_c\mathbf{\Gamma})'$ (see also section 5.4). In order to work out the projections, we need the norms of the row vectors of $\mathbf{D}_c\mathbf{\Gamma}$, which are given by $\mathbf{D}_c\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}_c = \mathbf{D}_c$, where the latter equality only holds in the full space, and including the trivial column of ones. Row vectors with norm one are thus obtained as $\mathbf{D}_c^{-1/2}\mathbf{D}_c\mathbf{\Gamma} = \mathbf{D}_c^{1/2}\mathbf{\Gamma}$. The estimated profiles for a supplementary point, expressed as a column vector, are thus given by:

$$\mathbf{D}_c^{1/2}\mathbf{\Gamma}\mathbf{x}. \tag{7.26}$$

Letting column vector $\mathbf{p}$ be the true supplementary profile, we try to minimize:

$$(\mathbf{D}_c^{1/2}\mathbf{\Gamma}\mathbf{x} - \mathbf{p})'(\mathbf{D}_c^{1/2}\mathbf{\Gamma}\mathbf{x} - \mathbf{p}). \tag{7.27}$$

Doing similar algebra as before, without particular restrictions for the norm of $\mathbf{x}$, we find that the solution is given by

$$\frac{\mathbf{x}}{\parallel \mathbf{x} \parallel} = \frac{1}{\sqrt{\mathbf{p}'\mathbf{D}_c^{1/2}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}_c^{1/2}\mathbf{p}}}\mathbf{\Gamma}'\mathbf{D}_c^{1/2}\mathbf{p} = \frac{1}{\parallel \mathbf{p} \parallel}\mathbf{\Gamma}'\mathbf{D}_c^{1/2}\mathbf{p}. \tag{7.28}$$

We note that, when the elements of the supplementary profile $\mathbf{p}$ are first divided by the square root of their respective column masses, then (7.28) gives the same solution as the transition formulae, up to a constant factor. When we choose the appropriate norm for the solution vector, the supplementary point found will coincide exactly with the one obtained by using the transition formulae. The squared norms of the species point vectors in the CA solution are given by the diagonal elements of $\mathbf{F}\mathbf{F}' = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}$, so that the solution vector $\mathbf{x}$ has to be rescaled to make its norm $\mathbf{p}'\mathbf{D}_c^{-1}\mathbf{p}$.

7.6 Relationships with other Methods

In this subsection we comment on the relationships of our approach with other multivariate methods. The other methods we consider are indirect gradient analysis as proposed by Dargie (1984), canonical correspondence analysis (CCA, (Ter Braak, 1986)) and weighted principal component analysis.

7.6.1 Indirect Gradient Analysis

Dargie (1984) described, in the context of multidimensional scaling, a procedure for finding a direction of maximal correlation between habitat variables and ordination axes as:

$$\theta = \arctan\left(\frac{b_2}{b_1}\right), \tag{7.29}$$

where θ is the angle with respect to the ordination axis, and b_1 and b_2 are the regression coefficients of $\mathbf{z}$ on the ordination axes. The length of this direction

is rescaled to reflect R^2. If the regression is weighted, and CA is used as the ordination method, Dargie's proposal will give the same solution as (7.7).

7.6.2 Canonical Correspondence Analysis

Canonical correspondence analysis (Ter Braak, 1986) is a technique that also provides a biplot of species and sites and environmental vectors, and is described in detail in chapter 9. CCA can be seen as a CA, where the standard site coordinates have been restricted to be linear combinations of environmental variables. In the particular case that the number of variables is as large or larger than the number of sites minus one ($Q \geq J-1$), the CCA solution is equal to the CA solution, but will still give us a representation in the biplot for the environmental variables. The environmental vectors obtained this way have the same direction as the ones obtained by (7.7) and (7.13), but a different length. To go short, the CCA solution can be obtained by the singular value decomposition (cf. (Jongman et al., 1987, section 5.9), chapter):

$$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}')\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} = \mathbf{A}\mathbf{T}\mathbf{W}', \tag{7.30}$$

with $\mathbf{A}'\mathbf{A} = \mathbf{I}$ and $\mathbf{W}'\mathbf{W} = \mathbf{I}$. Species coordinates and variable coordinates are given by $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{A}\mathbf{T}$ and $\mathbf{\Omega} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W}$ respectively. Since we consider conditions where CA and CCA solution coincide (inertia decomposition, species and site coordinates being the same in both types of analysis), premultiplying (7.30) by $\mathbf{D}^{-1}\mathbf{U}'$ and simplifying using (7.1) leads to:

$$\mathbf{\Omega}' = \mathbf{\Gamma}'\mathbf{D}_c\mathbf{Z}. \tag{7.31}$$

This is the same solution as described before (7.13), if the rows of $\mathbf{\Omega}$ are normalized.

7.6.3 Weighted Principal Component Analysis

The factorization of $\mathbf{Z}$ in (7.14), and the fact that the cosine of any angle between two column vectors of $\mathbf{V}_e$ approximates a (weighted) correlation is reminiscent of principal component analysis and suggests that the analysis is close to a weighted principal component analysis of $\mathbf{Z}$.

In a weighted principal component analysis (WPCA) of $\mathbf{Z}$ one extracts eigenvalues and eigenvectors of the weighted correlation matrix of the variables, or one can use a corresponding singular value decomposition:

$$\mathbf{D}_c^{\frac{1}{2}}\mathbf{Z} = \bar{\mathbf{U}}\bar{\mathbf{D}}\bar{\mathbf{V}}'. \tag{7.32}$$

Standardized principal components, uncorrelated in the weighted sense, are given by $\mathbf{D}_c^{\frac{1}{2}}\bar{\mathbf{U}}$, and satisfy similar identification conditions as the standard column coordinates in CA: $(\mathbf{D}_c^{-\frac{1}{2}}\bar{\mathbf{U}})'\mathbf{D}_c(\mathbf{D}_c^{-\frac{1}{2}}\bar{\mathbf{U}}) = \mathbf{I}_q$ and $\mathbf{\Gamma}'\mathbf{D}_c\mathbf{\Gamma} = \mathbf{I}_{J-1}$ respectively. This implies that, if $Q = J-1$, and if we consider the full space of the solution, then a particular column vector in CA will lie with its head on the

same sphere as the corresponding site vector (a case) in WPCA, as well as that the angles between site vectors are the same in the CA and the WPCA. However, the first principal component captures the direction of maximal variance of the site scores, a direction that does not necessarily coincide with the direction of maximum spread of row profiles as captured by the first principal axis in CA. Principal axes of both biplots will therefore usually not coincide. In practice it means that if Gabriel's biplot is put on top of the CA biplot, and rotated to make the site points coincide, then the variable vectors of the WPCA will coincide with the ones obtained by our regression approach. Because of the sign indeterminacy of eigenvectors in both analyses, one has to choose a particular reflection of the WPCA (or CA) output before this can be done. Because of this and the fact that the above only applies to *full space* solutions with a number of variables that is one less then the number of sites ($Q = J - 1$), this method of calculating the solution is of very little practical use; it will work for data sets that have a perfect representation in 2 dimensional space. The equivalence under these particular circumstances just described can easily be verified by applying a procrustes rotation to the joint set of coordinates of variables and sites in the two types of analysis. The procrustes rotation then gives a perfect fit with RSS=0 and the scaling factor equals 1.

7.7 An Example with Artificial data

In this section we present some examples. First, we consider a small artificial data set illustrating a perfect fit. The data are shown in table 7.1. The first three rows list the raw data (abundances and environmental variables Z1 and Z2), the second three rows the species profiles, and the last two rows represent the weighted averages of the species with respect to Z1 and Z2, as well as the weighted averages of the variables. Note that chemical gradients are present in the data, as the concentrations of Z1 and Z2 increase over the 3 sites.

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Z1	Z2
Site A	10	5	15	30	10	4.0	1.0
Site B	20	5	15	20	20	6.0	4.0
Site C	30	10	15	10	10	1.0	6.0
Site A	0.1667	0.2500	0.3333	0.5000	0.2500	-	-
Site B	0.3333	0.2500	0.3333	0.3333	0.5000	-	-
Site C	0.5000	0.5000	0.3333	0.1667	0.2500	-	-
wa Z1	3.1667	3.0000	3.6667	4.1667	4.2500	3.71	-
wa Z2	4.5000	4.2500	3.6667	2.8333	3.7500	-	3.73

TABLE 7.1: ARTIFICIAL ABUNDANCE AND ENVIRONMENTAL DATA

Since the abundance matrix is a 5 by 3 table, the CA solution has two dimensions, and a two-dimensional biplot will represent 100 percent of the inertia of this table, and species profiles can be perfectly recovered by projecting the

species points onto the column vectors. In particular, the first dimension captures 82.2% of the total inertia, and the second dimension 17.8%. The upper left graph in figure 7.2 shows ordinary CA output, the asymmetric map of the species profiles, with vectors pointing to the site vertices.

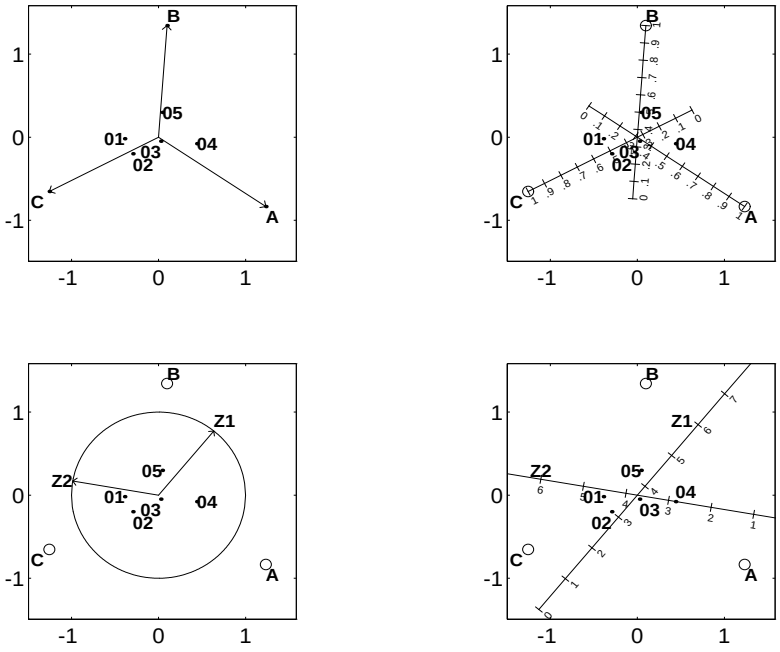

FIGURE 7.2: CA BIPLOTS WITH SUPPLEMENTARY VECTORS

The upper right graph shows the same CA biplot, but now the site vectors have been automatically calibrated, and one can recover the profiles of the species perfectly (e.g. species 4 projected onto vector A gives us exactly 0.5, cf. table 7.1). The lower left graph shows the CA output with the two added supplementary vectors Z1 and Z2, with their tips on the unit circle, indicating that the biplot represents all their weighted variance, and that the presentation of environmental data is therefore also perfect. Note how the three sites line up along the vectors Z1 and Z2: Z1 increases over the sites in order C,A and B and Z2 over the sites in order A,B and C, which is in accordance with the raw data values. The sites also line up along the first principal axis in order A,B and C, and we could say that the CA has "picked up" the Z2 gradient in our data set. The lower right graph shows again the same biplot, but now the variable vectors have also been calibrated so that one unit on the variable vector is one unit in the original scale of the variable. Projecting the site points onto these vectors will now recover the raw environmental data exactly, and projecting the species points will recover their weighted averages exactly (e.g. Site B projects onto value 6 with respect to Z1 and 4 with respect to Z2, projecting species 5 yields a weighted average of about 4.2 with respect to variable

Z1, and a weighted average around 3.8 with respect to variable Z2, cf. table 7.1)

An increasing number of species could have been added to the data table without any loss in the quality of the representation. However, if more sites are included in the analysis, then there will be extra dimensions in the solution and the projections just explained can only be approximate. The weighted correlation between the variables Z1 and Z2 is -0.4947, and is correctly reflected by an angle of 120 degrees between the two vectors.

7.8 Real Data Applications

We apply the methodology described in sections 7.2 and 7.3 to the survey data from 1992, with 166 species, 12 sites and 10 environmental variables: PEL, THC, TOC, Ba, heavy metals Cd, Cu, Fe, Pb, and Zn. We also include the distance from the platform as a variable. The left graph in figure 7.3 shows the two dimensional CA solution. The first dimension captures 28.4% of the inertia and the second dimension captures 23.5%, giving an overall quality of the display of 51.9%. The interpretation of the CA map, irrespective of the chemical data, has been commented on previously (cf. section 6.2.5): the horizontal axis separates the non-polluted reference station (40) from the rest. The second dimension captures a difference between stations 40, 24 and 15. These 3 stations are the best-represented ones in the map. Many species in the map are relatively ill-represented. A few ones with a high quality (≥ 0.9) in the display have been labelled. These are *Amphiura filiformis*, *Myriochele oculata* and *Chaetozone setosa*. *Amphiura filiformis* and *Myriochele oculata* are the main contributors to the first principal axis, whereas the second axis is mainly determined by *Chaetozone setosa* and *Myriochele oculata*.
From the map one infers that *Amphiura filiformis* is relatively more abundant at the reference station, while *Chaetozone setosa* is high on station 15 and *Myriochele oculata* on 24. These are the stations closest to the platform. This suggests *Chaetozone setosa* and *Myriochele oculata* could be considered indicators of pollution.

The right graph of figure 7.3 shows the same data, after zooming in a bit, and with added environmental vectors. This figure shows that nearly all variables have a considerable amount of their weighted variance accounted for in 2 dimensions. Only Pelite is ill-represented. The display of these variables greatly helps the interpretation of the theoretical CA axes. It is clear that all heavy metals, TOC and THC are associated with the second CA axis, whereas the horizontal principal axis has a relatively high negative correlation with Distance. The distance vector reflects the fact that reference station 40 is far away, 12 and 8 are at intermediate distance, and the other stations are close to the platform. We could globally resume the CA diagram by saying that the first axis is distance, and the second axis pollution. It is clear that, apart from Pelite, all the other environmental variables are correlated. The obtuse angle between distance and most chemical variables shows, as expected, distance to be negatively correlated with these variables. The biplot shows a whole bunch of vectors pointing up along the vertical axis. The biplot explains 59.16% of the weighted variance of the supplementary environmental variables (distance excluded). Individual

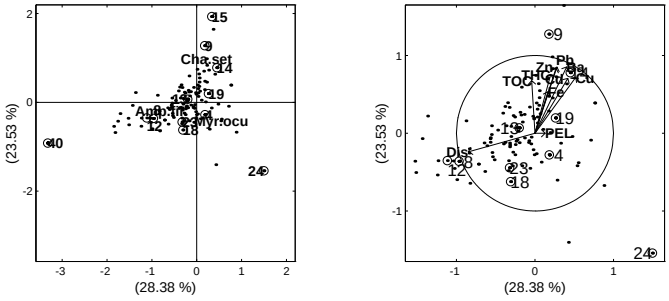

FIGURE 7.3: CA BIPLOT AND CA BIPLOT WITH SUPPLEMENTARY VECTORS

qualities of representation for the variables are: PEL 1.68%, THC 61.26%, TOC 45.91%, Ba 88.99%, Cd 67.39%, Cu 78.61%, Fe 31.24%, Pb 77.04%, Zn 80.31%, and Distance 77.73%. Regarding the weighted averages, the biplot explains 87.1% of their variance (distance excluded). For reasons of space the right graph of figure 7.3 does not show the species labels, though by comparison with the left graph it is clear that *Chaetozone setosa* projects high on the pollutants, whereas *Myriochele oculata* and *Amphiura filiformis* project low. This confirms that *Chaetozone setosa* indicates pollution, though our previous interpretation of *Myriochele oculata* seems wrong: it is low on the pollution vectors. Most clearly, *Amphiura filiformis* is low on all pollutants and high on station 40, suggesting this species dislikes contamination.

	Pel	THC	TOC	Ba	Cd	Cu	Fe	Pb	Zn	Dis
PEL	1.00	0.25	-0.06	0.43	0.51	0.59	0.29	0.30	0.31	-0.96
THC	0.13	1.00	0.95	0.98	0.96	0.93	1.00	1.00	1.00	-0.51
TOC	0.16	0.65	1.00	0.88	0.83	0.77	0.93	0.93	0.93	-0.21
Ba	0.08	0.88	0.72	1.00	1.00	0.98	0.99	0.99	0.99	-0.66
Cd	0.00	0.84	0.54	0.90	1.00	1.00	0.97	0.97	0.98	-0.72
Cu	0.03	0.83	0.49	0.91	0.97	1.00	0.95	0.95	0.95	-0.78
Fe	-0.11	0.33	0.44	0.61	0.46	0.48	1.00	1.00	1.00	-0.54
Pb	-0.07	0.82	0.66	0.94	0.91	0.90	0.71	1.00	1.00	-0.55
Zn	-0.05	0.81	0.56	0.90	0.94	0.94	0.57	0.95	1.00	-0.56
Dis	-0.19	-0.24	-0.01	-0.50	-0.41	-0.49	-0.40	-0.36	-0.36	1.00

TABLE 7.2: REAL VERSUS ESTIMATED CORRELATIONS

Table 7.2 shows the weighted correlations between the variables. Below diagonal

elements are the correlations based on the data, above diagonal elements are the correlations estimated from the biplot. Note that there are some sign reversals for PEL. The estimated correlations are nearly always (much) larger than the correct correlations. The biplot in figure 7.3 therefore considerably exaggerates the correlations between the variables. This is likely to happen in any biplot, notably when there are several uncorrelated variables. If we imagine three variables that are uncorrelated, then it is already impossible to depict there correlations correctly in two dimensions, since it is impossible to draw three vectors all at right angles with each other in a two dimensional plane. Thus, it is inevitable that a 2-D biplot with three uncorrelated variables suggests correlations being too high.

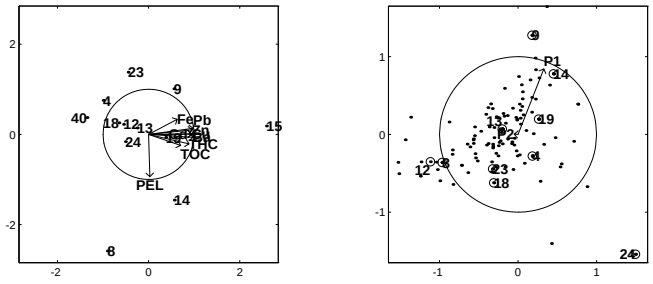

FIGURE 7.4: WPCA BIPLOT AND CA BIPLOT WITH ADDED PRINCIPAL COMPONENTS

We can try to reduce the amount of variables by a weighted principal component analysis (WPCA), which would give us a few orthogonal directions to add to the CA biplot. The biplot of the principal component analysis and the CA biplot with added principal components are displayed in figure 7.4. The variable distance has been excluded from the WPCA. Gabriel's biplot of the WPCA also shows the high intercorrelations between TOC, THC and all heavy metals. The first principal component explains 69.7% of the total weighted variance, and can be described as pollution due to these variables. The second principal component can be identified as Pelite, and accounts for 12.6% of the weighted variance. Station 40 is again singled out as a non-polluted station, being low on all measured variables, whereas station 15 seems to be a very polluted station. Station 8 and 14 are relatively high on Pelite. The right graph of figure 7.4 shows again the CA solution, but now two vectors representing the first and second principal component have been added. The second axis of the graph is

clearly associated with the first principal component (P1). The second principal component is ill-represented in the graph, as was Pelite in graph 7.3.

7.9 Conclusions

In this chapter we have treated theory and application of representing supplementary continuous variables in a CA biplot. Such representations turn out to be very useful in interpreting the CA solution. We have also shown that it is often possible to depict supplementary variables with good quality, notably when the CA biplot explains a large percentage of inertia and the correlations between the principal axes and the supplementary variables are high.

The analysis proposed remains in the realm of *indirect* gradient analysis, since first the species data are optimally represented and then environmental variables are added. Chapter 9 is dedicated to the theory of canonical correspondence analysis (CCA), which is a form of direct gradient analysis since it uses the species data and chemical variables simultaneously, and which is related to the proposed indirect analysis in this chapter (cf. section 7.6.2).

Chapter 8

Optimal Directions for Supplementary Variables in Principal Component Analysis

8.1 Introduction

In the previous chapter we have seen how we can represent environmental variables in a biplot in an optimal manner, in the context of correspondence analysis. The abundance matrix considered could also be analyzed by principal component analysis (PCA), and so we are also tempted to search for the optimal representation of external variables in a biplot obtained by PCA. This is in fact a topic of a more general interest beyond the particular ecological context considered here, as it concerns the representation of supplementary variables in a PCA biplot. PCA is performed on a particular set of variables, and it can be of interest to depict another variable, deliberately not included in the PCA, in a PCA biplot posterior to the analysis. A clear example of this is also given by the data at hand. In chapter 6 we considered the PCA of the chemical data. The variable distance is evidently not included in such an analysis, since it is a different type of variable. However, its representation in the PCA biplot can be very informative, as it might reveal that stations high on a particular chemical are close or far away from the platform. We therefore also consider the representation of supplementary variables in a PCA biplot in more detail. Expressions for calculating supplementary variable vectors in PCA are derived below, and we illustrate the results with an example from the Ekofisk oil field.

8.2 PCA in a Nutshell

We consider the PCA of a standardized data matrix $\mathbf{X}$. Principal components can be obtained by calculating eigenvectors of the covariance matrix. When data are standardized, as we assume here, the covariance matrix equals the correlation matrix. Phrased in terms of a singular value decomposition, we do a low rank approximation to the standardized data:

$$\mathbf{X} = \tilde{\mathbf{U}}\mathbf{T}\tilde{\mathbf{V}}', \tag{8.1}$$

with $\tilde{\mathbf{U}}'\tilde{\mathbf{U}} = \mathbf{I}$ and $\tilde{\mathbf{V}}'\tilde{\mathbf{V}} = \mathbf{I}$. We use a tilde ($\tilde{}$) to avoid possible confusion with previously used matrices in CA or CCA ($\mathbf{U}$, $\mathbf{F}$, etc.). Standardized principal components ($\tilde{\mathbf{F}}$) and coordinates for the variable vectors ($\tilde{\mathbf{H}}$) can then be obtained as:

$$\tilde{\mathbf{F}} = \sqrt{n}\tilde{\mathbf{U}} \qquad \tilde{\mathbf{H}} = (1/\sqrt{n})\tilde{\mathbf{V}}\mathbf{T}, \tag{8.2}$$

where n is the sample size. If the variables are standardized then $\tilde{\mathbf{H}}$ contains the correlations of the standardized principal components with the variables because: $(1/n)\mathbf{X}'\tilde{\mathbf{F}} = (1/n)\tilde{\mathbf{V}}\mathbf{T}\tilde{\mathbf{U}}'\tilde{\mathbf{U}}\sqrt{n} = (1/\sqrt{n})\tilde{\mathbf{V}}\mathbf{T}$. Principal components are linear combinations of the original variables with maximal variance, and can thus be obtained by a linear transformation of the data matrix. Using (8.1) we find:

$$\tilde{\mathbf{F}} = \sqrt{n}\tilde{\mathbf{U}} = \sqrt{n}\mathbf{X}\tilde{\mathbf{V}}\mathbf{T}^{-1} = \mathbf{X}\mathbf{C}, \tag{8.3}$$

where matrix $\mathbf{C} = \sqrt{n}\tilde{\mathbf{V}}\mathbf{T}^{-1}$ is known as the standardized score coefficient matrix. Postmultiplying the original data by this matrix gives the standardized principal components.

The results of a PCA are often represented in a graph, Gabriel's biplot (1971), by plotting the first two columns of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{H}}$. In this graph, the cosine of an angle between two variable vectors approximates their correlation because:

$$\tilde{\mathbf{H}}\tilde{\mathbf{H}}' = \frac{1}{n}\tilde{\mathbf{V}}\mathbf{T}^2\tilde{\mathbf{V}}' = \frac{1}{n}\mathbf{X}'\mathbf{X}. \tag{8.4}$$

Let $\mathbf{h}_i$ be the i^{th} row of $\tilde{\mathbf{H}}$, and $\mathbf{x}_i$ be the i^{th} column of $\mathbf{X}$. We then find:

$$\begin{aligned} \cos(\mathbf{h}_i, \mathbf{h}_j) &= \frac{\mathbf{h}_i{}'\mathbf{h}_j}{\parallel \mathbf{h}_i \parallel \parallel \mathbf{h}_j \parallel} = \frac{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_j}{\frac{1}{\sqrt{n}} \parallel \mathbf{x}_i \parallel \frac{1}{\sqrt{n}} \parallel \mathbf{x}_j \parallel} \\ &= \frac{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_j}{\sqrt{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_i}\sqrt{\frac{1}{n}\mathbf{x}_j{}'\mathbf{x}_j}} = r(\mathbf{x}_i, \mathbf{x}_j). \end{aligned} \tag{8.5}$$

This result is a full space result. The correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ will be recovered exactly when all columns of $\tilde{\mathbf{H}}$ are considered. In a two-dimensional biplot correlations will be represented perfectly if the data matrix consists of two variables only. With more variables correlations can only be recovered approximately, and the analysis was not designed to optimize this property explicitly. We note that in (8.5) we have $\parallel \mathbf{h}_i \parallel = \sqrt{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_i}$, which means (again

in the full space) that the length of a variable vector represents the standard deviation of the corresponding original variable (here 1). However, because $\tilde{\mathbf{H}}$ is a matrix containing correlations, it is more accurate to say that, by Pythagoras, the length of a variable vector in a two-dimensional biplot is described by:

$$\sqrt{r^2(\mathbf{x}_i, \tilde{\mathbf{F}}_1) + r^2(\mathbf{x}_i, \tilde{\mathbf{F}}_2)}, \tag{8.6}$$

where $\tilde{\mathbf{F}}_1$ and $\tilde{\mathbf{F}}_2$ indicate the first and the second principal component respectively. This states that the length corresponds to a multiple correlation coefficient. Thus, the length of the vector indicates the amount of variance of $\mathbf{x}_i$ explained by a regression onto the first two principal components. The amount of variance explained is used as a quality measure, thus the longer the vector, the better it is represented.

More details on PCA are provided by many textbooks and papers on multivariate analysis, see for instance Dillon (1984, chapter 2) for a introduction, Mardia (1979), Jöreskog (1993), Rao (1964) or the book by Jolliffe (1986) for more theoretical details.

8.3 Supplementary Variables

Say we have a supplementary variable $\mathbf{z}$ that we want to represent by a vector $\boldsymbol{\nu}$ in the biplot. As in chapter 7, we assume it is possible to calibrate such a supplementary vector, and we minimize projections errors of the cases. These projections are given by:

$$\hat{\mathbf{z}} = \alpha \tilde{\mathbf{F}} \boldsymbol{\nu}, \tag{8.7}$$

where α is a normalization factor. We try to minimize:

$$(\mathbf{z} - \hat{\mathbf{z}})'(\mathbf{z} - \hat{\mathbf{z}}) = (\mathbf{z} - \alpha \tilde{\mathbf{F}} \boldsymbol{\nu})'(\mathbf{z} - \alpha \tilde{\mathbf{F}} \boldsymbol{\nu}). \tag{8.8}$$

Note that we project the points corresponding to the cases onto the supplementary variable vector. In the context of the PCA of an abundance matrix, this means that we are projecting the species points, whereas in chapter 7 we minimized projections errors of the site points. In most applications, it will be more natural to project the cases points, corresponding to the rows of the data matrix. In the interpretation of a PCA biplot one projects cases onto variable vectors to approximately recover the data. The natural step is do to the same with respect to the supplementary variable, and therefore to minimize objective function (8.8). It seems not to make much sense to project a variable vector onto a supplementary variable vector. When we consider abundance data, this is somewhat different. If species are considered cases and sites variables, then the natural thing would be to project the sites, and thus to project variables rather than cases. The species points might however, also be projected as they might give an idea of the preferred environment for the species. Therefore, both projections were considered in the context of CA in chapter 7. For the moment, we continue to consider the projection of cases, this being probably more useful in general.

The derivation of the solution, without a unit norm constraint for $\boldsymbol{\nu}$, is analogous to problem (7.4) described in chapter 7. The Lagrangian is given by:

$$L(\boldsymbol{\nu}, \alpha) = \mathbf{z}'\mathbf{z} - 2\alpha\mathbf{z}'\tilde{\mathbf{F}}\boldsymbol{\nu} + \alpha^2\boldsymbol{\nu}'\tilde{\mathbf{F}}'\tilde{\mathbf{F}}\boldsymbol{\nu} \tag{8.9}$$

Setting $\partial L/\partial\boldsymbol{\nu} = \mathbf{0}$ and $\partial L/\partial\alpha = 0$ we find, after some algebra, that the solution is given by:

$$\boldsymbol{\nu} = \frac{1}{\alpha}(\tilde{\mathbf{F}}'\tilde{\mathbf{F}})^{-1}\tilde{\mathbf{F}}'\mathbf{z} = \frac{1}{\alpha}(n\mathbf{I})^{-1}\tilde{\mathbf{F}}'\mathbf{z} = \frac{1}{\alpha n}\tilde{\mathbf{F}}'\mathbf{z}. \tag{8.10}$$

From $\partial L/\partial\alpha = 0$ it follows that:

$$\alpha = \frac{1}{n}\frac{\mathbf{z}'\tilde{\mathbf{F}}\boldsymbol{\nu}}{\boldsymbol{\nu}'\boldsymbol{\nu}}, \tag{8.11}$$

which after substitution in (8.10) gives the solution:

$$\frac{\boldsymbol{\nu}}{\| \boldsymbol{\nu} \|} = \frac{1}{\sqrt{\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}}}\tilde{\mathbf{F}}'\mathbf{z}, \tag{8.12}$$

where we used the property that $\boldsymbol{\nu}/\| \boldsymbol{\nu} \|$ has norm one. Strictly speaking, the solution given by (8.12) is not identified, because if vector $\boldsymbol{\nu}$ is a solution, then any multiple of $\boldsymbol{\nu}$ is also a solution. We can require the norm of $\boldsymbol{\nu}$ to be one, such that the solution is given by:

$$\boldsymbol{\nu} = \frac{1}{\sqrt{\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}}}\tilde{\mathbf{F}}'\mathbf{z}. \tag{8.13}$$

We notice that we could also have introduced the norm one restriction straight at the beginning and minimize:

$$L(\boldsymbol{\nu}, \alpha, \lambda) = \mathbf{z}'\mathbf{z} - 2\alpha\mathbf{z}'\tilde{\mathbf{F}}\boldsymbol{\nu} + \alpha^2\boldsymbol{\nu}'\tilde{\mathbf{F}}'\tilde{\mathbf{F}}\boldsymbol{\nu} + \lambda(1 - \boldsymbol{\nu}'\boldsymbol{\nu}), \tag{8.14}$$

instead of (8.9). When working out the solution of this minimization problem, the Lagrange multiplier λ turns out to be zero, and the solution vector found is the same as given by (8.13).

Standardized principal components have zero mean and unit variance. If we assume $\mathbf{z}$ also to be standardized, then $(1/n)\tilde{\mathbf{F}}'\mathbf{z}$ is a vector of correlation coefficients between the supplementary variable and standardized principal components. Because $\sqrt{\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}}$ is a positive constant, the solution vector $\boldsymbol{\nu}$ is a vector that is proportional to the vector of correlations. If the supplementary variable $\mathbf{z}$ happens to correlate perfectly with one of the principal components, it must be uncorrelated with all others. As a consequence $\boldsymbol{\nu}$ will be an elementary vector consisting of a sole 1 and all other elements equal to zero. Such a variable will thus coincide precisely with one particular axis in the solution of a PCA. We proceed to discuss some properties of the solution, where we assume the norm of $\boldsymbol{\nu}$ to be one.

First, we note that the matrix of second order derivatives of the Lagrangian with respect to $\boldsymbol{\nu}$ is given by $\partial^2 L/\partial\boldsymbol{\nu}^2 = 2\alpha^2\tilde{\mathbf{F}}'\tilde{\mathbf{F}} = 2\alpha^2 n\mathbf{I}$, which is a positive definite matrix. The solution described by (8.13) thus indeed corresponds to a minimum.

8.4 Quality of Representation

The values of variable $\mathbf{z}$ estimated in the biplot become:

$$\hat{\mathbf{z}} = \alpha\tilde{\mathbf{F}}\boldsymbol{\nu} = \frac{1}{n}\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}. \tag{8.15}$$

This equation has a geometrical interpretation. Vector $\hat{\mathbf{z}}$ can be considered to be the projection of vector $\mathbf{z}$ onto the space spanned by the principal components. The associated projector matrix is given by $\tilde{\mathbf{F}}(\tilde{\mathbf{F}}'\tilde{\mathbf{F}})^{-1}\tilde{\mathbf{F}}' = \tilde{\mathbf{F}}(n\mathbf{I})^{-1}\tilde{\mathbf{F}}' = (1/n)\tilde{\mathbf{F}}\tilde{\mathbf{F}}'$. Or, in other words, $\hat{\mathbf{z}}$ is also given by the fitted values of the regression of $\mathbf{z}$ on $\tilde{\mathbf{F}}$.

We can now evaluate the quality of representation of the supplementary variable. Our measure of quality is the amount of variance of the supplementary variable accounted for by the display. In a formula this equals:

$$\frac{\frac{1}{n}\hat{\mathbf{z}}'\hat{\mathbf{z}}}{\frac{1}{n}\mathbf{z}'\mathbf{z}} = \frac{\mathbf{z}'\tilde{\mathbf{F}}(\tilde{\mathbf{F}}'\tilde{\mathbf{F}})\tilde{\mathbf{F}}'\mathbf{z}\frac{1}{n^2}}{\mathbf{z}'\mathbf{z}} = \frac{\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}\frac{1}{n}}{\mathbf{z}'\mathbf{z}} = \frac{\frac{1}{n}\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}\frac{1}{n}}{\frac{1}{n}\mathbf{z}'\mathbf{z}} = \sum_{k=1}^{K} r_k^2(\mathbf{z}, \tilde{\mathbf{F}}_k). \tag{8.16}$$

When $\mathbf{z}$ is centred on the mean, $\tilde{\mathbf{F}}'\mathbf{z}\frac{1}{n}$ is the vector of covariances between principal components and the supplementary variable. Thus, the quality of representation of the variable is the sum of the squared correlations with the principal components, and corresponds to the amount of variance of $\mathbf{z}$ explained by a regression onto principal components.

The length of the supplementary vector $\boldsymbol{\nu}$ in 2D also has a particular interpretation. We work out the length in 2D of the solution vector:

$$\| \boldsymbol{\nu}_{(2)} \| = \frac{\sqrt{\mathbf{z}'\tilde{\mathbf{F}}_{(2)}\tilde{\mathbf{F}}'_{(2)}\mathbf{z}}}{\sqrt{\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}}} = \frac{\sqrt{r^2(\mathbf{z}, \tilde{\mathbf{F}}_1) + r^2(\mathbf{z}, \tilde{\mathbf{F}}_2)}}{\sqrt{\sum r_i^2(\mathbf{z}, \tilde{\mathbf{F}}_i)}} \tag{8.17}$$

Thus, the length corresponds to the square root of the quotient of two amounts of variance explained in the regression of $\mathbf{z}$ on principal components. The quotient is the amount of variance explained in the regression of $\mathbf{z}$ on the first two principal components divided by the amount of variance explained by the regression on all principal components. A long vector tells us that the regression onto the first principal components is successful. This interpretation of the vector length of $\boldsymbol{\nu}$ is maybe not very attractive. The ordinary non-supplementary variable vectors have lengths that reflect the percentage of their variance explained, and so indicate their quality. It would be nice to maintain the same interpretation for supplementary variable vectors. Because any multiple of $\boldsymbol{\nu}$ is also a solution of the minimization problem posed (cf. 8.12), we might as well rescale the vector $\boldsymbol{\nu}$ in order to obtain a more attractive interpretation. We could rescale $\boldsymbol{\nu}$ in such a way that its length *does* reflect the amount of variance explained by the first two principal components, just like ordinary variable vectors in the display. This can be achieved if we choose the norm of $\boldsymbol{\nu}$ in (8.12) to be $(1/n)\sqrt{\mathbf{z}'\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{z}}$, so that solution (8.13) changes to:

$$\boldsymbol{\nu} = \frac{1}{n}\tilde{\mathbf{F}}'\mathbf{z}, \tag{8.18}$$

which is a vector of correlations between principal components and the supplementary variable, if the latter is assumed to be standardized. The 2D length of this vector is $\sqrt{1/(n^2)\mathbf{z}'\tilde{\mathbf{F}}_{(2)}\tilde{\mathbf{F}}'_{(2)}\mathbf{z}} = \sqrt{r^2(\mathbf{z},\tilde{\mathbf{F}}_1) + r^2(\mathbf{z},\tilde{\mathbf{F}}_2)}$ and is the square root of the amount of variance explained, just like for ordinary variables (cf. (8.6)).

8.5 Angles between Variables

As noted before (cf. 8.5), the cosine of an angle between two variable vectors in a PCA biplot approximates the correlation between the variables. How about the angle between a supplementary variable vector and the other variable vectors obtained by PCA? Can we also interpret the cosine of those angles to approximate correlations between the supplementary variables and the other variables? Surprisingly enough, this turns out to be the case. In order to verify this, we define a new minimization problem. We search again for an optimal direction $\boldsymbol{\nu}$ in the PCA biplot that depicts the correlations between $\mathbf{z}$ and $\mathbf{X}$ as best as possible. The rows of $\tilde{\mathbf{H}}$ contain the vectors representing the variables obtained in the PCA. If we indicate correlations between variables by the cosinus of the angle between their vectors, then the correlation between $\mathbf{z}$ and the i^{th} variable used in PCA is estimated by:

$$\cos(\boldsymbol{\nu}, \mathbf{h}_i) = \frac{\mathbf{h}_i'\boldsymbol{\nu}}{\| \mathbf{h}_i \| \| \boldsymbol{\nu} \|} \tag{8.19}$$

The rows of $\tilde{\mathbf{H}}$ are unit norm when the data are standardized (cf. (8.4)). The correlations between $\mathbf{z}$ and $\mathbf{X}$ are estimated in the biplot by $\alpha\tilde{\mathbf{H}}\mathbf{v}$, and we can minimize the sum of squared errors of estimated correlations minus real correlations:

$$(\alpha\tilde{\mathbf{H}}\boldsymbol{\nu} - \frac{1}{n}\mathbf{X}'\mathbf{z})'(\alpha\tilde{\mathbf{H}}\boldsymbol{\nu} - \frac{1}{n}\mathbf{X}'\mathbf{z}), \tag{8.20}$$

what amounts to minimizing:

$$L(\alpha, \boldsymbol{\nu}) = \boldsymbol{\nu}'\tilde{\mathbf{H}}'\tilde{\mathbf{H}}\boldsymbol{\nu} - \frac{2\alpha}{n}\boldsymbol{\nu}\tilde{\mathbf{H}}'\mathbf{X}'\mathbf{z} + (1/n^2)\mathbf{z}'\mathbf{X}\mathbf{X}'\mathbf{z}. \tag{8.21}$$

Setting first order derivatives to zero, we find from $\partial L/\partial\alpha = 0$ that:

$$\alpha\boldsymbol{\nu}'\tilde{\mathbf{H}}'\tilde{\mathbf{H}}\boldsymbol{\nu} = (1/n)\boldsymbol{\nu}'\tilde{\mathbf{H}}'\mathbf{X}'\mathbf{z} \tag{8.22}$$

and from $\partial L/\partial\boldsymbol{\nu} = \mathbf{0}$ that:

$$\begin{aligned} \boldsymbol{\nu} &= \frac{1}{\alpha n}(\tilde{\mathbf{H}}'\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}'\mathbf{X}'\mathbf{z} = \frac{1}{\alpha n}(\tilde{\mathbf{H}}'\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}'(\tilde{\mathbf{V}}\mathbf{T}\tilde{\mathbf{U}}')\mathbf{z} \\ &= \frac{1}{\alpha n}(\tilde{\mathbf{H}}'\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}'\tilde{\mathbf{H}}\mathbf{F}'\mathbf{z} = \frac{1}{\alpha n}\mathbf{F}'\mathbf{z}, \end{aligned} \tag{8.23}$$

and by substituting (8.22) in (8.23) we find that:

$$\frac{\boldsymbol{\nu}}{\| \boldsymbol{\nu} \|} = \frac{1}{\sqrt{\mathbf{z}'\mathbf{F}\mathbf{F}'\mathbf{z}}}\mathbf{F}'\mathbf{z}. \tag{8.24}$$

Thus, it turns out that the solution of this minimization problem is identical to the solution we found before, when we minimized projection errors of the cases onto the supplementary vector, as is described by equation (8.13). We conclude that correlations between the supplementary variable and the ordinary variables are optimally represented.

8.6 A Different Scaling

Results of a PCA are not always reported using standardized principal components as considered so far. Another type of scaling consists of a biplot of non-standardized principal components $\boldsymbol{\Psi}$ and vectors $\boldsymbol{\Theta}$ and can also be calculated from the decomposition in (8.1) as:

$$\boldsymbol{\Psi} = \tilde{\mathbf{U}}\mathbf{T} \qquad \boldsymbol{\Theta} = \tilde{\mathbf{V}}. \tag{8.25}$$

The different types of scaling of the results are described by Gabriel (1971) and Jolliffe (1986, section 5.3). In the scaling used in (8.25) angles between variable vectors do no longer approximate correlations. However, this scaling has the advantage that a biplot really shows the larger dispersion of the first principal component, as $\boldsymbol{\Psi}$ is not standardized.

In this scaling, the objective is to minimize $(\mathbf{z} - \alpha\boldsymbol{\Psi}\boldsymbol{\nu})'(\mathbf{z} - \alpha\boldsymbol{\Psi}\boldsymbol{\nu})$. The optimal direction for a supplementary variable is in this scaling is given by:

$$\frac{\boldsymbol{\nu}}{\| \boldsymbol{\nu} \|} = \frac{1}{\sqrt{\mathbf{z}'\boldsymbol{\Psi}\mathbf{T}^{-4}\boldsymbol{\Psi}'\mathbf{z}}}\mathbf{T}^{-2}\boldsymbol{\Psi}'\mathbf{z}. \tag{8.26}$$

This vector does not correspond to a vector of correlation coefficients between principal components and environmental variables. When this type of scaling is used in PCA, it is thus *not* correct to plot correlation coefficients. We rewrite (8.26):

$$\frac{\boldsymbol{\nu}}{\| \boldsymbol{\nu} \|} = \frac{1}{\sqrt{\mathbf{z}'\boldsymbol{\Psi}\mathbf{T}^{-4}\boldsymbol{\Psi}'\mathbf{z}}}\mathbf{T}^{-2}\boldsymbol{\Psi}'\mathbf{z} = \frac{1}{\sqrt{\mathbf{z}'\tilde{\mathbf{U}}\mathbf{T}^{-2}\tilde{\mathbf{U}}'\mathbf{z}}}\mathbf{T}^{-1}\tilde{\mathbf{U}}'\mathbf{z} = \frac{1}{\sqrt{\mathbf{z}'\tilde{\mathbf{F}}\mathbf{T}^{-2}\tilde{\mathbf{F}}'\mathbf{z}}}\mathbf{T}^{-1}\tilde{\mathbf{F}}'\mathbf{z}, \tag{8.27}$$

where the vector $\tilde{\mathbf{F}}'\mathbf{z}$ is proportional to the correlations between $\tilde{\mathbf{F}}$ and $\mathbf{z}$. However, the multiplication by $\mathbf{T}^{-1}$ makes that its elements are divided by the standard deviations of the principal components in $\mathbf{T}$.

In a two-dimensional biplot, this means that if the eigenvalues of the first two principal components are equal, the optimal direction will coincide with the correlation vector. If there are differences in the eigenvalues, which is usually the case, the two directions will differ, and plotting correlations is mistaken.

8.7 Supplementary Cases

In the framework of this chapter, we complement the material above with an indication of how supplementary cases might be added to a PCA biplot. As indicated above, standardized principal components can be obtained from the original data by postmultiplying these by the standardized score coefficient matrix. If we have a matrix with supplementary cases to depict in the biplot, $\mathbf{X}_{sup}$, then first we center this matrix on the means of the original variables (in matrix $\mathbf{X}$) used in the PCA:

$$\mathbf{X}_{sup} \leftarrow \mathbf{X}_{sup} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}, \tag{8.28}$$

where we note that the new $\mathbf{X}_{sup}$ will in general not have columns with zero mean. Next, the data in $\mathbf{X}_{sup}$ are "standardized" by dividing each variable by the standard deviations of the variables in $\mathbf{X}$. The variance of the columns of $\mathbf{X}_{sup}$ will neither be one. The corresponding coordinates for the supplementary cases ($\mathbf{F}_{sup}$) are now obtained as:

$$\mathbf{F}_{sup} = \mathbf{X}_{sup}\mathbf{C}. \tag{8.29}$$

When we use non-standardized principal components, $\mathbf{\Psi}$, then the latter can be obtained from the data by postmuliplying by $\mathbf{V}$ since $\mathbf{\Psi} = \mathbf{UT} = \mathbf{XV}$. With this scaling, supplementary cases can thus be represented in a PCA-biplot by:

$$\mathbf{F}_{sup} = \mathbf{X}_{sup}\mathbf{V}, \tag{8.30}$$

where $\mathbf{X}_{sup}$ has been centred and standardized as described above.

8.8 An Example

In this section we repeat the PCA of the chemical data from 1990 shown earlier in section 6.3.1, p. 74, where we now use distance as a supplementary variable, and the outlying station 40 as a supplementary case. The biplot of this analysis is shown in figure 8.1.

Using formula (8.29) we find that the coordinates for station 40 to be (0.04,-10.01). Station 40 thus remains a highly outlying point (not shown in figure 8.1) low on most of the variables, except Barium. This interpretation is largely consistent compared with the analysis where station 40 was included as an active point (cf. graph 6.8 p. 74), except for its position with respect to Barium.

The supplementary variable distance is shown by an arrow in figure 8.1, and coincides with the direction separating the inner ring stations 36, 37, 30 and 31 from the outer stations 20, 16, 10, 1 and 6 (cf. figures 2.1, 2.2 on pp. 6 and 7). Distance is seen to be correlated with C18 and degradation parameter Pristane (cf. figure 6.9 p. 75). The distance vector has a 2D length of 0.81. This means that 66% ($0.81^2 = 0.66$) of the variance in distance is accounted for by the first two principal components. Figure 8.1 can be complemented with other geographical information such as East-West or North-South distances, but these variables had a very low quality and are not shown.

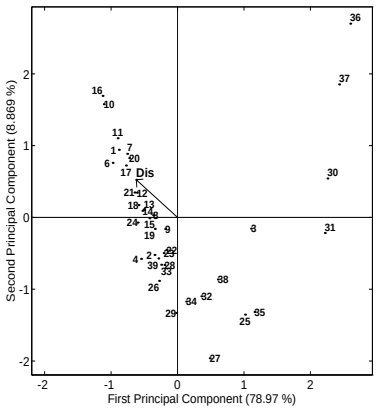

FIGURE 8.1: PCA FOR 1990, STATION 40 AND DISTANCE SUPPLEMENTARY

Figure 8.2 shows the results of analysis, but now using result (8.26) with non-standardized principal components. The larger dispersion along the horizontal axis is now very clear. The length of vector Distance has been multiplied by a factor 10 to make it more visible. The vector length could also have been rescaled to reflect R^2 as in the previous analysis, making the vector fall within the unit circle. However, if the principal components have a large variance, the unit circle becomes very small, making the vectors difficult to see and interpret.

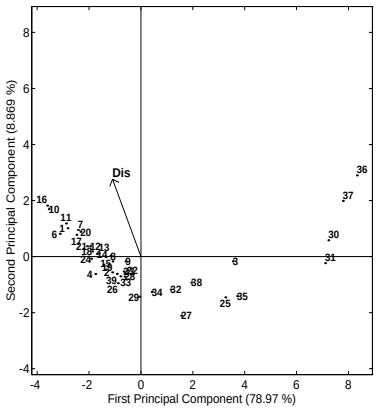

FIGURE 8.2: PCA FOR 1990, WITH PC'S NOT STANDARDIZED

Chapter 9

Theory of Canonical Correspondence Analysis

9.1 Introduction

In the previous chapter, we first performed correspondence analysis to get an optimal picture of the species data, and then tried to fit environmental information to a CA biplot in an optimal way. This procedure belongs to the realm of *indirect gradient analysis*, where latent gradients are extracted from the species data, and environmental data are related to these in a second step.

Canonical correspondence analysis, first described by Ter Braak (1986), is a method for *direct gradient analysis*. In direct gradient analysis, environmental information is used simultaneously with the species data when theoretical gradients are extracted. Over the last decade, canonical correspondence analysis (CCA) has become an important multivariate technique in ecology (Palmer, 1993). One of the mainstays of the method is the assumption of a nonlinear relationship between (linearly combined) environmental axes and species abundance, known as the unimodal response model (Ter Braak, 1985). Many applications of CCA can be found in the ecological literature (Birks and Austin, 1992; Ter Braak, 1994). The behaviour of CCA under varying circumstances (multicollinearity, noise, etc.) has been investigated in various simulation studies (Palmer, 1993; Johnson and Altman, 1999; McCune, 1997).

From a more theoretical point of view, it is possible to arrive at the basic equations of CCA from different perspectives, most of them being described in the literature. For instance, CCA is a maximization of the dispersion of the species scores using a linear restriction on the site scores (Ter Braak, 1987, section 5.5; Johnson, 1999). CA with linear restrictions has also been described by Böckenholt and Takane (1994) and Takane (1991). Alternatively, CCA has been stated to be a weighted least squares approximation to the weighted averages of the species with respect to the environmental variables (Ter Braak, 1986). It is also possible to cast CCA in the framework of reciprocal averaging, where the recip-

rocal averaging algorithm is combined with the regression of site scores onto environmental variables. CCA has also been formulated as a weighted principal component analysis of a matrix of weighted averages (Ter Braak, 1987).

The purpose of this chapter is to give a detailed and transparent mathematical exposition of CCA, parting from the viewpoint of ordinary correspondence analysis (CA). We describe CCA by projecting (scaled) standardized residuals onto a space spanned by environmental variables. A detailed treatment of many theoretical aspects of CCA will be given (singular value decomposition, biplots, bounds for singular values, use of Moore-Penrose inverse, etc), several of which are, to our knowledge, not described in the literature. We will also show that it is possible to do CCA on the basis of a distance matrix. Some illustrative examples using artificial data are presented, reserving applications of CCA to the survey data from the Ekofisk oil field for the next chapter. Special attention will be paid to the issue of the interpretation of the graphical output of CCA.

9.2 Theory of CCA

We start again with the singular value decomposition that is at the heart of ordinary correspondence analysis (CA), and then introduce linear constraints. CA can be performed by the s.v.d. ((5.3), p. 55):

$$\mathbf{D}_r^{-1/2}(\mathbf{P}-\mathbf{r}\mathbf{c}')\mathbf{D}_c{}^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}', \tag{9.1}$$

where $\mathbf{P}$ is the $I \times J$ abundance matrix (species by sites) divided by its grand total ($\mathbf{P}$ being called the correspondence matrix), $\mathbf{r}$ and $\mathbf{c}$ are column vectors containing the row sums and column sums of $\mathbf{P}$ respectively, $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices built from these vectors. Right and left singular vectors are orthogonal, satisfying $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$. Matrix $(\mathbf{P} - \mathbf{r}\mathbf{c}')$ is the matrix containing the deviations from the independence model (no association between rows and columns). In later formulae we will use $\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}'$ to indicate this matrix of deviations. The LHS of (9.1) is known as the matrix of standardized residuals (van der Heijden, 1987, p. 31; Gabriel and Odoroff, 1990, p. 483), divided by a factor of $\sqrt{n}$, where n is the grand total of the abundance matrix. We will refer to matrix $\tilde{\mathbf{P}}$ as the matrix of scaled standardized residuals. Principal and standard coordinates for rows $(\mathbf{F}, \mathbf{\Phi})$ and columns $(\mathbf{G}, \mathbf{\Gamma})$ are obtained as:

$$\begin{aligned} \mathbf{F} &= \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}, \quad & \mathbf{G} &= \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}, \\ \mathbf{\Phi} &= \mathbf{D}_r^{-1/2}\mathbf{U}, & \mathbf{\Gamma} &= \mathbf{D}_c^{-1/2}\mathbf{V}, \end{aligned} \tag{9.2}$$

where the notation of Greenacre (1984) has been adopted. We want to constrain the standard site coordinates to be linear combinations of the environmental variables. Such a constrained analysis can be performed by projecting the rows of the matrix of scaled standardized residuals onto the space spanned by the environmental variables. CCA is, in fact, the CA of these projections[1]. The

[1] This does not mean that one can obtain the CCA solution simply by the use of a program for ordinary CA. This is because a program for CA will usually presuppose that data offered are in raw form, and consequently the program will first divide by the grand total, do the centring operation, operations which are a nuisance in this case. A computer program for doing CCA is given in appendix A.2.

situation is akin to the relationship described by Tenenhaus (1998, chapter 4) between redundancy analysis (van den Wollenberg, 1977) and principal component analysis. Let $\mathbf{Z}$ be the $J \times Q$ matrix of environmental variables, where we assume the columns of $\mathbf{Z}$ to be centred on the weighted means ($\mathbf{c}'\mathbf{Z} = \mathbf{0}$), and standardized by dividing by the square root of the weighted variance. Weighting sites by the square root of their total abundance, the constrained analysis can be performed by postmultiplication of the LHS of (9.1) by the symmetric idempotent projector matrix $\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}$. The constrained analysis can be performed by the s.v.d.:

$$(\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2})(\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}) = \mathbf{U}_l\mathbf{D}_l\mathbf{V}_l^{'}. \tag{9.3}$$

We use the subindex l to distinguish the matrices on the RHS from their counterparts in ordinary CA, as given by (9.1), and to stress that they are linearly constrained dimensions. Introduction of linear constraints in CA by the use of projection matrices has also been described by Böckenholt and Takane (1991). Coordinates for rows (species) and columns (sites) are obtained by the expressions:

$$\begin{aligned} \mathbf{F}_l &= \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l, \quad & \mathbf{G}_l &= \mathbf{D}_c^{-1/2}\mathbf{V}_l\mathbf{D}_l, \\ \mathbf{\Phi}_l &= \mathbf{D}_r^{-1/2}\mathbf{U}_l, & \mathbf{\Gamma}_l &= \mathbf{D}_c^{-1/2}\mathbf{V}_l. \end{aligned} \tag{9.4}$$

The reader will notice that the only difference between these expressions and their counterparts in ordinary CA resides in the subindex l. When there are more sites than variables, as is often the case, not all dimensions in the analysis will be restricted. This seems not to be recognized in many applied studies, as many authors state that as many axes can be extracted as there are variables (Johnson and Altman, 1999, p. 41; Ter Braak, 1986, p. 1167), though their existence is recognized in a later paper by Ter Braak (1994, p. 130). There will be Q restricted dimensions and $J - 1 - Q$ unrestricted dimensions. Unconstrained dimensions in the analysis can be obtained by projecting the rows of the matrix of scaled standardized residuals onto the space orthogonal to the one spanned by the environmental variables. This can be achieved by postmultiplying the LHS of (9.1) by the symmetric idempotent projector matrix $(\mathbf{I}-\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2})$. In order to extract these unconstrained axes, we do a second s.v.d.:

$$(\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2})(\mathbf{I} - \mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}) = \mathbf{U}_u\mathbf{D}_u\mathbf{V}_u^{'}, \tag{9.5}$$

where we now use subindex u to stress that these are unconstrained axes. Species and site coordinates in the unconstrained dimensions are obtained by the same formulae as in (9.4), but changing subindex l for subindex u. Thus, all species and site coordinates can in principle be obtained by two singular value decompositions, and there is no strict need to use a reciprocal averaging algorithm. In his original paper on CCA, Ter Braak (1986, appendix) describes CCA as a decomposition of a species by variables matrix rather than the species by sites matrices considered so far. The s.v.d. described by Ter Braak is:

$$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\tilde{\mathbf{P}}\mathbf{Z})(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} = \mathbf{U}_l\mathbf{D}_l\mathbf{W}', \tag{9.6}$$

with $\mathbf{U}_l'\mathbf{U}_l = \mathbf{I}$ and $\mathbf{W}'\mathbf{W} = \mathbf{I}$. This decomposition is easily obtained from equation (9.3) by postmultiplying by $\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$ and setting $\mathbf{W} =$

$(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{D}_c^{1/2}\mathbf{V}_l$, and focuses only on the constrained dimensions. From s.v.d. (9.6) it is clear that CCA is invariant with respect to multiplication of the environmental data by a scalar. Thus it does not matter whether an environmental variable is expressed in say milligrams per kg or in nanograms per kg, the results of the analysis will be the same. Just like CA, the analysis is also invariant with respect to scalar multiplication of the abundance matrix. When using (9.6), principal and standard coordinates for the species $(\mathbf{F}_l, \mathbf{\Phi}_l)$, variables $(\mathbf{H}, \mathbf{\Omega})$ and sites $(\mathbf{G}_l, \mathbf{\Gamma}_l)$ are found as:

$$\begin{array}{ll} \mathbf{F}_l = \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l, & \mathbf{\Phi}_l = \mathbf{D}_r^{-1/2}\mathbf{U}_l, \\ \mathbf{H} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W}\mathbf{D}_l, & \mathbf{\Omega} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W}, \\ \mathbf{G}_l = \mathbf{\Gamma}_l\mathbf{D}_l, & \mathbf{\Gamma}_l = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W}. \end{array} \tag{9.7}$$

The standard site coordinates are linear combinations of the environmental variables, $\mathbf{\Gamma}_l = \mathbf{Z}\mathbf{B}$, where the matrix with the coefficients of the linear combinations can be obtained as:

$$\mathbf{B} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W}. \tag{9.8}$$

If matrix $\mathbf{\Gamma}_l$ is known, $\mathbf{B}$ can be also be obtained as the matrix of regression coefficients:

$$\mathbf{B} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}_l. \tag{9.9}$$

We note that from $\mathbf{\Gamma}_l\mathbf{D}_c\mathbf{\Gamma}_l = \mathbf{I}$ follows that $\mathbf{B}\mathbf{B}'\mathbf{Z}'\mathbf{D}_c\mathbf{Z}\mathbf{B}\mathbf{B}' = \mathbf{B}\mathbf{B}'$, and if the coefficient matrix $\mathbf{B}$ is of full rank then $\mathbf{B}\mathbf{B}' = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}$. We see that (9.6) is a low rank approximation to the weighted averages of the species (weighted by the square root of their abundance), but postmultiplied by the square root of the inverse of the correlation matrix of environmental variables. If the environmental variables are uncorrelated in the weighted sense, then CCA amounts to a least squares fit to the matrix of weighted averages. However, environmental variables are often correlated, and we arrive at the conclusion that *CCA is, strictly speaking, not a least squares fit to the weighted averages*, in contrast to the first paper about CCA (Ter Braak, 1986, p. 1172).

When environmental variables are standardized, we see that $\mathbf{\Omega}$ represents a matrix of correlation coefficients between environmental variables and standard site coordinates:

$$\mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}_l = \mathbf{Z}'\mathbf{D}_c\mathbf{Z}\mathbf{B} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W} = \mathbf{\Omega}. \tag{9.10}$$

We note that the environmental variables and the unrestricted site coordinates $(\mathbf{\Gamma}_u)$ are uncorrelated. Because the standard site coordinates are uncorrelated in the weighted sense, we have $\mathbf{\Gamma}_u{}'\mathbf{D}_c\mathbf{\Gamma}_l = \mathbf{\Gamma}_u{}'\mathbf{D}_c\mathbf{Z}\mathbf{B} = \mathbf{0}$. Postmultiplication by $\mathbf{B}'(\mathbf{B}\mathbf{B}')^{-1}$ gives $\mathbf{\Gamma}_u{}'\mathbf{D}_c\mathbf{Z} = \mathbf{0}$. From (9.6) we find the loss function of CCA:

$$\| \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} - \mathbf{U}_{(2)}\mathbf{D}_{(2)}\mathbf{W}'_{(2)} \|_E^2 = \| \mathbf{U}_{(r)}\mathbf{D}_{(r)}\mathbf{W}'_{(r)} \|_E^2 =$$

$$\mathrm{tr}(\mathbf{W}_{(r)}\mathbf{D}_{(r)}\mathbf{U}'_{(r)}\mathbf{U}_{(r)}\mathbf{D}_{(r)}\mathbf{W}'_{(r)}) = \mathrm{tr}(\mathbf{D}^2_{(r)}),$$

For the sake of completeness, we note that s.v.d. (9.6) can be rewritten as the spectral decomposition:

$$\mathbf{T}'\mathbf{T} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} = \mathbf{W}\mathbf{D}_l^2\mathbf{W}', \tag{9.11}$$

where we use $\mathbf{T}$ to indicate the LHS of (9.6). After some manipulation, this can be rewritten as:

$$(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}\mathbf{B} = \mathbf{B}\mathbf{D}_l^2, \tag{9.12}$$

with $\mathbf{B} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W}$. This shows that the coefficients of the linear combinations $\mathbf{B}$ can be obtained as eigenvectors of the matrix $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}$, and this is precisely the approach described by Johnson and Altman (1999, p. 41). This result is also reminiscent of canonical correlation analysis, where one searches linear combinations that maximize the correlation between two sets of variables.

9.2.1 Dimensions in the Solution

In ordinary CA the solution has $min(I-1, J-1)$ dimensions, when the trivial one is omitted. The CCA solution has actually the same number of dimensions. Q of these dimensions are restricted, and the remaining ones are not. We can distinguish three situations with respect to the number of variables (Q) and the number of sites (J).

- $J-1 > Q$. This is the normal situation in CCA. Matrix $\mathbf{Z}$ has rank Q, and matrix $\mathbf{Z}'\mathbf{D}_c\mathbf{Z}$, the weighted variance-covariance matrix, is of full rank Q and invertible.
- $J-1 = Q$. In this case precisely all dimensions are linearly constrained. $\mathbf{Z}$ is still of rank Q, and $\mathbf{Z}'\mathbf{D}_c\mathbf{Z}$ is still of full rank Q. The solution is now identical to that of ordinary CA.
- $J-1 < Q$. $\mathbf{Z}$ has rank J, and $\mathbf{Z}'\mathbf{D}_c\mathbf{Z}$ has rank J as well, and is singular if $J < Q$. The solution now equals ordinary CA if we use the Moore-Penrose inverse inverse of $\mathbf{Z}'\mathbf{D}_c\mathbf{Z}$ in the calculations. An analytical proof of this is given in section 9.2.2.

In the second case, $J-1 = Q$, the species and site coordinates obtained from the CCA will equal their CA counterparts. In order to illustrate the equivalence with CA more clearly, imagine that the standard site coordinates obtained with CA *are* exact linear combinations of the environmental variables, e.g. that we have $\mathbf{\Gamma} = \mathbf{Z}$. Then, by (9.10), the variable vectors turn out to be elementary vectors coinciding precisely with the axes of the display. Substituting $\mathbf{Z} = \mathbf{\Gamma}$ in equation (9.9) reduces the coefficient matrix $\mathbf{B}$ to identity, and so shows the standard site coordinates of CCA equal there CA counterparts: $\mathbf{\Gamma}_l = \mathbf{Z}\mathbf{B} = \mathbf{\Gamma}\mathbf{I} = \mathbf{\Gamma}$. On a personal computer, these things are easily verified by feeding $\mathbf{\Gamma}$ obtained by CA as environmental data into a program for CCA.

We note that in the last two cases ($J-1 \leq Q$), it still remains useful to perform CCA rather than CA, because CCA also provides a representation of the environmental data (the variable vectors). In CA there exists a trivial dimension in the solution with an associated singular value of 1. How is this

in decomposition (9.6)? In the current layout with $\mathbf{Z}$ centred on the weighted means and containing environmental variables only, there is no trivial dimension. However, if we do CCA by a reciprocal averaging algorithm (cf. section 9.3.3), a trivial dimension does pop up, and has an associated singular value of 1. This is because in such an algorithm we are regressing on $\mathbf{Z}$, and thus a first column of ones is included in $\mathbf{Z}$ for estimating the intercept. If $\mathbf{Z}$ contains a leading column of ones, then the trivial dimension can be omitted from the solution by using the centred correspondence matrix ($\tilde{\mathbf{P}}$), as we did when we used the s.v.d. in (9.6). It is not sufficient only to centre $\mathbf{Z}$ on the weighted mean, in that case the trivial dimension will remain. Obviously, it does not hurt to centre both. When $\mathbf{Z}$ does not contain a first column of ones, then $\mathbf{Z}$ *must* be centred on the weighted mean, and it is not sufficient to centre only the correspondence matrix. Again, it does not hurt to centre both. A proof for the existence of the trivial dimension is given in section 9.2.3.

9.2.2 Use of the Moore-Penrose Inverse in CCA

Generalized inverses, the Moore-Penrose inverse in particular, play an important role in multivariate analysis. Introductions to the Moore-Penrose inverse can be found in Searle (1982, chapter 8), Magnus and Neudecker (1994, chapter 2), Graybill (1983, chapter 6) and Rao (1971). In this section we show that, when we use the Moore-Penrose inverse in case the covariance matrix of environmental variables is singular, the CCA solution will reduce to the CA solution. This has been published as a linear algebra problem (Graffelman, 1999c). We consider the singular value decomposition of CCA as given by (9.6), and $\mathbf{Z}$ centred by subtracting weighted means without leading column of ones. Then, from (9.6), (9.8) and (9.9):

$$\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} = \mathbf{U}_l\mathbf{D}_l\mathbf{W}' = \mathbf{U}_l\mathbf{D}_l\mathbf{\Gamma}_l'\mathbf{D}_c\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}.$$

Postmultiply by $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{D}_c^{1/2}$ to obtain:

$$(\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2})\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2} = \mathbf{U}_l\mathbf{D}_l\mathbf{\Gamma}_l'\mathbf{D}_c\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}. \tag{9.13}$$

Consider now the case where $Q > J$ and $\mathbf{Z}'\mathbf{D}_c\mathbf{Z}$ is singular. We define $\mathbf{X} = \mathbf{D}_c^{1/2}\mathbf{Z}$, and use the Moore-Penrose inverse of $\mathbf{Z}'\mathbf{D}_c\mathbf{Z}$, which we denote by $(\mathbf{X}'\mathbf{X})^+$. Then we replace $\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}$ by $\mathbf{X}(\mathbf{X}'\mathbf{X})^+\mathbf{X}'$
$= \mathbf{X}\mathbf{X}^+(\mathbf{X}')^+\mathbf{X}' = \mathbf{X}\mathbf{X}^+(\mathbf{X}^+)'\mathbf{X}' = \mathbf{X}\mathbf{X}^+(\mathbf{X}\mathbf{X}^+)'$, where proofs of the properties $(\mathbf{X}'\mathbf{X})^+ = \mathbf{X}^+(\mathbf{X}')^+$ and $(\mathbf{X}')^+ = (\mathbf{X}^+)'$ can be found in Graybill (1983, pp. 108-110).

We use two of the Moore-Penrose conditions, $\mathbf{X}^+ = \mathbf{X}^+\mathbf{X}\mathbf{X}^+$ and $\mathbf{X}^+\mathbf{X} = (\mathbf{X}^+\mathbf{X})'$, substituting the latter in the first: $\mathbf{X}^+ = (\mathbf{X}^+\mathbf{X})\mathbf{X}^+ = (\mathbf{X}^+\mathbf{X})'\mathbf{X}^+ = \mathbf{X}'(\mathbf{X}^+)'\mathbf{X}^+ = \mathbf{X}'(\mathbf{X}\mathbf{X}')^+$. Because $\mathbf{X}$ has full row rank J, $(\mathbf{X}\mathbf{X}')$ is non-singular and thus $(\mathbf{X}\mathbf{X}')^+ = (\mathbf{X}\mathbf{X}')^{-1}$ so that $\mathbf{X}^+ = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$. As a consequence the term $\mathbf{X}\mathbf{X}^+ = \mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} = \mathbf{I}$.
Consequently $\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2} = \mathbf{X}\mathbf{X}^+(\mathbf{X}\mathbf{X}^+)' = \mathbf{I}\mathbf{I}' = \mathbf{I}$.

Equation (9.13) thus reduces to $\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2}\mathbf{I} = \mathbf{U}_l\mathbf{D}_l\mathbf{\Gamma}_l{}'\mathbf{D}_c^{1/2}\mathbf{I}$, from which it follows that $\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{U}_l\mathbf{D}_l\mathbf{\Gamma}_l{}'\mathbf{D}_c^{1/2}$. Assuming no repetitions of singular values and ignoring the indeterminacy of the sign of the singular vectors, the singular value decomposition will be unique and we have $\mathbf{U} = \mathbf{U}_l$, $\mathbf{D} = \mathbf{D}_l$ (thus implying $\mathbf{F} = \mathbf{F}_l$) and $\mathbf{V} = \mathbf{D}_c^{1/2}\mathbf{\Gamma}_l$. Because $\mathbf{V} = \mathbf{D}_c^{1/2}\mathbf{\Gamma} = \mathbf{D}_c^{1/2}\mathbf{\Gamma}_l$, also $\mathbf{\Gamma} = \mathbf{\Gamma}_l$. □

9.2.3 The Trivial Dimension

In this section we prove the existence of a trivial dimension in CCA with an associated singular value of 1, if $\mathbf{Z}$ contains a leading column of ones. The s.v.d. (9.6), with $\mathbf{P}$ not centred, corresponds with the eigenvalue-eigenvector problem:

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1/2}\mathbf{u} = \lambda\mathbf{u}, \tag{9.14}$$

which, by premultiplication by $\mathbf{D}_r^{-1/2}$ becomes:

$$\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{w} = \lambda\mathbf{w}, \tag{9.15}$$

where $\mathbf{w} = \mathbf{D}_r^{-1/2}\mathbf{u}$. We partition $\mathbf{Z}$ as $[\mathbf{1} \mid \mathbf{Z}]$. Substitution in (9.15) gives:

$$\mathbf{D}_r^{-1}\mathbf{P}[\mathbf{1} \mid \mathbf{Z}]\begin{bmatrix} 1 & \mathbf{c}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{c} & \mathbf{Z}'\mathbf{D}_c\mathbf{Z} \end{bmatrix}^{-1}[\mathbf{1} \mid \mathbf{Z}]'\mathbf{P}'\mathbf{w} = \lambda\mathbf{w}. \tag{9.16}$$

Assuming $\mathbf{Z}$ to be centred on the weighted means, we have to find the inverse of the partitioned matrix:

$$\begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{Z}'\mathbf{D}_c\mathbf{Z} \end{bmatrix}^{-1}. \tag{9.17}$$

Inverses of partitioned matrices occur often in multivariate statistics. Expressions for the inverse of a partitioned matrix exist and are described in many textbooks on linear algebra (Magnus and Neudecker, 1994, p. 11). Using these results, the inverse of the matrix above in (9.17) is:

$$\begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1} \end{bmatrix}, \tag{9.18}$$

making that (9.16) reduces to:

$$(\mathbf{1}\mathbf{r}' + \mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}')\mathbf{w} = \lambda\mathbf{w}. \tag{9.19}$$

If we choose $\mathbf{w} = \mathbf{1}$, then the second term on the LHS is a zero matrix because $\mathbf{Z}'\mathbf{P}'\mathbf{1} = \mathbf{Z}'\mathbf{c} = \mathbf{0}$, and we have $\mathbf{1}\mathbf{r}'\mathbf{1} = \mathbf{1} = \lambda\mathbf{1}$. Thus $\lambda = 1$ is an eigenvalue of (9.15) with associated eigenvector $\mathbf{1}$, and correspondingly $\lambda = 1$ is also an eigenvalue of (9.14) with associated eigenvector $\mathbf{D}_r^{1/2}\mathbf{1}$. This dimension is uninteresting because the species coordinates show no variation with respect to this axis.

9.2.4 Inertia Decomposition and Inertia Bounds

In ordinary CA the total inertia of the abundance matrix is given by (cf. equation (5.13), page 58) :

$$\mathrm{tr}(\tilde{\mathbf{P}}\mathbf{D}_c^{-1}\tilde{\mathbf{P}}'\mathbf{D}_r^{-1}). \tag{9.20}$$

If we sum (9.3) and (9.5), we see that in CCA the matrix of scaled standardized residuals is decomposed as:

$$\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2} = \mathbf{U}_l\mathbf{D}_l\mathbf{V}_l^{'} + \mathbf{U}_u\mathbf{D}_u\mathbf{V}_u^{'}. \tag{9.21}$$

Consequently, the total inertia as calculated in ordinary CA, can be decomposed in a constrained and an unconstrained part as:

$$\mathrm{tr}(\tilde{\mathbf{P}}'\mathbf{D}_r^{-1}\tilde{\mathbf{P}}\mathbf{D}_c^{-1}) = \mathrm{tr}(\mathbf{D}_l^2) + \mathrm{tr}(\mathbf{D}_u^2). \tag{9.22}$$

Analogous to CA, principal inertias are also weighted variances of the principal coordinates, and so we find for CCA $\mathbf{F}_l'\mathbf{D}_r\mathbf{F}_l = \mathbf{D}_l^2$ for the restricted dimensions, and $\mathbf{F}_u'\mathbf{D}_r\mathbf{F}_u = \mathbf{D}_u^2$ for the unconstrained dimensions. In a similar manner we have for the sites $\mathbf{G}_l'\mathbf{D}_c\mathbf{G}_l = \mathbf{D}_l^2$ and $\mathbf{G}_u'\mathbf{D}_c\mathbf{G}_u = \mathbf{D}_u^2$.

Just as in ordinary CA, principal inertias can be further decomposed into contributions of the rows (species) and the columns (sites) to the principal inertias, and we can also work out contributions of the principal axes to the row or column inertias, calculate the qualities of representation of the rows in a subspace of certain dimension, and so on. The whole inertia decomposition for rows and columns can be concisely expressed by the respective hadamard products:

$$\mathbf{D}_r([\mathbf{F}_l \mid \mathbf{F}_u] \odot [\mathbf{F}_l \mid \mathbf{F}_u]), \qquad \mathbf{D}_c([\mathbf{G}_l \mid \mathbf{G}_u] \odot [\mathbf{G}_l \mid \mathbf{G}_u]), \tag{9.23}$$

where the columns sums of these matrices give the principal inertias, and the row sums give row and column inertias respectively.

We note that the inertia of the restricted dimensions can also be obtained from (9.3) as:

$$\mathrm{tr}(\mathbf{D}_r(\mathbf{D}_r^{-1}\tilde{\mathbf{P}})\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{D}_r^{-1}\tilde{\mathbf{P}})') = \mathrm{tr}(\mathbf{D}_l^2). \tag{9.24}$$

and also from (9.7) by:

$$\mathbf{H}'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{H} = \mathbf{D}_l\mathbf{W}'\mathbf{W}\mathbf{D}_l = \mathbf{D}_l^2. \tag{9.25}$$

We have shown in section 5.5 that principal inertias in CA are always in the interval $[0, 1]$. Since CCA is a linearly restricted form of CA, we expect the principal inertias in CCA also to be in the $[0, 1]$ interval, and to be smaller or at best equal to principal inertias obtained in CA. We proceed to give a formal algebraical proof of this. It turns out that this problem can be expressed in a more general way as finding bounds for the eigenvalues of the product of a symmetric idempotent and a non-negative definite matrix. As such, the problem has been published (Graffelman and van de Velden, 1999). We first formulate and solve the general problem and then show how it is embedded in CCA.

Let $\mathbf{A}$ be an $n \times n$ non-negative definite matrix, and let $\mathbf{M}$ be a symmetric idempotent matrix. Show that:

$$\lambda_i(\mathbf{AM}) \leq \mu_i(\mathbf{A}) \qquad (i = 1, \ldots, n), \tag{9.26}$$

where $\lambda_i(\cdot)$ and $\mu_i(\cdot)$ represent the eigenvalues of the respective matrices in decreasing order of magnitude.

Since $\mathbf{M}$ is idempotent, it can be factored as $\mathbf{GG}'$, where the $n \times k$ matrix $\mathbf{G}$ is semi-orthogonal, $\mathbf{G}'\mathbf{G} = \mathbf{I}_k$, and k is the rank of $\mathbf{M}$. The eigenvalue equation $\mathbf{AMw} = \lambda\mathbf{w}$ gives $\mathbf{AGG}'\mathbf{w} = \lambda\mathbf{w}$, and can be rewritten as $\mathbf{G}'\mathbf{AGz} = \lambda\mathbf{z}$, where $\mathbf{z} = \mathbf{G}'\mathbf{w}$. Matrices $\mathbf{AM}$ and $\mathbf{G}'\mathbf{AG}$ thus have the same (non-zero) eigenvalues. Moreover, because $\mathbf{A}$ is non-negative definite, $\mathbf{G}'\mathbf{AG}$ is also non-negative definite, hence all eigenvalues λ_i of $\mathbf{AM}$ are larger than or equal to zero.

For the rank of $\mathbf{AM}$ we have: $r(\mathbf{AM}) = r$, where $r = \min(r(\mathbf{A}), r(\mathbf{M}))$. Since the eigenvector-eigenvalue decomposition of $\mathbf{AM}$ can be rephrased as the spectral decomposition of a symmetric matrix ($\mathbf{G}'\mathbf{AG}$ above), we conclude that $\mathbf{AM}$ has exactly $n - r$ zero eigenvalues. Hence, for $i = r + 1, ..., n$, the inequality $\lambda_i \leq \mu_i$ is trivial. In order to prove the result for $i = 1, \ldots, r$ we shall use the following known result (Magnus and Neudecker, 1994, pp. 205-207):

$$\mu_i = \max_{\mathbf{T}'\mathbf{x}=\mathbf{0}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} \leq \max_{\mathbf{C}'\mathbf{x}=\mathbf{0}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} \qquad (i = 1, \ldots, n),$$

where $\mathbf{C}$ is any $n \times (i-1)$ matrix and $\mathbf{T}$ is an $n \times (i-1)$ matrix of orthogonal eigenvectors corresponding to the $i-1$ largest eigenvalues. Since the introduction of extra constraints never increases the maximum, we find:

$$\mu_i = \max_{\mathbf{T}'\mathbf{x}=\mathbf{0}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} \geq \max_{\substack{\mathbf{T}'\mathbf{x}=\mathbf{0} \\ \mathbf{x}=\mathbf{Gy}}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} = \max_{\mathbf{T}'\mathbf{Gy}=\mathbf{0}} \frac{\mathbf{y}'\mathbf{G}'\mathbf{AGy}}{\mathbf{y}'\mathbf{y}} \geq \lambda_i,$$

where we have used the semi-orthogonality of $\mathbf{G}$. □

The result $\lambda_i \leq \mu_i$ directly carries over to CCA if we rewrite s.v.d. (9.6) as a spectral decomposition:

$$\mathbf{T}'\mathbf{T} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{PZ}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} = \mathbf{WD}_l^2\mathbf{W}'. \tag{9.27}$$

We premultiply by $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$, postmultiply by $\mathbf{W}$ and set $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W} = \mathbf{X}$ to find:

$$(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{PZX} = \mathbf{XD}, \tag{9.28}$$

and premultiply by $\mathbf{D}_c^{1/2}\mathbf{Z}$ and set $\mathbf{D}_c^{1/2}\mathbf{ZX} = \mathbf{Y}$ to find:

$$\mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}\mathbf{D}_c^{-1/2}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{PD}_c^{-1/2}\mathbf{Y} = \mathbf{YD}, \tag{9.29}$$

where $\mathbf{M} = \mathbf{D}_c^{1/2}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c^{1/2}$ is easily shown to be idempotent, and $\mathbf{A} = \mathbf{D}_c^{-1/2}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{PD}_c^{-1/2}$ is the non-negative definite matrix subject to a spectral decomposition in ordinary CA (cf. equation (5.9), page 57).

Thus we have shown that the spectral decomposition of CCA can be written in the form $\mathbf{MAv} = \lambda\mathbf{v}$. Matrix $\mathbf{MA}$ has the same eigenvalues as $\mathbf{AM}$ in (9.26) because premultiplication by $\mathbf{A}$ gives $\mathbf{AMAv} = \lambda\mathbf{Av}$ and thus $\mathbf{AMz} = \lambda\mathbf{z}$ with $\mathbf{z} = \mathbf{Av}$.

9.2.5 Quality of Representation

From equation (9.6) we see that the squared singular values (eigenvalues) obtained in CCA can be used as a measure of quality of the representation of matrix $\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$, and from the previous equations (9.20) and (9.22) it is clear that $\mathbf{D}_l^2$ contains the inertias of the species points in the restricted dimensions. To indicate the quality of the display of the abundance data, it is most fair to express the fraction of inertia captured with respect to the total as would be obtained by ordinary CA. The quality of an n-dimensional representation is given by:

$$\frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{Q} d_{li}^2 + \sum_{i=1+Q}^{J-1} d_{ui}^2}, \tag{9.30}$$

d_i indicating the i^{th} singular value in the analysis, whether constrained or not. As indicated in section 9.2, these eigenvalues are weighted variances of the principal coordinates of the species (or sites). They thus indicate the *fraction of inertia of the abundance matrix* that is captured by a low dimensional display. This is however, in contrast to what Ter Braak writes in his original paper on CCA (Ter Braak, 1986, pp. 1172). We cite: "... the measure of goodness of fit expresses the percentage variance of the weighted averages ...". From (9.6) we have however:

$$\mathrm{tr}(\mathbf{D}_l^2) = \mathrm{tr}((\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{PZ}), \tag{9.31}$$

which does not correspond to a weighted variance of the weighted averages, as the covariance matrix of the latter would be described by:

$$(\mathbf{D}_r^{-1}\mathbf{PZ})'\mathbf{D}_r(\mathbf{D}_r^{-1}\mathbf{PZ}) = \mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{PZ}. \tag{9.32}$$

Comparing these two equations, we see that the trace of (9.32) will equal (9.31) if the environmental variables are uncorrelated in the weighted sense. Thus, *the interpretation of the eigenvalues as fractions of the (weighted) variance of the weighted averages is correct if and only if the environmental variables are uncorrelated in the weighted sense.* This will hardly ever occur in practice, as environmental variables tend to be correlated. Weighted uncorrelatedness can be achieved if we, prior to performing CCA, reduce the amount of environmental variables by a weighted principal component analysis. Thus, we conclude that the proper interpretation of the eigenvalues is that they are inertias of the abundance matrix. A numerical example in section 9.5 will help to further clarify this issue.

CCA is usually employed with the idea to get a picture of the species optima with respect to the environmental variables. If the eigenvalues indicate only how well the abundance data are displayed, then it is thus very important to work out another statistic: how much of the variance in the weighted averages of

the species is accounted for by a low dimensional display? Estimated weighted averages of the species in the display are given by $\mathbf{F}_l\mathbf{\Omega}'$ (see also (9.39)), because the rows of $\mathbf{\Omega}$ are of norm one if the variables are standardized. The weighted variance accounted for by a 2-D map is thus given by the diagonal elements of the matrix:

$$(\mathbf{F}_{l(2)}\mathbf{\Omega}'_{(2)})'\mathbf{D}_r(\mathbf{F}_{l(2)}\mathbf{\Omega}'_{(2)}) = \mathbf{\Omega}_{(2)}\mathbf{F}'_{l(2)}\mathbf{D}_r\mathbf{F}_{l(2)}\mathbf{\Omega}'_{(2)} = \mathbf{\Omega}_{(2)}\mathbf{D}^2_{(2)}\mathbf{\Omega}'_{(2)}. \quad (9.33)$$

The fraction of weighted variance explained is 2D thus given by $\mathrm{tr}(\mathbf{\Omega}_{(2)}\mathbf{D}^2_{(2)}\mathbf{\Omega}'_{(2)})/\mathrm{tr}(\mathbf{\Omega}\mathbf{D}^2_l\mathbf{\Omega}')$. As noted before in equation (9.10), matrix $\mathbf{\Omega}$ represents a matrix of weighted correlations, and thus we can write the variance fraction explained by a k-dimensional solution in scalar form as:

$$\frac{\sum_{i=1}^{Q}\sum_{l=1}^{k} {r_{il}}^2 {d_{ll}}^2}{\sum_{i=1}^{Q}\sum_{l=1}^{Q} {r_{il}}^2 {d_{ll}}^2}, \quad (9.34)$$

where r_i is the weighted correlation between environmental variables and restricted site coordinates. Notice that this development is entirely analogous to what we did when we considered the quality of display of the weighted averages in the indirect approach. (cf. section 7.3 p. 87). From (9.34) it is clear that if there are only one or two variables involved, the display of the weighted averages will be perfect (e.g. with two variables $Q = k = 2$). This is illustrated with a numerical example in section 9.5. We can also calculate the qualities of representation of the weighted averages for each variable separately, by using just one row of matrix $\mathbf{\Omega}$. The quality of representation of the weighted averages in k dimensions with respect to the i^{th} variable only is then given by:

$$\sum_{l=1}^{k} {r_{il}}^2 {d_{ll}}^2 / \sum_{l=1}^{Q} {r_{il}}^2 {d_{ll}}^2. \quad (9.35)$$

Last, we evaluate the quality of representation of the matrix of environmental variables, $\mathbf{Z}$. This matrix is approximated by the projections of the site points onto the variable vectors, given by $\mathbf{\Gamma}_l\mathbf{\Omega}'$. The weighted variance explained by a 2D map is then $(\mathbf{\Gamma}_{l(2)}\mathbf{\Omega}'_{(2)})'\mathbf{D}_c\mathbf{\Gamma}_{l(2)}\mathbf{\Omega}'_{(2)} = \mathbf{\Omega}_{(2)}\mathbf{\Omega}'_{(2)}$. The fraction of variance explained then becomes:

$$\mathrm{tr}(\mathbf{\Omega}_{l(2)}\mathbf{\Omega}'_{l(2)})/\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}) = \mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}\mathbf{W}\mathbf{W}')/\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}), \quad (9.36)$$

where we used that $\mathbf{W} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{\Omega}$. Thus, if environmental variables are uncorrelated (in the weighted sense), then we have $\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}) = \mathrm{tr}(\mathbf{I}_q) = Q$ and $\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}\mathbf{W}\mathbf{W}') = \mathrm{tr}(\mathbf{W}\mathbf{W}') = \mathrm{tr}(\mathbf{W}'\mathbf{W}) = \mathbf{I}_{(k)} = k$. The quality of the display is then just k/Q, with k the number of dimensions chosen for representation (usually two) and Q the number of variables. For instance, a CCA with 3 uncorrelated environmental variables will always explain 2/3 of the weighted variance of the environmental variables in a two-dimensional biplot.

Since we know that $\mathbf{\Omega}$ contains the correlations between the environmental variables and the restricted axes of the CCA solution, we can also write (9.36) as:

$$\mathrm{tr}(\mathbf{\Omega\Omega}')/\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}) = \mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}_l\mathbf{\Gamma}_l'\mathbf{D}_c\mathbf{Z})/\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z}) \tag{9.37}$$

Matrix $\mathbf{\Gamma}_l\mathbf{\Gamma}_l'\mathbf{D}_c$ is, when all columns of $\mathbf{\Gamma}_l$ are considered, an idempotent centring matrix, centring $\mathbf{Z}$ on the weighted mean (the situation is analogous to equation (5.18) on page 59). However, $\mathbf{Z}$ is already centred on the weighted mean, and thus $\mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}_l\mathbf{\Gamma}_l'\mathbf{D}_c\mathbf{Z}) = \mathrm{tr}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})$. Consequently, we also arrive at the conclusion that when we have only one or two variables, $\mathbf{Z}$ can be displayed perfectly in two dimensions.

Equation (9.36) can also be written in scalar form as:

$$\frac{1}{Q}\sum_{i=1}^{Q}\sum_{j=1}^{2} r_{ij}^2\,, \tag{9.38}$$

showing that if the correlations of the variables with the first two axes of the display are high, then we will explain a relatively large percentage of the variance of $\mathbf{Z}$.

9.2.6 Biplots and Calibrations in CCA

Asymmetric maps in ordinary CA are biplots (Greenacre, 1993a; 1993b). Vectors in a biplot can be calibrated, and tickmarks can be drawn along the vector, once the length of one unit along a variable or site vector has been calculated (Greenacre, 1993b, pp 107-108).

Equations (9.3) and (9.6) form approximations of species by sites and species by variables matrices respectively. Equation (9.6) shows that we when we plot the first two columns of $\mathbf{F}_l$ and $\mathbf{\Omega}$, we approximate the matrix of weighted averages of the species with respect to the variables:

$$\mathbf{D}_r^{-1}\tilde{\mathbf{P}}\mathbf{Z} = \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l\mathbf{W}'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2} = \mathbf{F}_l\mathbf{\Omega}'. \tag{9.39}$$

We can obtain a biplot of the weighted averages. Rewriting this in scalar notation we can recover the weighted average of species i on variable q as:

$$\sum_{j=1}^{J}(\frac{p_{ij}}{r_i} - c_j)z_{jq} \approx f_{i1}\omega_{q1} + f_{i2}\omega_{q2}\,. \tag{9.40}$$

The left hand side of this equation expresses the differences in weighted average of species i with respect to variable q from the over-all weighted average of variable q.

When we look only at the restricted dimensions of decomposition (9.21) we get an approximation of the centred row profiles when we plot the first two columns of $\mathbf{F}_l$ and $\mathbf{\Gamma}_l\mathbf{D}_c$:

$$\mathbf{D}_r^{-1}\tilde{\mathbf{P}} = \mathbf{D}_r^{-1/2}(\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{D}_c^{-1/2})\mathbf{D}_c^{1/2} \approx \mathbf{D}_r^{-1/2}(\mathbf{U}_l\mathbf{D}_l\mathbf{V}_l')\mathbf{D}_c^{1/2} = \mathbf{F}_l(\mathbf{D}_c\mathbf{\Gamma}_l)', \tag{9.41}$$

what we can write in scalar notation as:

$$(\frac{p_{ij}}{r_i} - c_j) \approx f_{i1}(c_j \gamma_{1j}) + f_{i2}(c_j \gamma_{2j}), \tag{9.42}$$

where now the profiles of the species across the stations are approximated in 2D by the scalar products of species and rescaled site vectors.

It is tempting also to project site points onto the variable vectors in order to estimate values for environmental variables at the sites. There is some justification for this since:

$$\mathbf{\Gamma}_l \mathbf{\Omega}' = \mathbf{ZB\Omega}' = \mathbf{ZBW}'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2} = \mathbf{Z}. \tag{9.43}$$

Projecting standard site coordinates onto the variables axes reconstitutes exactly our environmental data matrix. But equation (9.43) is a full space result, that is, we will recover $\mathbf{Z}$ when we consider all columns of $\mathbf{\Gamma}_l$ and $\mathbf{\Omega}$. If we use a subset of the columns of $\mathbf{\Gamma}_l$ and $\mathbf{\Omega}$, we will "approximate" $\mathbf{Z}$, but we don't know how good or how bad, as in the analysis we did no explicit least squares approximation to $\mathbf{Z}$.

In order to investigate if the display of the environmental data matrix $\mathbf{Z}$ is optimal in any sense, consider the following argument. The s.v.d (9.6) decomposes a species by variables matrix, producing in the first place coordinates for species and variables. We can consider adding the site points in a supplementary manner, and try to represent them as best as possible, *given the species scores and variable vectors*. That is to say, we optimize the display of $\mathbf{Z}$ conditional on the fact that first the LHS of (9.6) is optimally represented. This amounts to adapting a minimization problem solved previously ((7.28) p. 90 or (8.26) p. 105) for this situation. One row of the environmental data matrix $\mathbf{Z}$, here indicated by the $Q \times 1$ column vector $\mathbf{z}_j$, is represented as a supplementary vector $\mathbf{g}$ in the biplot. $\mathbf{z}_j$ is estimated in the biplot by $\alpha\mathbf{\Omega g}$, as the rows of $\mathbf{\Omega}$ have norm one. The minimization problem is thus:

$$\mathbf{e}'\mathbf{e} = (\mathbf{z}_j - \alpha\mathbf{\Omega g})'(\mathbf{z}_j - \alpha\mathbf{\Omega g}), \tag{9.44}$$

which has the solution:

$$\frac{\mathbf{g}}{\| \mathbf{g} \|} = \frac{1}{\sqrt{\mathbf{z}_j'\mathbf{\Omega}(\mathbf{\Omega}'\mathbf{\Omega})^{-2}\mathbf{\Omega}'\mathbf{z}_j}}(\mathbf{\Omega}'\mathbf{\Omega})^{-1}\mathbf{\Omega}'\mathbf{z}_j. \tag{9.45}$$

If we are willing to minimize $\mathbf{e}'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{e}$ rather than $\mathbf{e}'\mathbf{e}$, then we have the solution vector:

$$\frac{\mathbf{g}}{\| \mathbf{g} \|} = \mathbf{\Omega}'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{z}_j, \tag{9.46}$$

which, in matrix form, gives us all site coordinates as $\mathbf{D}_g\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{\Omega}$, $\mathbf{D}_g$ taking care of the normalization of the rows. This is precisely the matrix of standard site coordinates obtained in CCA, since by (9.7) $\mathbf{\Gamma}_l = \mathbf{ZB} = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W} = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{\Omega}$, but then with rows normalized to one. Thus, we find that the representation of $\mathbf{Z}$ is optimal, conditional on the display of the species by variables matrix, and using a transformation of the errors by multiplying them by $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$.

9.2.7 Geometrical Properties: Distances, Angles and Vector Lengths.

It is known that in ordinary CA, the Euclidean distances between principal coordinates in an asymmetric map represent χ^2-distances between the row profiles. In CCA, the distance interpretation between species points (in the restricted dimensions) is as follows. If we call the weighted averages of the species, $\mathbf{T} = \mathbf{D}_r^{-1}\tilde{\mathbf{P}}\mathbf{Z}$, then we have:

$$\mathbf{T}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{T}' = \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l\mathbf{W}'\mathbf{W}\mathbf{D}_l\mathbf{U}_l'\mathbf{D}_r^{-1/2} = \mathbf{F}_l\mathbf{F}_l'. \tag{9.47}$$

It follows that the Euclidean distance between the principal coordinates of the species represents the weighted Mahalanobis distance between the weighted averages of the species with respect to the environmental variables: $d_M^2(\mathbf{t}_i, \mathbf{t}_{i'}) = (\mathbf{t}_i - \mathbf{t}_{i'})'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{t}_i - \mathbf{t}_{i'}) = (\mathbf{f}_i - \mathbf{f}_{i'})'(\mathbf{f}_i - \mathbf{f}_{i'}) = d_E^2(\mathbf{f}_i, \mathbf{f}_{i'})$.

From (9.21), and using the centred row profiles $\mathbf{R} = \mathbf{D}_r^{-1}\tilde{\mathbf{P}}$, we have:

$$\mathbf{R}\mathbf{D}_c^{-1}\mathbf{R}' = \mathbf{F}_l\mathbf{F}_l' + \mathbf{F}_u\mathbf{F}_u'. \tag{9.48}$$

This means that the Euclidean distance between the principal coordinates of the species also represents the χ^2-distance between the row profiles: $d_{\chi^2}^2(\mathbf{r}_i, \mathbf{r}_{i'}) = (\mathbf{r}_i - \mathbf{r}_{i'})'\mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{r}_{i'}) = (\mathbf{f}_i - \mathbf{f}_{i'})'(\mathbf{f}_i - \mathbf{f}_{i'}) = d_E^2(\mathbf{f}_i, \mathbf{f}_{i'})$. Thus, Euclidean distances between species points have a double distance interpretation.

We can also consider the distances between the sites, irrespective of their species composition, but just on the basis of their chemical constitution, as in a PCA. We have, using $\mathbf{B}\mathbf{B}' = (\mathbf{Z}\mathbf{D}_c\mathbf{Z})^{-1}$:

$$\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{Z}\mathbf{B}\mathbf{B}'\mathbf{Z}' = \mathbf{\Gamma}_l\mathbf{\Gamma}_l{}'. \tag{9.49}$$

This means that the Euclidean distance between the standard site scores represents a weighted Mahalanobis distance between the sites, using only environmental information. $d_M^2(\mathbf{z}_i, \mathbf{z}_{i'}) = (\mathbf{z}_i - \mathbf{z}_{i'})'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{z}_i - \mathbf{z}_{i'}) = (\gamma_i - \gamma_{i'})'(\gamma_i - \gamma_{i'}) = d_E^2(\gamma_i, \gamma_{i'})$. The joint ordination diagram of species, sites and variables is called a triplot, and we consider the interpretation of the angles in the triplot between the variable vectors. Because

$$\mathbf{\Omega}\mathbf{\Omega}' = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z}\mathbf{B})(\mathbf{B}'\mathbf{Z}'\mathbf{D}_c\mathbf{Z}) = \mathbf{Z}'\mathbf{D}_c\mathbf{Z}, \tag{9.50}$$

which is again a full space result, we find:

$$\cos(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) = \frac{\boldsymbol{\omega}_i'\boldsymbol{\omega}_j}{\parallel \boldsymbol{\omega}_i \parallel \parallel \boldsymbol{\omega}_j \parallel} = \frac{\mathbf{z}_i'\mathbf{D}_c\mathbf{z}_j}{\sqrt{\mathbf{z}_i'\mathbf{D}_c\mathbf{z}_i}\sqrt{\mathbf{z}_j'\mathbf{D}_c\mathbf{z}_j}}, \tag{9.51}$$

and so the cosine of the angle between two variable vectors represents a weighted correlation coefficient between environmental variables, where the weights are the square roots of the total abundances at the sites. Finally, the length of a variable vector is proportional to the weighted variance of an environmental variable, because:

$$\parallel \boldsymbol{\omega}_j \parallel = \sqrt{\boldsymbol{\omega}_j'\boldsymbol{\omega}_j} = \sqrt{\mathbf{z}_j'\mathbf{D}_c\mathbf{z}_j}. \tag{9.52}$$

9.2.8 Invariance of CCA

It has been mentioned before that CCA is invariant under scalar multiplication of the environmental data and under scalar multiplication of the species data. With scale invariant we mean that "results" remain the same if we multiply the data by a scalar. Mardia (1979) has noted that canonical correlation analysis (CCR) is invariant under non-singular linear transformations of the data. In this section we investigate whether CCA is invariant under non-singular linear transformations of the environmental data. Consider we do a linear transformation of the environmental data by postmultiplying $\mathbf{Z}$ by a $Q \times Q$ non-singular matrix $\mathbf{Q}$, such that we get new environmental data $\tilde{\mathbf{Z}} = \mathbf{ZQ}$. Calling the LHS of (9.6) $\mathbf{T}$, we can rewrite (9.6) as the spectral decomposition of $\mathbf{TT}'$, and substitute $\tilde{\mathbf{Z}}$ for $\mathbf{Z}$ to find:

$$\mathbf{TT}' = \mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\mathbf{D}_c\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{P}}'\mathbf{D}_r^{-1/2} = \mathbf{U}_l\mathbf{D}_l^2\mathbf{U}_l', \tag{9.53}$$

and substituting $\tilde{\mathbf{Z}} = \mathbf{ZQ}$ we find:

$$\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{ZQ}(\mathbf{Q}'\mathbf{Z}'\mathbf{D}_c\mathbf{ZQ})^{-1}\mathbf{Q}'\mathbf{Z}'\tilde{\mathbf{P}}'\mathbf{D}_r^{-1/2}\mathbf{U}_l = \mathbf{U}_l\mathbf{D}_l^2, \tag{9.54}$$

and we see that $\mathbf{Q}$ disappears, as this reduces to:

$$\mathbf{D}_r^{-1/2}\tilde{\mathbf{P}}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\tilde{\mathbf{P}}'\mathbf{D}_r^{-1/2}\mathbf{U}_l = \mathbf{U}_l\mathbf{D}_l^2. \tag{9.55}$$

Thus, the eigenvalues, and consequently the decomposition in principal inertias remains unaltered. The eigenvectors $\mathbf{U}_l$ also remain the same, and consequently the species coordinates will neither change. We can do the same for the spectral decomposition of $\mathbf{T}'\mathbf{T}$ (cf. equation (9.11)), what also leads to the conclusion that the eigenvalues remain unaltered, but that the vectors for the environmental variables do alter. In short, the only matrices of the analysis that are altered under linear transformation of the environmental variables are $\mathbf{\Omega}$ and $\mathbf{B}$.

9.3 Relationships with Other Methods

In this section we comment on the relationship between CCA and principal coordinates analysis, weighted principal component analysis and explain a reciprocal averaging algorithm for CCA.

9.3.1 Principal Coordinates Analysis

In this section we develop a distance-based approach to canonical correspondence analysis, and show this to be equivalent to the analysis based on cases by variables matrices. This section was previously presented at the Spanish Biometry Conference (Graffelman, 1999a).

Multivariate methods like principal component analysis, multiple regression, canonical correlation analysis and others usually operate on data coded in a cases by variables matrix. Cluster analysis, multidimensional scaling methods and principal coordinates analysis on the other hand require a (symmetric) distance or similarity matrix, and the object of the analysis is to represent distances

between cases and/or objects as well as possible in a two-dimensional plane.

Some of the methods traditionally based on a cases by variables matrix have been shown to have equivalent distance-based formulations. Notably, Gower (1966) has shown that is possible to do a principal component analysis by doing principal coordinates analysis (PCO) using a matrix of Euclidean distances. The equivalence between PCO and PCA is well-known, and was the subject of a recently published linear algebra problem (van de Velden et al., 1999; Graffelman, 1999b). Digby and Kempton (1987, p. 90) wrote that it is possible to approximate correspondence analysis (CA) by doing a PCO on a matrix of χ^2-distances. However, Greenacre (1984) showed that correspondence analysis is exactly equivalent to "two dual principal coordinates analyses", if we weight and double-centre the distance matrices in the right way.

As PCA and CA can be formulated in a distance-based manner, it should also be possible to perform the canonical form of correspondence analysis in a distance-based manner, although it is not immediately evident which distances one needs to consider, and what distance measure one should use. In the next section a distance-based approach to CCA will be developed, and its equivalence to the usual approach will be shown.

We start again with the singular value decomposition:

$$\mathbf{T} = \mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2} = \mathbf{U}_l\mathbf{D}_l\mathbf{W}', \tag{9.56}$$

where $\mathbf{Z}$ is assumed to be centred on the weighted mean. Principal coordinates of the species are now found as $\mathbf{F}_l = \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l$, standard coordinates of the sites are given by $\mathbf{\Gamma}_l = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W}$, and the variables are represented by $\mathbf{\Omega} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W}$. Note that the site coordinates are standardized and uncorrelated in the weighted sense because $\mathbf{\Gamma}_l'\mathbf{D}_c\mathbf{\Gamma}_l = \mathbf{W}'\mathbf{W} = \mathbf{I}$. Note also that, if the variables are standardized by dividing them by the square root of their weighted variance, then the coordinates for the variables are actually weighted correlations between $\mathbf{\Gamma}_l$ and $\mathbf{Z}$, since $\mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}_l = \mathbf{Z}'\mathbf{D}_c\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W} = \mathbf{\Omega}$. Equation (9.56) shows that CCA is, in fact, a weighted least squares fit to the matrix of (centred) weighted averages of the species with respect to the environmental variables, postmultiplied by the square root of the inverse of the correlation matrix. These weighted averages are estimates of the optima of the species for the variables, where the responses of the species with respect to the variables are assumed to follow a Gaussian curve.

We continue by exposing a distance-based approach, and shows that it is equivalent to (9.56). The development is similar to the distance-based approach to CA, described by Greenacre (1984 pp. 81-82). First, we construct the matrix of weighted averages ("optima") of the species with respect to the Q environmental variables:

$$\mathbf{X} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}, \tag{9.57}$$

and consider the distances between the optima of the species. Rather than using Euclidean distances, we use a weighted Mahalanobis distance. The squared

distance between the optima of two species i and i', taking into account all environmental variables, can then be described by:

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_{i'})'(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{x}_i - \mathbf{x}_{i'}). \tag{9.58}$$

This weighted Mahalanobis distance is scale-invariant, that is, changing the scale of measurement of the variables will not affect the distance between the species optima. Consider also the matrix $\mathbf{S}$ of scalar products between the optima, using the Mahalanobis metric: $\mathbf{S} = \mathbf{X}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{X}'$. The $I \times I$ distance matrix $\mathbf{\Delta}_M$ between the optima of the species can then be obtained as:

$$\mathbf{\Delta}_M = \mathbf{s}\mathbf{1}' + \mathbf{1}\mathbf{s}' - 2\mathbf{S}, \tag{9.59}$$

where $\mathbf{s} = \text{diag}(\mathbf{S})$. What follows next is essentially a weighted PCO, with a particular centring of $\mathbf{\Delta}_M$. In PCO, the first step is to double-centre the distance matrix, by subtracting row and column means and adding the overall mean. This can be achieved by pre and post multiplication of the distance matrix by an idempotent centring matrix, $\mathbf{Q} = (\mathbf{I} - \mathbf{1}\mathbf{r}')$ so that:

$$\mathbf{Q}\mathbf{\Delta}_M\mathbf{Q}' = \mathbf{Q}\mathbf{s}\mathbf{1}'\mathbf{Q}' + \mathbf{Q}\mathbf{1}\mathbf{s}'\mathbf{Q}' - 2\mathbf{Q}\mathbf{S}\mathbf{Q}' = -2\mathbf{Q}\mathbf{S}\mathbf{Q}', \tag{9.60}$$

because $\mathbf{Q}\mathbf{1} = (\mathbf{I} - \mathbf{1}\mathbf{r}')\mathbf{1} = \mathbf{1} - \mathbf{1}\mathbf{r}'\mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}$. After double-centring, the weighted means of the columns and of the rows are zero, because $\mathbf{r}'\mathbf{Q} = \mathbf{r}'(\mathbf{I} - \mathbf{1}\mathbf{r}') = \mathbf{r}' - \mathbf{r}' = \mathbf{0}$. Equation (9.60) can be rewritten as $-\frac{1}{2}\mathbf{Q}\mathbf{\Delta}_M\mathbf{Q}' = \mathbf{Q}\mathbf{S}\mathbf{Q}'$. Notice that if $\mathbf{S}$ is calculated using a centred $\mathbf{Z}$, then the double-centred distance matrix is given simply by $-2\mathbf{S}$, the transformation of $\mathbf{S}$ not being necessary. Next, we weight the species by their total abundance, so that we obtain:

$$-\frac{1}{2}\mathbf{D}_r^{1/2}\mathbf{Q}\mathbf{\Delta}_M\mathbf{Q}'\mathbf{D}_r^{1/2} = \mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})'\mathbf{D}_r^{1/2} \tag{9.61}$$

As in PCO, we do a least squares fit to the scalar product matrix on the RHS of (9.61), where the optimal plane is found by the spectral decomposition:

$$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})'\mathbf{D}_r^{1/2} = \mathbf{U}_l\mathbf{D}_l^2\mathbf{U}_l' = \mathbf{T}\mathbf{T}' \tag{9.62}$$

We can rewrite this as:

$$(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z})' = \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l^2\mathbf{U}_l'\mathbf{D}_r^{-1/2} = \hat{\mathbf{U}}'\mathbf{D}_l^2\hat{\mathbf{U}}, \tag{9.63}$$

where $\hat{\mathbf{U}} = \mathbf{D}_r^{-1/2}\mathbf{U}$. The principal coordinates in this weighted analysis are $\mathbf{F} = \hat{\mathbf{U}}\mathbf{D}_l = \mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l$. We tentatively called the eigenvectors of (9.62) $\mathbf{U}_l$, as they relate to the left singular vectors in (9.56). Thus, the principal coordinates found in the PCO of weighted mahalanobis distances between optima correspond exactly with the species coordinates of CCA.

The analysis is not yet complete. CCA also produces coordinates for sites and variables. How can these be obtained in the distance-based approach? In order to find the coordinates of the sites, we consider a second distance matrix $\mathbf{\Delta}_{M2}$. It is a $J \times I$ distance matrix between the optima of J hypothetical species that occur only at one particular site ("vertex species" representing the sites) and the I "ordinary" species. Double-centring this matrix, we get:

$$-\frac{1}{2}\mathbf{R}\boldsymbol{\Delta}_{M2}\mathbf{Q}' = (\mathbf{IZ})(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}(\mathbf{D}_r^{-1}\mathbf{PZ})', \tag{9.64}$$

Where $\mathbf{R}$ is the idempotent centring matrix($\mathbf{I}-\mathbf{1c}'$). The weighted row means and column means of (9.64) are both zero ($\mathbf{r}'\mathbf{Q}=\mathbf{0}'$, $\mathbf{c}'\mathbf{R}=\mathbf{0}'$). We consider the J rows of (9.64) as supplementary vectors, and project them onto the optimal plane provided by PCO in equation (9.62). We therefore need the projector matrix $\mathbf{P}_r = \mathbf{U}_l(\mathbf{U}_l'\mathbf{U}_l)^{-1}\mathbf{U}_l'$, and the coordinates of the sites, with respect to basis $\mathbf{U}_l$ are obtained as:

$$\mathbf{G} = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1/2}\mathbf{U}_l. \tag{9.65}$$

These are principal coordinates, the standard coordinates $\boldsymbol{\Gamma}$ being obtained by postmultiplying by $\mathbf{D}_l^{-1}$. Because $\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1/2}\mathbf{U}_l\mathbf{D}_l^{-1} = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{T}'\mathbf{U}_l\mathbf{D}_l^{-1} = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W}\mathbf{D}_l\mathbf{U}_l'\mathbf{U}_l\mathbf{D}_l^{-1} = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W}$, the equivalence with the coordinates obtained by the singular value decomposition in (9.56) is clear.

The last step is to obtain coordinates for the environmental variables. Given an ordination diagram with species and site coordinates, optimal directions for the environmental variables can be found by plotting (weighted) correlation coefficients between the variables and the standard coordinates of the sites, the standard coordinates of the variables are so obtained as:

$$\boldsymbol{\Omega} = \mathbf{Z}'\mathbf{D}_c\boldsymbol{\Gamma}_l = \mathbf{Z}'\mathbf{D}_c\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}\mathbf{W} = (\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{1/2}\mathbf{W}, \tag{9.66}$$

where $\mathbf{Z}$ now has been assumed to be standardized in the weighted sense.

9.3.2 Principal Component Analysis

In one of the earlier papers on CCA, Ter Braak (1987, p. 76, appendix) mentions CCA to be equivalent to a weighted principal component analysis (WPCA) applied to a matrix of weighted averages. In a previous section, it was noted that angles between variable vectors in a CCA represent weighted correlations, vector length variances, which also suggest that we are close to some kind of principal component analysis. In this section we elaborate the relationship between PCA and CCA in more detail. A weighted principal component analysis can be performed by the singular value decomposition:

$$\mathbf{D}_w\mathbf{X} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}', \tag{9.67}$$

where we use hats (^) to stress that these matrices refer to PCA results, and not to CCA or CA results considered previously. $\mathbf{D}_w$ is a diagonal matrix built from a vector of case weights $\mathbf{w}$, and $\mathbf{X}$ is an $n \times p$ data matrix of continuous variables. The standardized principal components are given by $\mathbf{D}_w^{-1/2}\hat{\mathbf{U}}$, and coordinates for variable vectors are $\hat{\mathbf{V}}\hat{\mathbf{D}}$, allowing us to construct a biplot by plotting the first two columns of these matrices. If we take $\mathbf{X}$ to be a matrix of weighted averages, and weight the species by the square root of their total abundance, then we have:

$$\mathbf{D}_r^{-1/2}\mathbf{PZ} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'. \tag{9.68}$$

The similarity of this decomposition with the previously described s.v.d. of CCA is striking (cf. (9.6), page 111). If we assume $\mathbf{Z}$ to be centred on the weighted mean and standardized by dividing by the square root of the weighted variance, then the only difference between (9.68) and (9.6) is the postmultiplication by the square root of the inverse of the weighted correlation matrix $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$ in the latter. Thus, if the environmental variables happen to be uncorrelated in the weighted sense, then $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$ reduces to an identity matrix, and by the uniqueness of the s.v.d. we have that the principal components coincide with rescaled CCA species coordinates, and the variable vectors of both methods are also related by a rescaling.

However, note that the preceding does not mean that we can actually perform CCA by merely stacking a matrix of weighted averages into a PCA program. Standard software for (weighted) PCA will usually centre or standardize the data prior to subsequent analysis. This modifies the matrix of weighted averages, and so we no longer do an s.v.d. of the weighted averages. Note that in this case it is not possible to simply "undo" the centring operation by a linear transformation of the data. "Undoing" the centring operation here implies that there exist a matrix $\mathbf{Q}$ such that $(\mathbf{I}-\mathbf{1r}')\mathbf{Q}=\mathbf{I}$. Premultiplication of the data by $\mathbf{Q}$ would so prevent the centring. However, this is not possible because the centring matrix $(\mathbf{I}-\mathbf{1r}')$ is singular, and its inverse, the desired matrix $\mathbf{Q}$, does not exist. ($(\mathbf{I}-\mathbf{1r}')$ is singular because $\mathbf{r}'(\mathbf{I}-\mathbf{1r}')=\mathbf{0}$, thus a linear combination of the rows of the centring matrix sums to the zero vector, meaning that one row of the centring matrix is a linear combination of the others. More precisely, the rank of the centring matrix is $I-1$ (the rank of an idempotent equals the trace so $\mathrm{tr}(\mathbf{I}-\mathbf{1r}')=\mathrm{tr}(\mathbf{I})-1=I-1$). In order to perform CCA by PCA, when environmental variables are uncorrelated, one needs a PCA routine that does not centre or standardize the data but leaves this to the user.

9.3.3 Reciprocal Averaging

It is well known that CA can performed by using a reciprocal averaging algorithm (Hill, 1974; Greenacre, 1984). The same is also true for CCA (Ter Braak, 1986), and the reciprocal averaging algorithm underlies the CANOCO software (Ter Braak, 1988). For the purpose of illustration and later reference, we describe a simplified version of this algorithm in box 9.1. In short, we start with a I-dimensional vector of random species scores, which are standardized by subtracting the weighted mean and dividing by the square root of the weighted variance, where the weights are abundances of the species at the sites. Site scores are calculated as weighted averages (routine `wa`) of the species scores and vice verse, until the scores no longer change. As a criterion for convergence, we require the sum of squared differences between the scores of two successive iterations to be smaller than some particular value. After taking weighted averages, scores need to be standardized again to prevent a decrease in variance.

After convergence, vector `species` will contain the standard coordinates of the species (the first column of $\mathbf{\Phi}_l$) and vector `sites` will contain the standard coordinates of the sites (the first column of $\mathbf{\Gamma}_l$) as given in (9.7). The algorithm can be extended in order to obtain the second and higher order dimensions by

including extra steps that require coordinates in the higher dimensions to be uncorrelated with previously extracted coordinates. A more general version of the algorithm has been described by Ter Braak and Prentice (1988).

```
1.  species := random(I,1);
2.  standardize(species);
3.  convergence := false;
4.  while not(convergence) do
5.        sites := wa(species);
6.        oldscores := sites;
7.        standardize(sites);
8.        siteslc := fitregr(sites,Z);
9.        species := wa(siteslc);
10.       standardize(species);
11.       convergence := (ssq(sites-oldscores) < 0.0001);
12. end;
```

BOX 9.1: A RECIPROCAL AVERAGING ALGORITHM FOR CCA

In step 8 the site scores are assigned the fitted values of the regression of the site scores onto the environmental variables. This step precisely restricts the standard site scores to be linear combinations of the environmental variables.

Notice that after convergence we have *two sets of site scores*: the ones which are linear combinations of the variables (`siteslc`) and the ones which are not, but are weighted averages of the species scores (`sites`). In the literature these are known as LC site scores and WA site scores respectively (Palmer, 1993; McCune, 1997).

Note also that if step 8 would be made inactive, and the scores passed to routine `wa` in step 9 would be `sites` rather than `siteslc`, this algorithm will converge to the first dimension of the ordinary CA solution.

The approach to CCA in this chapter is exclusively based on the singular value decomposition from which we obtain the LC site scores, but not the WA scores. In our approach, WA site scores (principal coordinates) can be obtained by applying the transition equations from ordinary CA to the standard species scores (cf. (5.11) p. 57). Standard WA sites scores are then calculated by post-multiplying by the square root of the inverse of the inertias of the restricted dimensions.

9.4 Transition Equations

In ordinary CA, species scores and site scores are related to each other by the transition formulae. These formulae express that principal coordinates of species and sites are weighted averages of the standard coordinates of sites and species

respectively (cf. equations (5.10) and (5.11) on page 57).

CCA provides three sets of coordinates in the restricted dimensions, and so in principle we can look for three sets of transition equations: between species and sites, species and variables, and sites and variables. With some algebraic manipulation these can all be derived from previous equations (9.6) and (9.7). First, the transition from species to sites and from sites to species:

$$\mathbf{G}_l = (\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c)(\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{\Phi}_l), \qquad \mathbf{F}_l = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{\Gamma}_l. \tag{9.69}$$

From this we can see that principal site coordinates can be considered weighted averages of standard species coordinates ($\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{\Phi}$), but projected onto the subspace spanned by the environmental variable with the idempotent projector matrix ($\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}_c$). On the other hand, principal coordinates of the species are weighted averages of the standard site coordinates just like in ordinary CA.

Formulae (9.69) can be used for supplementary point calculation. If the abundances of the species at a supplementary site are expressed as a profile, this profile can be substituted for $\mathbf{D}_c^{-1}\mathbf{P}'$ in the first equation, and the principal coordinates of the supplementary site are found. In the same manner, a supplementary profile of a species over the sites can be substituted into the second equation in order to obtain a supplementary species point.

For the sake of completeness, we also give equations relating species and variables, and sites and variables, though these seem not to be very interpretable, at least not as weighted average relationships. The relationship between variables and species can be expressed as:

$$\mathbf{F}_l = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{\Omega}, \qquad \mathbf{H} = \mathbf{Z}'\mathbf{D}_c(\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{\Phi}_l), \tag{9.70}$$

where the latter of the two can be seen as weighted covariances between variables and weighted averages of standard species coordinates. Considering sites and variables, we find the equations:

$$\mathbf{\Omega} = \mathbf{Z}'\mathbf{D}_c\mathbf{\Gamma}_l, \qquad \mathbf{\Gamma}_l = \mathbf{Z}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}\mathbf{\Omega}. \tag{9.71}$$

9.5 An Example with Artificial Data

We have created an artificial data set of five species and five sites, with three environmental variables in order to illustrate some of the previously exposed theory. First we will do an ordinary CA of the abundance data, followed by a CCA of the abundance data with two of the environmental variables. Since the representation of the weighted averages of the species in a 2-dimensional biplot is perfect when one has only two environmental variables (cf. equation (9.34)), we continue to analyse the same table with an extra third environmental variable. This allows us to investigate errors that are obtained when projecting site points and species points onto environmental variable vectors.
The data is shown in table 9.2, and is represented in different forms. The first set of rows gives the raw data, the counts for the five species (Sp_1 through Sp_5)

	Sp_1	Sp_2	Sp_3	Sp_4	Sp_5	V_1	V_2	V_3
A	4	5	2	1	0	5	2	1
B	6	2	2	4	0	7	1	6
C	8	1	2	8	4	10	3	12
D	10	0	2	4	6	12	0	6
E	12	0	2	1	8	14	1	1
A	0.1000	0.6250	0.2000	0.0556	0.0000	-1.7627	0.5647	-1.0869
B	0.1500	0.2500	0.2000	0.2222	0.0000	-1.1052	-0.3505	0.0929
C	0.2000	0.1250	0.2000	0.4444	0.2222	-0.1189	1.4800	1.5086
D	0.2500	0.0000	0.2000	0.2222	0.3333	0.5386	-1.2658	0.0929
E	0.3000	0.0000	0.2000	0.0556	0.4444	1.1961	-0.3505	-1.0869
V_1	10.7500	6.1250	9.6000	9.7222	12.4444	10.3617	-	-
V_2	1.2500	1.8750	1.4000	1.7222	1.1111	-	1.3830	-
V_3	5.2000	3.6250	5.2000	8.1111	5.1111	-	-	5.6064
V_1	0.1277	-1.3929	-0.2504	-0.2102	0.6847	1.0000	-0.3746	-0.1210
V_2	-0.1217	0.4503	0.0156	0.3105	-0.2488	-0.3746	1.0000	0.5288
V_3	-0.0959	-0.4675	-0.0959	0.5910	-0.1169	-0.1210	0.5288	1.0000

TABLE 9.2: ARTIFICIAL DATA SET, ABUNDANCES AND ENVIRONMENTAL VARIABLES

at the five sites (A,B,C,D,E), and the raw measurements for the environmental variables V_1, V_2 and V_3. The second set of rows lists the row profiles of the species and the standardized environmental variables. The third row block gives the weighted averages of the species with respect to the raw environmental data followed by the weighted averages of the variables. The last set of rows gives the weighted averages with respect to the standardized environmental variables, and the last diagonal block of the table gives the weighted correlation coefficients between the variables. We see that V_1 is increasing over the 5 sites, whereas V_2 shows no clear gradient. Sp_1 seems to respond to V_1, having higher abundances for stations where V_1 is high. Sp_2 prefers the lower values of V_1, whereas Sp_5 prefers the higher ones. Sp_3 is indifferent with respect to the environmental variables, and Sp_4 shows a unimodal response.

9.5.1 CA of the Artificial Data

The CA asymmetric map of the row profiles, as presented in figure 9.1 is made without using environmental data, but in its interpretation, external environmental information can be used, according to the methodology developed in chapter 7. Here the first principal axis separates the stations in order of increasing concentration of V_1 (from right to left). It also separates the species 1 and 5, which are high on V_1, from 2, which is low on V_1. The first principal axis could therefore be labelled as decreasing concentration of V_1. This is also justified by the high negative correlation (-0.96) between the first axis and V_1. Note that the configuration of the site points takes the form of a horseshoe.
The inertia decomposition of CA (table 9.3) shows that we capture 97.8% of the inertia in a two dimensional plane, which means that we have a map of high quality. This is confirmed by inspection of the detailed CA statistics (not shown). Except for species 1 and site B, all points have a quality of over 0.9. A special algorithm has been developed in order to automatically calibrate the oblique vertex vectors in the asymmetric map. Standard software for correspondence analysis does not allow such automated calibration. The

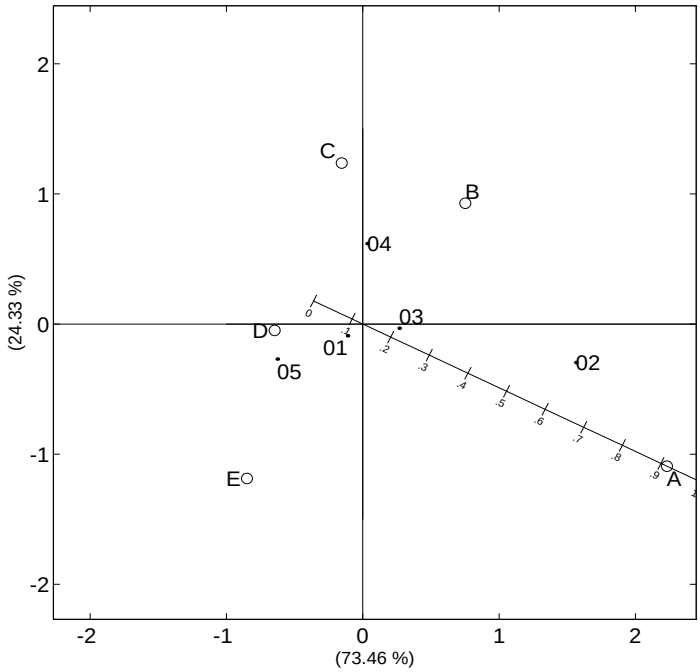

FIGURE 9.1: CA ASYMMETRIC MAP FOR ABUNDANCE DATA

calibration is shown in figure 9.1, only for site A, in order to keep the graphic readable. The exact proportions of the species Sp_1 through Sp_5 with respect to the site A, based on the original data, are [0.100 0.625 0.200 0.056 0.000] respectively. In figure 9.1 these numbers are very well approximated by projecting Sp_1 through Sp_5 onto the calibrated site vector for site A. This gives us the values [0.109 0.614 0.208 0.050 -0.014]. If we calculate these projections exactly, with respect to all the site vectors, we find small errors which can be assembled into an inertia component of 0.0089. This precisely equals the inertia of the third and fourth dimension which is not represented in the display.

9.5.2 CCA of the Artificial Data

We now present the results of a CCA of the same data described in table 9.2, using only the first two variables, V_1 and V_2. We begin by considering the inertia decomposition of CCA given in table 9.4.
Notice that the total inertia of the abundance matrix in the 2D CCA map (72.6%) is less than the total inertia in the CA map. We have lost about 25.2% of the total inertia by restricting the solution. This illustrates that the linearly constrained optimal plain will always capture less inertia than the optimal CA plane (or at best as much). Another noteworthy point is that the inertia in the third dimension is much higher than the inertia in the second dimension. In CA the amount of inertia always decreases when we look at the next dimension.

Dim.	Inertia	%	Cum. %
1	0.2964	73.46	73.46
2	0.0982	24.33	97.79
3	0.0089	2.20	99.99
4	0.0000	0.01	100.00
Total	0.4035	100	

TABLE 9.3: INERTIA DECOMPOSITION CA

Dim.	Inertia	%	Cum. %
1	0.2770	68.65	68.65
2	0.0159	3.95	72.60
3	0.1058	26.22	98.81
4	0.0048	1.19	100.00
Total	0.4035	100.00	

TABLE 9.4: INERTIA DECOMPOSITION CCA

This also holds in CCA when restricted dimensions and unrestricted dimensions are considered separately.

Panel A in figure 9.2 shows the biplot of a CCA for the data in table 9.2. The origin of the map represents the weighted averages of the variables. At the same time the origin also represents the average species profile. We see that variable V_1 virtually coincides with the horizontal axis, this axis is could be labelled V_1, whereas the vertical axis has high negative correlation with V_2 (-0.92). Note that both variables vectors have the head of their arrows on the unit circle, which means that all their weighted variance is accounted for by the display.

In panel B of figure 9.2 we show the same CCA output, but now we have calibrated the variable vectors. Increments of half a unit have been marked off on both vectors. When we project species points onto the two environmental variables, the weighted averages of the species are recovered perfectly, there is no error (cf. fourth set of rows of table 9.2). When we project site points onto environmental vectors we see that there is neither error in the values of the environmental variables we recover (cf. second diagonal block of table 9.2). Because we standardized the environmental data, the values we recover are weighted averages of the species with respect to standardized environmental variables, and when we project site points, we recover standardized values for the environmental variables. In panel B, due to the standardization, one unit on vector V_1 is the same as one unit on vector V_2. If we prefer to recover our original data, then it is perfectly possible to change the calibration of the variable vectors in order to do so. It that case, the origin represents the weighted averages of the variables in their original units (10.36 and 1.38 for V_1 and V_2 respectively), and the calibration of the variable vectors is changed. This is shown in panel C, where we can now recover the raw environmental data values of table 9.2 (first set of rows, second set of columns) and the weighted averages in the original

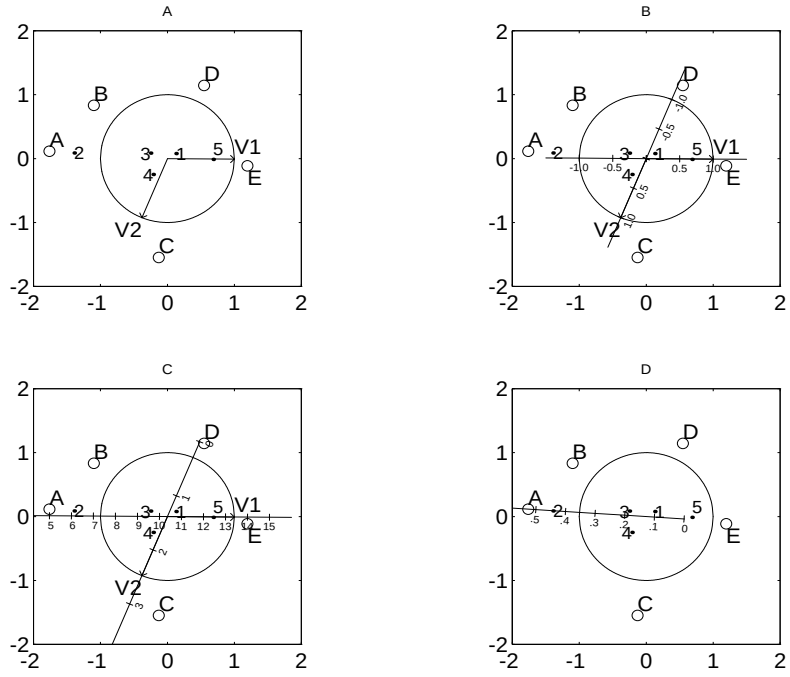

FIGURE 9.2: CCA BIPLOTS

scale (third set of rows) without error.

In the same map, we can also calibrate the site vectors. This calibration is shown in panel D of figure 9.2 for site A. One can approximately recover the profiles of the species by projecting them onto the site vectors. In this case there are errors. For instance, Sp_2 projects onto a value of about 0.45 for site A whereas its true value is 0.625. The most erroneous interpretation in the map seems to be that of Sp_2 with respect to site E. Table 9.5 lists the profiles based on the data and the profiles recovered from the map.
These errors were calculated by actually carrying out all possible projections in the map. The errors can be assembled into a component of inertia with the formula:

$$\sum_{i=1}^{I} r_i \sum_{j=1}^{J} e_{ij}^2 / c_j \qquad \text{or} \qquad \text{tr}(\mathbf{D}_r \mathbf{E} \mathbf{D}_c^{-1} \mathbf{E}'), \tag{9.72}$$

Where $\mathbf{E}$ is a $I \times J$ matrix of errors. For the errors in table 9.5 this gives an inertia of 0.1106, which is exactly the quantity of inertia outside the 2 dimensional plane (cf. table 9.4).

The site coordinates we plotted so far have always been the ones that are linear combinations of the environmental variables ($\mathbf{\Gamma}_l$), and are called LC scores.

		A	B	C	D	E
Data	Sp_1	0.1000	0.1500	0.2000	0.2500	0.3000
	Sp_2	0.6250	0.2500	0.1250	0.0000	0.0000
	Sp_3	0.2000	0.2000	0.2000	0.2000	0.2000
	Sp_4	0.0556	0.2222	0.4444	0.2222	0.0556
	Sp_5	0.0000	0.0000	0.2222	0.3333	0.4444
Map	Sp_1	0.1000	0.1377	0.2108	0.2715	0.2800
	Sp_2	0.4421	0.3877	0.2560	0.0792	-0.1649
	Sp_3	0.1851	0.2005	0.2200	0.2251	0.1692
	Sp_4	0.1717	0.1528	0.3455	0.1404	0.1896
	Sp_5	-0.0265	0.0352	0.2279	0.3181	0.4453

TABLE 9.5: SPECIES PROFILES OF ORIGINAL DATA AND RECOVERED FROM THE MAP

This seems a natural choice, since we are doing a restricted analysis. However, the singular value decomposition at the heart of the method is decomposing a species by variables matrix (cf. equation (9.6)). This is the matrix whose display is optimized, and the site points are in fact added in a second step, once they are calculated as the weights **B** are known (cf. (9.8),(9.9)) We might also consider to plot site points that are not linear combinations (the WA scores) if this is convenient for some reason, as it will not affect the optimal display of the species by variables matrix. The confusion about the type of site points to use has also been noticed by Palmer (1993) and McCune (1997). We thus extend our work with the artificial data set considering also site coordinates that are not linear combinations.

In graph 9.3. we present again the same CCA biplot, but now both sets of site coordinates are plotted. The WA site scores are indicated in lower case and with a cross (**x**), and are connected to their corresponding LC scores (open circles) by a dotted line. We see that the projections of these site coordinates onto the environmental variables is no longer free of error. With respect to variable V_1 the use of these WA coordinates seems not too bad, the ordering of the sites being correct. With respect to the second variable however, the order is destroyed and large errors are found in the projections.

We can also project the species points onto the site vectors pointing to the WA scores. These vectors can also be calibrated, and approximations of the profiles of the species can be obtained just as we did before. And with use of formula (9.72) these errors can again be compiled into an inertia component, which takes a value of 0.0308 for the artificial data under consideration. For this data set the abundance matrix thus has a better representation when we use WA scores.

We can of course not generalize about this beyond the particular data set we have analyzed. However, the simulation results described by McCune (1997) point in the same direction: the species data are better displayed when we use WA site coordinates rather than LC site coordinates.

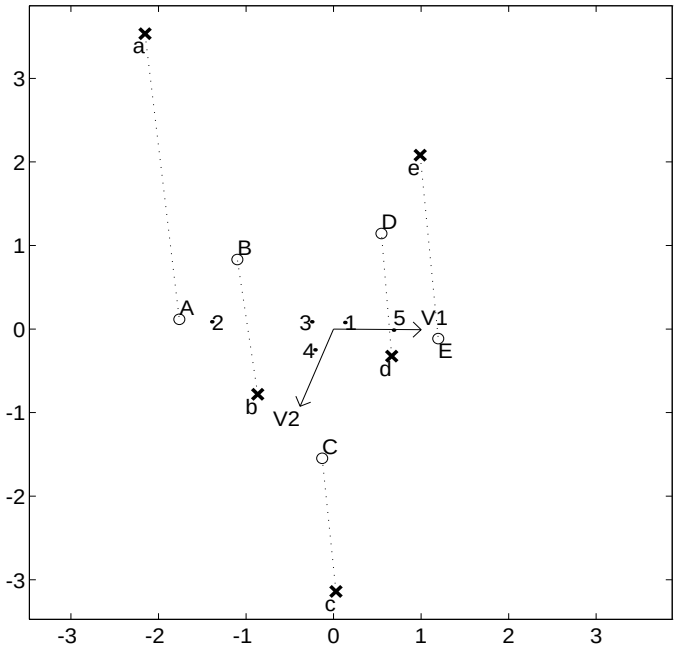

FIGURE 9.3: CCA BIPLOT WITH LC AND WA SITE SCORES

9.5.3 CCA with Three Variables

We continue with a CCA for the same data table 9.2, but extended with the third variable, V_3. With three variables the display of the weighted averages of the species will no longer be perfect in a 2-dimensional map, thus allowing us to investigate errors produced when projecting species and site points onto the variable vectors. The new decomposition of inertia is shown in the first five columns of table 9.6.

Note that if we take the third variable into account, we can now represent 95.5% of the inertia in 2 dimensions, much more than with only 2 variables, but of course, still a bit less than the ordinary CA (97.8%). The CCA biplot of the data is shown in figure 9.4.

The graph shows that variables V_1 and V_3 have an excellent display in the map, but that V_2 has a worse representation. With 3 variables, there are now slight differences in the weighted averages of the species obtained from the map and from the real data. Both are listed in table 9.7, measured in standard deviations. The weighted variance of the weighted averages is 0.4239, whereas the weighted variance represented in the map is 0.4216, making that the quality of the display of the weighted averages of the species is 0.9947. We could also have calculated this by working out the quotient of the weighted sums of inertias and squared correlations in equation (9.34). The quality of the representation of the weighted averages for solutions of different dimensionality is given in the seventh column

Dim	Inertia	%	cpi[a]	cpri[b]	cvwa[c]	cpvwa[d]	cvev[e]	cpvev[f]
1	0.2936	72.76	72.76	75.58	0.3087	72.84	1.0516	35.05
2	0.0917	22.72	95.48	99.19	0.4216	99.47	2.2829	76.10
3	0.0032	0.78	96.27	100.00	0.4239	100.00	3.0000	100.00
4	0.0151	3.73	100.00	103.88	-	-	-	-
Total	0.4035	100.00						

[a]Cumulative percentage of inertia
[b]Cumulative percentage of inertia w.r.t. total inertia in constrained dimensions
[c]Cumulative explained weighted variance of weighted averages
[d]Cumulative percentage of explained weighted variance of weighted averages
[e]Cumulative explained weighted variance of environmental variables
[f]Cumulative percentage of explained weighted variance of environmental variables

TABLE 9.6: INERTIA DECOMPOSITION OF CCA WITH 3 VARIABLES

of table 9.6. The weighted averages of the species have an excellent representation in figure 9.4, and species preferences can be inferred from the map with confidence. We also want to contrast this with the fraction of the sum of the first two eigenvalues with respect to the total: 0.9548. This example thus also illustrates our point that the *eigenvalues do not indicate fractions of weighted variance in the weighted averages, but correspond to fractions of the total inertia of the abundance matrix.* (compare columns 4,5 and 7 in table 9.6)
The errors in the weighted averages can also be assembled into an inertia component with the formula:

$$\mathrm{tr}(\mathbf{D}_r\mathbf{E}(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1}{}'\mathbf{E}'). \tag{9.73}$$

This gives us a lost inertia of 0.0032, exactly the amount of inertia of the third

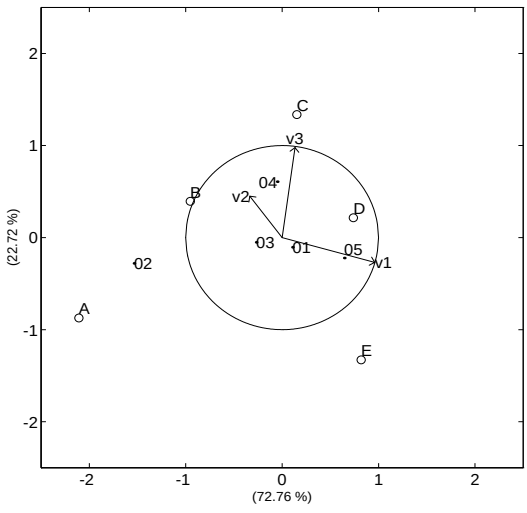

FIGURE 9.4: CCA MAP FOR ABUNDANCE DATA WITH 3 VARIABLE VECTORS

	Data			Map		
Species	V_1	V_2	V_3	V_1	V_2	V_3
1	0.1277	-0.1217	-0.0959	0.1315	-0.0835	-0.0884
2	-1.3929	0.4503	-0.4675	-1.3989	0.3901	-0.4792
3	-0.2504	0.0156	-0.0959	-0.2452	0.0678	-0.0857
4	-0.2102	0.3105	0.5910	-0.2122	0.2908	0.5871
5	0.6847	-0.2488	-0.1169	0.6779	-0.3162	-0.1300

TABLE 9.7: WEIGHTED AVERAGES OF SPECIES (ORIGINAL DATA AND RECOVERED FROM THE MAP)

dimension. In general, the formula above will give us the inertia in the remaining restricted dimensions (the dimensions outside the 2D plane that are restricted).

We evaluate projections of the site coordinates onto the environmental variable vectors. Environmental values for the sites and projections obtained from the map are shown in table 9.8, expressed in standardized units.

Site	V_1	V_2	V_3	V_1	V_2	V_3
A	-1.7627	0.5647	-1.0869	-1.7879	0.3150	-1.1355
B	-1.1052	-0.3505	0.0929	-1.0198	0.4976	0.2581
C	-0.1189	1.4800	1.5086	-0.2125	0.5507	1.3275
D	0.5386	-1.2658	0.0929	0.6509	-0.1510	0.3101
E	1.1961	-0.3505	-1.0869	1.1434	-0.8736	-1.1888

TABLE 9.8: ENVIRONMENTAL VALUES FOR THE SITES (ORIGINAL DATA AND RECOVERED FROM THE MAP)

We can express how well the environmental data in $\mathbf{Z}$ is represented in the display by calculating the fraction of weighted variance of the environmental variables accounted for. These fractions are shown as the last column in table 9.6 and shown that the map in figure 9.4 still captures 76.1% of the weighted variance of the environmental variables.

9.5.4 CCA with Three Principal Components

Finally, we want to illustrate our assertion that the eigenvalues of a CCA *do* indicate fractions of weighted variance *when environmental variables are uncorrelated.* We therefore repeat the analysis above, where we replace the three variables considered by the first three principal components obtained from a weighted principal component analysis of the environmental data. The graph of this analysis is shown in figure 9.5.
Note that the first two principal components show up as two nearly orthogonal directions in the display. The third component has a much shorter vector, as it is uncorrelated with the previous two and can no longer be correctly represented

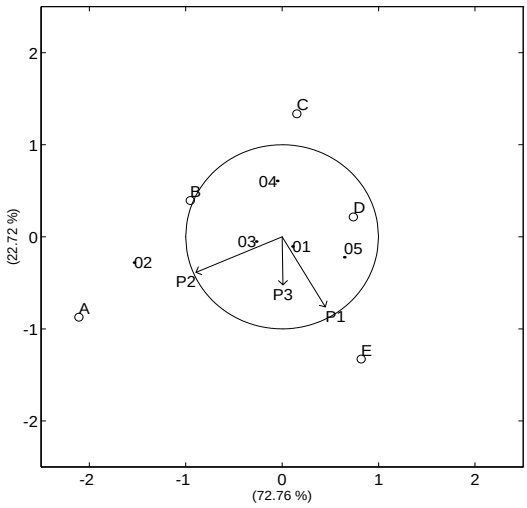

FIGURE 9.5: CCA USING PRINCIPAL COMPONENTS

in 2D. Inertia decomposition and fractions of explained variance are shown in table 9.9. This table shows that the inertia decomposition obtained is exactly the same one as when the original variables were used rather then principal components (cf. table 9.6). Principal components are linear combinations of the original variables, and can be obtained from the original data by postmultiplication with the score coefficient matrix. In section 9.2.8 CCA was shown to be invariant under linear transformation of the environmental data, and the fact that the inertia decomposition is the same when we use principal components merely illustrates this.

Dim	Inertia	%	cpi	cpri	cvwa	cpvwa	cvev	cpvev
1	0.2936	72.76	72.76	75.58	0.2936	75.58	1.0000	33.33
2	0.0917	22.72	95.48	99.19	0.3852	99.19	2.0000	66.67
3	0.0032	0.78	96.27	100.00	0.3884	100.00	3.0000	100.00
4	0.0151	3.73	100.00	103.88	-	-	-	-
Total	0.4035	100.00						

TABLE 9.9: INERTIA DECOMPOSITION OF CCA WITH 3 PRINCIPAL COMPONENTS

Cumulative fractions of inertia are expressed in two ways in table 9.9 (as well as in table 9.6). The fourth column gives the cumulative percentages of inertia explained, with respect to the total inertia in the abundance matrix, 0.4035. The fifth column gives the same information, but with respect to the total inertia in the restricted dimensions only (0.3884). We see that *the quality of the species data, if environmental variables are uncorrelated, and if expressed with*

respect to the inertia in the restricted dimensions, is the same as the amount of weighted variance explained of the weighted averages.
We also note that the fractions of weighted variance of $\mathbf{Z}$ accounted for by a one, two or three dimensional solution are 1/3, 2/3 and 1 respectively, which is in precise agreement with what we predict from theory (cf. (9.36), p. 119).

9.6 Conclusions

In this chapter we have gone through the theory of canonical correspondence analysis. In this section we briefly summarize the main results that are, to our knowledge not or not correctly described in the literature. First, we noted that a CCA has in fact $J-1$ dimensions in its solution, whereas most authors hold this to be Q, the number of environmental variables. Next, CCA is held to optimally represent species optima, these optima being estimated by the weighted averages of the species with respect to the environmental variables. We have shown that this is strictly speaking not the case due to the postmultiplication of the weighted averages by the square root of the inverse of the correlation matrix of environmental variables, an operation that renders CCA scale invariant. It is well known that when we have more variables than samples, the CCA solution equals the CA solution. This chapter has provided an analytical proof of this observation. We precisely stated the conditions under which the CCA solution includes a trivial dimension, and what we need to do omit this dimension. Principal inertias in CA are in the interval [0,1], and have provided a proof that the same holds in CCA. We have derived quality statistics that indicate the quality of all three matrices displayed in CCA, the abundance matrix, the environmental data matrix and the matrix of weighted averages. This chapter also shows that CCA can be seen as principal coordinates analysis of a matrix of weighted Mahalanobis distances between species optima.

Chapter 10

Applications of Canonical Correspondence Analysis

10.1 Introduction

This chapter deals with some applications of CCA to the Ekofisk database. Prior to analysis, we applied the square root transformation to the species data and the log transformation to the chemical data. This has the advantage that the influence of highly abundant species like *Myriochele oculata* and *Chaetozone setosa* is somewhat reduced, and that the influence of chemically aberrant stations will also be diminished. Many applications of CCA are can be found in the ecological literature, see for instance (Ter Braak, 1986; Johnson and Altman, 1999; Ter Braak, 1994).

10.2 CCA of 1992

Figure 10.1 shows the solution of a CCA of the data from 1992 (148 species, 11 stations and 9 variables). Station 40 has been eliminated. If station 40 is included, the first axis opposes this station to all others, with all contaminants pointing away from station 40. Station 40 is thus very different from the rest, as it is not contaminated, and shows up as an outlier in the analysis. Its deletion allows us to perceive more details about the contaminated stations. We can infer from the map that *Chaetozone setosa* is a species preferring contaminated conditions, with high concentrations of heavy metals, Barium and organic components, whereas species like *Amphiura filiformis*, *Timoclea ovata*, *Trichobranchus sp.* and *Nephtys hombergi* prefer less contaminated conditions. Though the projection of site points with respect to the environmental variables is not explicitly optimized in CCA, the analysis suggests stations 15,14 and 13 to be the most contaminated ones. This is a group of stations relatively close to the southern side of the platform (cf. figure 2.3 p. 8). On the other hand, we find the more remote stations 8, 12 and 18 in the upper right of the display, sug-

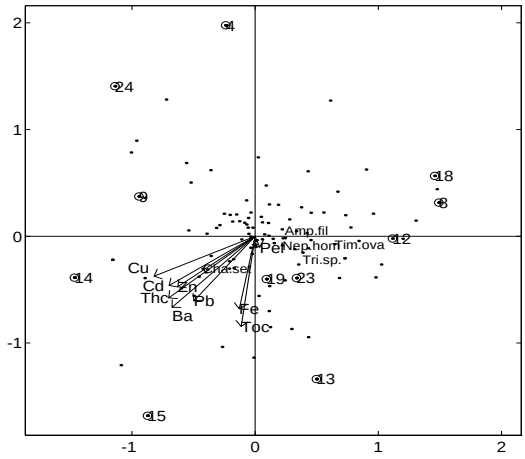

FIGURE 10.1: CCA OF 1992, ALL SPECIES.

gesting these stations are the lowest on the contaminants. The vectors for the pollutants thus seem to coincide with what can be seen as a distance-direction in the biplot.

Dim	Inertia[a]	%	V(WA)[b]	% V(WA)	V(**Z**)[c]	% V(**Z**)
1	0.1255	16.9	0.3479	40.8	2.7723	30.8
2	0.0952	29.6	0.6241	73.2	5.6746	63.1
3	0.0900	41.7	0.6454	75.7	5.9113	65.7
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total	0.7445	100.0	0.8521	100.0	9.0000	100.0

[a]Inertia of the matrix **N**, not cumulative
[b]Cumulative weighted variance of $\mathbf{D}_r^{-1}\mathbf{PZ}$
[c]Cumulative weighted variance of **Z**

TABLE 10.1: INERTIA AND VARIANCE DECOMPOSITION FOR CCA, 1992

The basic numerical results of the analysis are shown in table 10.1. The two-dimensional diagram is seen to capture 30% of the inertia of the abundance matrix, 73% of the weighted variance of the weighted averages (column V(WA)) of the species, and 63% of the weighted variance of the environmental variables (column V(**Z**)). The total inertia, 0.7445, can be partitioned into a restricted part of 0.6794 (91.3%) and an unconstrained part, 0.0650 (8.7%).

10.2.1 Reducing the Number of Species

Because there are so many species, it is impossible to show them all with their names in a biplot. In figure 10.1 we only labelled the ones that have a consid-

erable fraction of their inertia accounted for by the display (> 0.6) and have a total abundance larger than 20. Selecting only well-displayed species still produces a crowded display with a large amount of relatively well-displayed rare species, who merely happen to be close to the optimal plane. Alternatively, one could delete the rarer species (e.g. all species with a total abundance of less than 10), who usually have little influence in the analysis anyway. This has the disadvantage that we ignore some of the (expensive) biological information. We must however, use some rule to reduce the amount of species, simply because it is impossible to label 148 species and 11 stations in one plot. We report the results of another CCA, where we used only the 50 most abundant species. This means that we deleted all species with a total abundance of 13 or lower. Some of the station points become outliers, making it difficult to jointly plot stations and species.

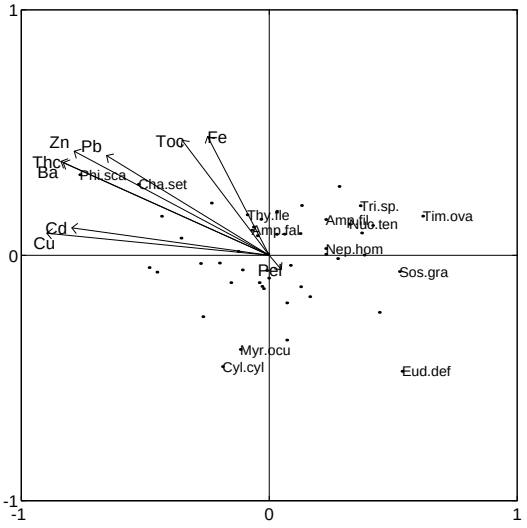

FIGURE 10.2: CCA OF 1992, 50 MOST ABUNDANT SPECIES

Figure 10.2 shows the graphical output of the CCA, zoomed in so that we can detect more details of the species. The labelled species have more than 50% of their inertia accounted for by two-dimensional map. The analysis so obtained is a bit more informative with respect to species preferences. In particular, we see that *Philine scabra* is another species with preference for highly contaminated conditions. *Ampharete falcata* and *Thyasira flexuosa* have relatively high optima for TOC and Fe. Species like *Eudorellopsis deformis* and *Sosane gracilis* seem to prefer, as most of the better represented species, less contaminated conditions. The species data used in this analysis has a total inertia of 0.2217, of which 0.2041 (92.1%) is in the restricted dimensions and 0.0176 (7.9%) is in the unrestricted dimensions. The two-dimensional biplot shown in figure 10.2 captures 48.4 percent of the inertia of the abundance matrix, 83.2% of the variance of the weighted averages, and 57.6% of the variance of the environmental variables.

Dim	Inertia	%	V(WA)	% V(WA)	V(**Z**)	% V(**Z**)
1	0.0678	30.6	0.2761	71.8	4.0724	45.2
2	0.0395	48.4	0.3200	83.2	5.1842	57.6
3	0.0263	60.2	0.3401	88.5	5.9478	66.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total	0.2217	100	0.3844	100	9.0000	100

TABLE 10.2: INERTIA AND VARIANCE DECOMPOSITION FOR CCA OF 50 MOST ABUNDANT SPECIES, 1992

CCA statistics of the analysis are given in table 10.2.

10.2.2 Partialling out Spatial Effects

In the first analysis of this chapter, (cf. figure 10.1), we commented on the presence of a distance direction in the biplot. In the analysis with the reduced set of species, this effect is still present. In order to quantify this, we report that in the latter analysis the first CCA axis has a high correlation of 0.81 with distance, whereas the second CCA axis has correlation of 0.28 with distance.

In a successive analysis, we try to partial out these spatial effects. This is done by regressing all pollutants onto the Euclidean distance from the platform, the east-west distance from the platform and the north-south distance from the platform. Next, CCA is performed using the residuals of this simultaneous multiple regression as environmental variables. These residuals are uncorrelated with the three distance variables, and allow us to perform an analysis where the distance-effects have been removed. The biplot of this analysis is shown

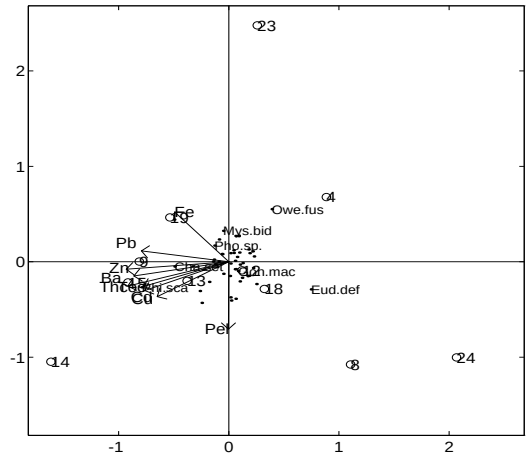

FIGURE 10.3: CCA OF 1992, DISTANCES PARTIALLED OUT

in figure 10.3. The quality of the display of the different data matrices has somewhat changed, compared to the previous analysis (see table 10.3). The species optima and the environmental variables are now better represented, whereas the display of the abundance matrix is worse. Variable Pel has gained considerably in quality of representation, and coincides with the second axis. The horizontal axis of the display has become more closely associated to organic and heavy metal pollution. Two of the more abundant species of the survey, *Mysella bidentata* and *Phoronis sp.* pop up in the analysis as species having high optima for Fe. The qualities of representation of the different matrices are given in table 10.3. The two-dimensional biplot of figure 10.3 captures 31.0% of the inertia of the abundance matrix, 85.4% of the variance of the weighted averages and 67.6% of the variance of the environmental data. Of the total amount of inertia, 0.2217, 0.1289 is in the restricted dimensions, and 0.0928 in the remaining dimensions.

Dim	Inertia	%	V(WA)	% V(WA)	V(**Z**)	% V(**Z**)
1	0.0427	19.3	0.2099	74.7	4.9145	54.6
2	0.0259	31.0	0.2401	85.4	6.0796	67.6
3	0.0173	38.8	0.2657	94.5	7.5616	84.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total	0.2217	100	0.2811	100	9.0000	100

TABLE 10.3: INERTIA AND VARIANCE DECOMPOSITION FOR CCA, DISTANCES PARTIALLED OUT

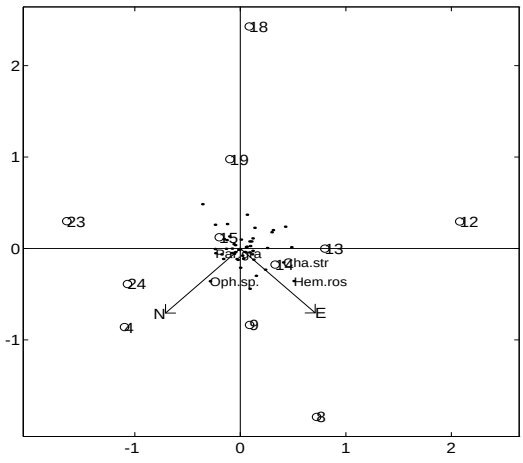

FIGURE 10.4: CCA OF 1992, USING N/S AND E/W DISTANCE ONLY

Alternatively, one can also do a CCA using only the spatial information. In particular, if we do a CCA using only East-West and North-South distance, we

will recover the station grid exactly, just because we have only two variables. With two variables, the display of the environmental data (the distances from the platform in this case) in a two-dimensional biplot is perfect (cf. section 9.2.5). This is shown in figure 10.4. If we rotate and reflect figure 10.4 in the right way, the station grid is the same as the geographical map in figure 2.3 on page 8, up to a stretching factor.

In this analysis, we see that most species cluster in the center of the display. A species like *Hemilamprops rosea* is seen to be more prevalent in the east of the field. Species who form part of the cluster in the center could be interpreted as being species who like pollution, since the origin now represents the platform. For instance, a species like *Chaetozone setosa*, known to prefer contaminated conditions on the basis of prior analysis, can be found here. On the other hand, according to Ter Braak (1987, pp. 74), species not who do not respond to any of the measured environmental variables also often end up in the center of the display.

10.2.3 Reducing the Number of Variables

If we would perform ordinary CA on the 1992 data, and add the variables as supplementary vectors, (cf. chapter 7) then we obtain an ordination that is very similar to the one in section 10.2.1. This analysis is shown in figure 10.5.

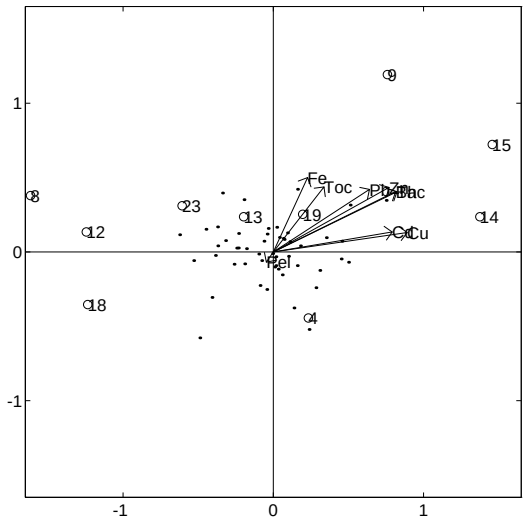

FIGURE 10.5: CA WITH SUPPLEMENTARY VARIABLES, 1992

Station 24 is not shown in figure 10.5, as it is very outlying. The ordination is seen to be similar to a reflection in the vertical axis of the one in section 10.2.1, figure 10.2. In the current circumstances, there is only a small difference between CCA and indirect gradient analysis. This is because the number of variables

(9) is large in comparison with the number of sites (11, after deletion of station 40), and the analysis is not very restrictive. A more restrictive analysis could be carried out if we would drop some environmental variables. This brings along a problem of variable selection, and we have no a priori reasons to keep or drop particular variables. Nearly all environmental variables are closely correlated. Rather then dropping one or more variables, we might as well try to reduce the amount of variables by a PCA, before doing a CCA. In this case, we perform a PCA of all heavy metals, in the hope that we can reduce these five variables to one or two heavy metal components. This turns out to work pretty well. A PCA of all the heavy metals gives a first principal component that explains 80.7% of the variance of the heavy metals, and that can be used to replace the heavy metal variables.
The biplot of this analysis is shown in figure 10.6. The biplot shows that the

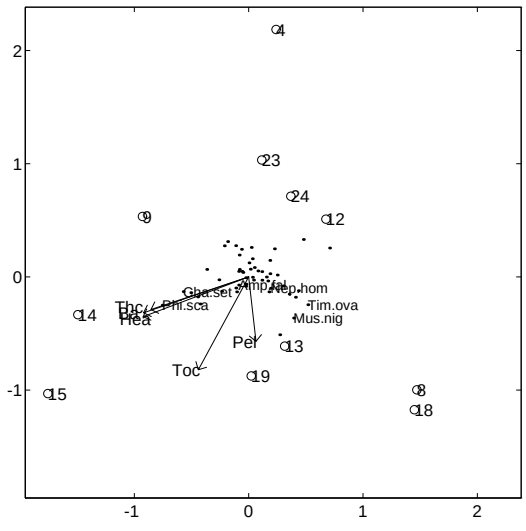

FIGURE 10.6: CCA OF 1992, WITH FEWER VARIABLES

heavy metal component (Hea) is highly correlated with Ba and THC. Species *Chaetozone setosa* and *Philine scabra* show again up as preferring a contaminated environment. *Timoclea ovata* and *Musculus niger* prefer less than average contamination, but have relatively higher optima for silt (Pel). The CCA statistics for this analysis are shown in table 10.4. The two-dimensional biplot in figure 10.6 explains 37.3% of the inertia of the species abundances, as much as 92.4% of the variance of the weighted averages, and 88.4% of the variance of the environmental data. Species preferences are now better displayed than in any previous analysis.
Of the total inertia of the species abundances, 0.2217, an amount of 0.1224 (55.5%) is in the restricted dimensions, whereas a component of 0.0993 (44.8%) is outside the restricted dimensions.

Dim	Inertia	%	V(WA)	% V(WA)	V(**Z**)	% V(**Z**)
1	0.0622	28.1	0.1634	79.4	2.6253	52.5
2	0.0206	37.3	0.1902	92.4	3.9294	78.6
3	0.0160	44.5	0.1980	96.2	4.4177	88.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total	0.2217	100	0.2057	100	5.0000	100

TABLE 10.4: INERTIA AND VARIANCE DECOMPOSITION FOR CCA AFTER DATA REDUCTION

The CANOCO program from Ter Braak (1998) provides facilities for ranking environmental variables in order of importance. We do not review the theory of these facilities, but merely give the results from their application. The CANOCO program provides a forward selection routine for the inclusion of environmental variables in the analysis. When all environmental variables are included in the CCA, the amount of inertia in the restricted dimensions can be calculated. The same can be done for a series of separate CCA's, each with one variable only. The environmental variable giving the highest amount of inertia is thought to be the most important one. In a next step, one can calculate the increase in inertia obtained when another variable is included in the analysis, given that the most important one is already included, and so on. This is the basis of the forward selection algorithm. The CANOCO program provides a Monte Carlo permutation test to assess the statistical significance of the variables.

When we use the 1992 data set of the 50 most abundant species, Cu and Zn are the only significant variables, and contribute most to the total amount of restricted inertia. However, there is a large group of variables that, when used as sole environmental variables, give a virtually equal total amount of (restricted) inertia. For instance, one-variable CCA's with Cd, Zn, Ba, THC or Cu give amounts of inertia (in the restricted dimension) of .05, .05, .06, .06, and .06 respectively. The choice of Cu as the "most important" environmental variable is thus rather arbitrary, it might as well be Ba or THC. For these data, the forward selection procedure is not conclusive, as there are several variables with a similar contribution.

10.3 Conclusions

In the different analyses performed, we have seen that CCA helps us to discover the preferences of the species, though usually only a few species are represented with good quality. Some rule for reducing the number of species is necessary, simply because we have to many species to be able to depict them in a single diagram. In a previous chapter we noted that several species have a monotone decreasing pattern along the environmental variables. Though not considered here, in such circumstances, we could also use reduncancy analysis for analyzing the data.

Chapter 11

An Alternative for Canonical Correspondence Analysis

11.1 Introduction

We noted in chapter 9 that CCA does, strictly speaking, not optimize the display of species optima, where the latter are estimated by a matrix of weighted averages. The singular value decomposition (9.6) on page 111 shows that the matrix of weighted averages is postmultiplied by the inverse of a variance-covariance matrix, and that CCA optimizes the display of the product of these two. The estimation of the species optima is an important aspect in ecological research. For this reason, we dedicate this chapter to an optimal display of the weighted averages, in an attempt to graphically depict these weighted averages as best as possible. In the next section, we develop the algebra for this, and in a later section, we give an application.

11.2 Optimal Display of Weighted Averages

We can do a low rank approximation to the matrix of weighted averages, where we maintain the weighting of the species by the square root of their total abundance:

$$\mathbf{T} = \mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}) = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'. \tag{11.1}$$

This is just the singular value decomposition of CCA, where the postmultiplication by $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$ has been left out. We use hats $\hat{}$ to distinguish the results of this type of analysis, maintaining the same notation ($\hat{\mathbf{F}}$ for species, $\hat{\mathbf{W}}$ for variables, $\hat{\mathbf{\Gamma}}$ for sites). The postmultiplication by $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$ has been justified by noting that it renders CCA scale-invariant with respect to scalar multiplication (Ter Braak, 1986, appendix). It does not matter whether environmental

variables are measured in milligrams or grams per kilo, the matrix decomposed by CCA is the same. However, in practice environmental variables are nearly always standardized. As a consequence, the analysis *is* already scale-invariant. Whether a variable is expressed in milligrams per kilo or grams per kilo, the standardized values of that variable will be the same. If data are always standardized, then there is no need that the statistical method we use takes special precautions. In other words, the postmultiplication by $(\mathbf{Z}'\mathbf{D}_c\mathbf{Z})^{-1/2}$ becomes unnecessary.

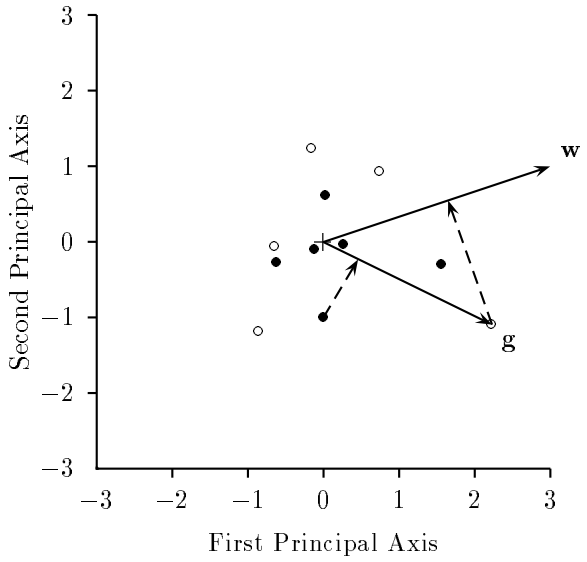

FIGURE 11.1: A PCA BIPLOT WITH A SUPPLEMENTARY SITE POINT

The singular value decomposition proposed in (11.1) is easily seen to be equivalent to the spectral decomposition:

$$\mathbf{T}'\mathbf{T} = (\mathbf{Z}'\mathbf{P}'\mathbf{D}_r^{-1})\mathbf{D}_r(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}) = \hat{\mathbf{V}}\hat{\mathbf{D}}^2\hat{\mathbf{V}}'. \tag{11.2}$$

This shows that we do in fact a spectral decomposition of the weighted covariance matrix of the weighted averages. The proposed analysis thus amounts to a weighted principal component analysis of the matrix of weighted averages. This analysis provides us a biplot of the matrix of weighted averages. Projecting species points onto variable vectors in such a biplot allows us to approximate the optima of the species as estimated by the weighted averages with respect to the environmental variables. With this approach, the display of the weighted averages is explicitly optimized. The representation of the samples (sites) is absent in this analysis. However, site coordinates can be added to a biplot in very much the same way as we added supplementary variables in CA (cf. chapter 7) or PCA (cf. chapter 8). In figure 11.1 we show such a (fictitious) biplot, with species points (•), a variable vector ($\mathbf{w}$), and supplementary site points (∘). The aim is to add the site vector $\mathbf{g}$ in an optimal way, where different criteria can

be used for what is meant by optimal.

First, we can try to optimize the representation of the species abundances. This amounts to minimizing the projection errors obtained when projecting the species points onto the added site vectors, as illustrated for one species in figure 11.1. Second, we can also try to optimally represent the environmental data matrix $\mathbf{Z}$, by minimizing projection errors for the sites onto the variable vectors. This is illustrated for one variable $\mathbf{w}$ and one site vector $\mathbf{g}$ in figure 11.1. It is also possible to minimize both projection errors simultaneously, in an attempt to display both $\mathbf{Z}$ and $\mathbf{N}$ as best as possible. This constitutes a compromise between the two alternatives just mentioned. In this chapter we develop the algebra for each of these methods, and give an example of an application.

We first have to specify the type of PCA we perform. In equation (11.2), we assume, for the sake of comparison with CCA, that the columns of $\mathbf{Z}$ are centred on their weighted means ($\mathbf{c}'\mathbf{Z}$), and standardized by dividing by the square root of their weighted variances. This does not mean that the matrix of weighted averages, $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{Z}$ is also standardized. If we want to consider a PCA of the correlation matrix, the latter matrix would first have to be standardized by postmultiplying by a diagonal matrix with the reciprocal of the square root of the variances of each of the columns of the matrix of weighted averages (note that this is a different kind of postmultiplication than the one performed in CCA). If we do not carry out this standardization, then we are performing a PCA of a covariance matrix, and not of a correlation matrix. An analysis based on the covariance matrix has the disadvantage that variables with a large variance dominate in the analysis. Here the matrix to be analyzed consists of weighted averages of the standardized values in $\mathbf{Z}$. The variances of these columns are not necessarily equal, but will be of the same order of magnitude. A PCA of such a covariance matrix is not likely to be dominated by a sole variable with a large variance. With the s.v.d. of equation (11.1), scores for the standardized principal components ($\hat{\mathbf{F}}$) and for the variable vectors ($\hat{\mathbf{W}}$) are given by:

$$\hat{\mathbf{F}} = \mathbf{D}_r^{-1/2}\hat{\mathbf{U}}, \qquad \hat{\mathbf{W}} = \hat{\mathbf{V}}\hat{\mathbf{D}}. \tag{11.3}$$

In the following two sections we derive expressions for adding site vectors to the PCA biplot considered, using the two different minimizations explained above.

11.3 Optimizing the Display of Abundances

The projection errors of the species points onto a hypothetical site vector $\mathbf{g}$ (a column vector), are given by $\alpha\hat{\mathbf{F}}\mathbf{g}$, and the objective function is:

$$\mathbf{e}'\mathbf{e} = (\alpha\hat{\mathbf{F}}\mathbf{g} - \mathbf{D}_r^{-1}\mathbf{p}_j)'(\alpha\hat{\mathbf{F}}\mathbf{g} - \mathbf{D}_r^{-1}\mathbf{p}_j), \tag{11.4}$$

where we assume to recover abundances as elements of profiles, and where $\mathbf{p}_j$ indicates the j^{th} column of the correspondence matrix, and α is a normalization factor. Note that $\mathbf{D}_r^{-1}\mathbf{p}_j$ is not a profile, but an $I \times 1$ column in the matrix of row profiles. We have to take care that the two vectors, estimates $\alpha\hat{\mathbf{F}}\mathbf{g}$ and data vector $\mathbf{D}_r^{-1}\mathbf{p}_j$ are centred on the same mean. This is guaranteed because the principal components have weighted mean zero ($\mathbf{r}'\hat{\mathbf{F}} = \mathbf{0}$) and the site vector

$\mathbf{D}_r^{-1}\mathbf{p}_j$ as well, if we assume the profiles have been centred on the average profile: $\mathbf{D}_r^{-1}\mathbf{p}_j \leftarrow \mathbf{D}_r^{-1}\mathbf{p}_j - \mathbf{1}c_j$, so that $\mathbf{r}'(\mathbf{D}_r^{-1}\mathbf{p}_j - \mathbf{1}c_j) = \mathbf{1}'\mathbf{p}_j - c_j = c_j - c_j = 0$. This minimization problem is entirely equivalent to the one previously described, when looking for optimal directions for supplementary variables in PCA. We apply solution (8.10) to find:

$$\frac{\mathbf{g}}{\parallel \mathbf{g} \parallel} = \frac{1}{\sqrt{\mathbf{p}_j{}'\mathbf{D}_r^{-1}\hat{\mathbf{F}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-2}\hat{\mathbf{F}}'\mathbf{D}_r^{-1}\mathbf{p}_j}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'\mathbf{D}_r^{-1}\mathbf{p}_j\,, \tag{11.5}$$

or, in matrix notation, we obtain all site vectors simultaneously as the rows of the matrix $\hat{\mathbf{\Gamma}} = \mathbf{D}_g\mathbf{P}'\mathbf{D}_r^{-1}\hat{\mathbf{F}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1}$ with $\mathbf{D}_g = \mathrm{diag}(\mathbf{P}'\mathbf{D}_r^{-1}\hat{\mathbf{F}}(\hat{\mathbf{F}}\hat{\mathbf{F}})^{-2}\hat{\mathbf{F}}'\mathbf{D}_r^{-1}\mathbf{P})^{-1/2}$. The data matrix $\mathbf{D}_r^{-1}\mathbf{P}$ is then approximated by $\hat{\mathbf{F}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'\mathbf{D}_r^{-1}\mathbf{P}$.

It is natural to weight the errors in the minimization above by the square root of the total abundance of each species, as this is also done in the analysis given by (11.1), and thus to minimize $\mathbf{e}'\mathbf{D}_r\mathbf{e}$. This simplifies the solution to:

$$\frac{\mathbf{g}}{\parallel \mathbf{g} \parallel} = \frac{\hat{\mathbf{F}}'\mathbf{p}_j}{\sqrt{\mathbf{p}_j'\hat{\mathbf{F}}\hat{\mathbf{F}}'\mathbf{p}_j}}, \tag{11.6}$$

which for all sites simultaneously gives $\hat{\mathbf{\Gamma}} = \mathbf{D}_g\mathbf{P}'\hat{\mathbf{F}}$. The profiles are then approximated by $\hat{\mathbf{F}}\hat{\mathbf{F}}'\mathbf{P}$. The quality of representation can be calculated as the fraction of the weighted variance in the species profiles explained by this approximation.

Given that an optimal direction has been found for the different sites, can we calculate how well the environmental data is represented? We can project the site vectors onto the environmental variables, and work out what part of the weighted variance of $\mathbf{Z}$ they make up. This is however, somewhat arbitrary, because it will depend on the norm we choose for the supplementary vectors $\mathbf{g}$. Therefore, it is difficult to say how well the environmental data in $\mathbf{Z}$ are represented in comparison with CCA.

11.4 Optimizing the Display of Environmental Data

It is evident that with two variables environmental data can be represented without error. With two variables, we can draw perpendiculars from the two variable vectors at the true chemical values measured at that site. The point where the two perpendiculars intersect is the optimal position for the site point. With more than two variables this is not possible any more, and we have to allow for error. We indicate one site, the j^{th} row of matrix $\mathbf{Z}$, as the $Q \times 1$ column vector $\mathbf{z}_j$. First, we take care that the observations in this vector are centred on the respective weighted means: $\mathbf{z}_j \leftarrow \mathbf{z}_j - \mathbf{Z}'\mathbf{c}$. If we consider one supplementary site vector $\mathbf{g}$, then its projections onto all variable vectors, given by the rows of $\hat{\mathbf{W}}$, are $\alpha\mathbf{D}_w\hat{\mathbf{W}}\mathbf{g}$, with $\mathbf{D}_w = \mathrm{diag}(\hat{\mathbf{W}}\hat{\mathbf{W}}')^{-1/2}$. These projections

now approximate a case (row) of the matrix of environmental variables, $\mathbf{Z}$. We minimize:

$$\mathbf{e}'\mathbf{e} = (\mathbf{z}_j - \alpha\mathbf{D}_w\hat{\mathbf{W}}\mathbf{g})'(\mathbf{z}_j - \alpha\mathbf{D}_w\hat{\mathbf{W}}\mathbf{g}). \tag{11.7}$$

Even though we represent a supplementary site, and not a variable as considered previously in chapters 7 and 8, the algebraical problem is very similar, and the solution is given by applying result (8.12):

$$\frac{\mathbf{g}}{\parallel \mathbf{g} \parallel} = \frac{1}{\sqrt{\mathbf{z}_j'\mathbf{D}_w\hat{\mathbf{W}}(\hat{\mathbf{W}}'\mathbf{D}_w^2\hat{\mathbf{W}})^{-2}\hat{\mathbf{W}}'\mathbf{D}_w\mathbf{z}_j}}(\hat{\mathbf{W}}'\mathbf{D}_w^2\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}'\mathbf{D}_w\mathbf{z}_j\,. \tag{11.8}$$

All supplementary site coordinates can be obtained simultaneously with the matrix expression $\hat{\mathbf{\Gamma}} = \mathbf{D}_\alpha\mathbf{Z}\mathbf{D}_w\hat{\mathbf{W}}(\hat{\mathbf{W}}'\mathbf{D}_w^2\hat{\mathbf{W}})^{-1}$, where $\hat{\mathbf{\Gamma}}$ is the J by Q matrix of site coordinates, and $\mathbf{D}_\alpha = \mathrm{diag}(\mathbf{Z}\mathbf{D}_w\hat{\mathbf{W}}(\hat{\mathbf{W}}'\mathbf{D}_w^2\hat{\mathbf{W}})^{-2}\hat{\mathbf{W}}'\mathbf{D}_w\mathbf{Z}')^{-1/2}$ takes care of the normalization of the rows of $\hat{\mathbf{\Gamma}}$. Matrix $\mathbf{Z}'$ is approximated by the projections $\mathbf{D}_w\hat{\mathbf{W}}\hat{\mathbf{\Gamma}}'\mathbf{D}_\alpha^{-1} = \mathbf{D}_w\mathbf{W}(\mathbf{W}'\mathbf{D}_w^2\mathbf{W})^{-1}\mathbf{W}'\mathbf{D}_w\mathbf{Z}'$. When the full space of the PCA solution is considered, the latter expression collapses to $\mathbf{Z}'$, data being recovered exactly. When using only the first two dimensions of a PCA, we only use the first two columns of $\hat{\mathbf{W}}$ and recover $\mathbf{Z}$ only approximately. The fraction of the weighted variance of this approximated $\mathbf{Z}$ with respect to the total weighted variance of $\mathbf{Z}$ is again used as a measure for the quality of representation.

We note that equations (11.6) and (11.8) again represent normalized regression coefficients. The response "variables" in those regressions do not need to be variables in the usual sense, they can as well correspond to cases in a data matrix.

11.5 An Example with Artificial Data

We use the same artificial data of chapter 9 in table 9.2 (page 130), and apply a PCA to the matrix of weighted averages, where we add sites as supplementary information. The result is shown in figure 11.2.
Since there are three variables, a maximum of three principal components can be extracted. Table 11.1 provides the quality of the display of the different matrices, for the CCA and for two different approaches considered above.
For matrix $\mathbf{N}$, the criterion for the quality of representation is the fraction of the weighted variance of the row profiles explained by the biplot. When optimizing the display of $\mathbf{N}$, 97.7% of the weighted variance of the profiles is captured by the two-dimensional solution, so we can recover species profiles with confidence. If we compare figure 11.2 with the profiles in table 9.2 on page 130, then we see that the figure is consistent with these numbers. For instance, in the graph the species line up along site A in order 2,3,1,4 and 5 which is the same as the order of the profiles values in table 9.2. Projections onto other site vectors are also largely in agreement with the data table. In fact, the ordination diagram of figure 11.2 highly resembles the ordination diagram obtained by CCA in figure

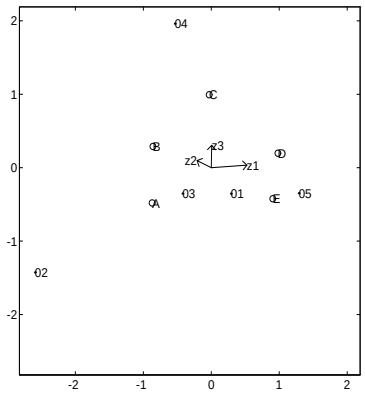

FIGURE 11.2: WPCA OF SPECIES OPTIMA WITH SUPPLEMENTARY SITES (OPTIMIZING **N**)

9.4 on page 136.

To facilitate the comparison, the ordination in figure 11.2 has been reflected in the horizontal axis. The qualitative interpretation of the two maps is not different. In CCA the dispersion of the sample points is seen to be larger. In the analysis in figure 11.2 we found the coordinates for the samples using the arbitrary norm one constraint, and with a different choice of norm a different degree of dispersion would have been obtained.

We can not compare this fraction of weighted variance of the profiles with the quality of the species data in CCA, since the latter method uses a different cri-

Method	Dim	CV(**N**)[a]	%CV(**N**)	Inertia	%I[b]	V(WA)[c]	%CV(WA)[d]	CV(**Z**)[e]	%CV(**Z**)
PCA +	1	0.0483	68.64	-	-	0.3190	75.26	-	-
opt. **N**	2	0.0688	97.70	-	-	0.1033	99.63	-	-
	3	0.0704	99.97	-	-	0.0016	100.00	-	-
PCA +	1	-	-	-	-	0.3190	75.26	1.3588	45.29
opt. **Z**	2	-	-	-	-	0.1033	99.63	2.5258	84.19
	3	-	-	-	-	0.0016	100.00	3.0000	100.00
CCA	1	-	-	0.2936	72.76	0.3087	72.84	1.0516	35.05
	2	-	-	0.0917	22.72	0.4216	99.47	2.2829	76.10
	3	-	-	0.0032	0.78	0.4239	100.00	3.0000	100.00
	4	-	-	0.0151	3.73	-	-		-
Total		-	-	0.4035	100.00				

[a]Cumulative variance of species profiles
[b]Percentage of inertia
[c]For CCA variances are cumulative
[d]Cumulative variance of weighted averages
[e]Cumulative variance of environmental variables

TABLE 11.1: VARIANCE DECOMPOSITION OF WPCA WITH 3 VARIABLES

terion for the fit of the species data, namely the inertias, as given in columns 5 and 6 of table 11.1.

Note that we indeed obtain slightly higher amounts of variance explained for the weighted averages than CCA does (columns 7 and 8). Thus, we have achieved a better representation of the species preferences. This was to be expected, as we now explicitly optimized these. We expect this to be true for any data set.

When we optimize, after the WPCA, the representation of **Z**, then the 2D solution captures 84% of the weighted variance present in the environmental data, whereas the CCA of the same data captures 76.1% of the variance of **Z**. The biplot of this analysis is shown in figure 11.3. The site points have changed their positions and should now be interpreted with respect to the variable vectors. The sites line up along the variables vectors in approximately the right order (cf. table 9.2). The better display of **Z** holds for this data set, and without more theoretical work, we cannot generalize about this beyond the particular data set studied.

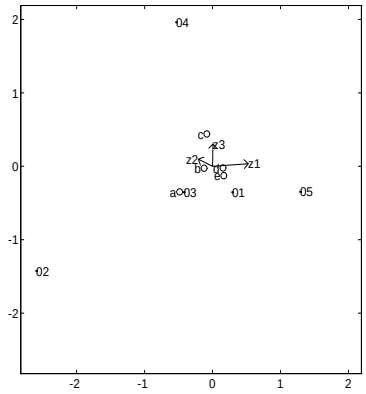

FIGURE 11.3: WPCA OF SPECIES OPTIMA WITH SUPPLEMENTARY SITES (OPTIMIZING **Z**)

11.6 An Application with Ekofisk Data

We applied weighted principal component analysis to the matrix of weighted averages of the 50 most abundant species. Figure 11.4 graphs the result of this analysis, where the site points were added to the biplot such as to represent the species profiles as best as possible.
To make the graph more interpretable, the length of the variable vectors was incremented by a factor 10, and the length of the site vectors was incremented by a factor 3. Since it is the relative position of the species with respect to both these sets of vectors that matters, this increase of vector length does not affect the interpretation, at least if we refrain from interpreting the vector length.

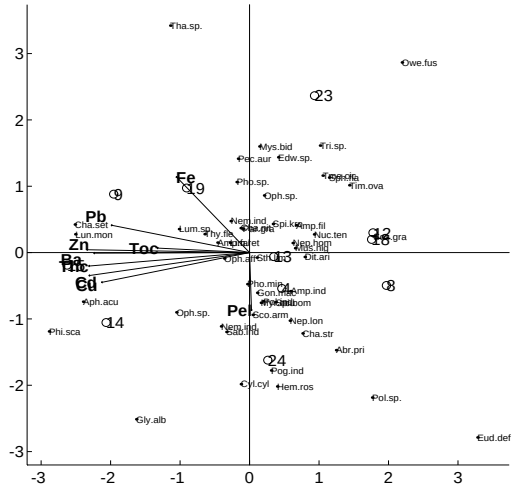

FIGURE 11.4: WPCA OF 1992; (OPTIMIZING **N**)

The corresponding CCA of this data set was discussed in section 10.2.1. The ordination obtained here strongly resembles the ones discussed in 10.2.1 and 10.2.2. *Chaetozone setosa* is again high on the contaminants, *Eudorellopsis deformis* low, *Mysella bidentata* and *Phoronis sp.* are again high on Fe. The qualitative interpretation of the two types of analysis is essentially the same. The statistics in table 11.2 show however that the WPCA explains more than 91% of the variance in the weighted averages, about 8.5 percent more than CCA. Graph 11.4 displays about 42% of the variance of the species profiles.

Method	Dim	CV(**N**)	%CV(**N**)	Inertia	%CI	V(WA)	%CV(WA)	CV(**Z**)	%CV(**Z**)
PCA +	1	0.0056	0.2793	-	-	0.3262	84.86	-	-
opt. **N**	2	0.0085	0.4239	-	-	0.0261	91.65	-	-
	3	0.0189	0.9430	-	-	0.0199	96.83	-	-
PCA +	1	-	-	-	-	0.3262	84.86	6.1508	68.34
opt. **Z**	2	-	-	-	-	0.0261	91.65	7.1999	80.00
	3	-	-	-	-	0.0199	96.83	8.1177	90.20
CCA	1	-	-	0.0678	30.6	0.2761	71.8	4.0724	45.2
	2	-	-	0.0395	48.4	0.3200	83.2	5.1842	57.6
	3	-	-	0.0263	60.2	0.3401	88.5	5.9478	66.1

TABLE 11.2: VARIANCE DECOMPOSITION OF WPCA FOR 1992 EKOFISK DATA

When we optimize the display of the environmental data after the PCA, stations 9, 14 and 15 appear as the most contaminated stations, just like in CCA. As shown in table 11.2, a 2D biplot of that analysis captures 80% of the variance of the environmental data, whereas the corresponding CCA captures 58%.

11.7 Conclusions

In this chapter we considered the PCA of the matrix of weighted averages as an alternative for CCA. Ordinations obtained by this method are very similar to the ones obtained by CCA. Amounts of weighted variance explained of the species optima are higher. If we choose to optimize the display of the environmental data after the PCA, then we can also account for more variance of the environmental data.

Chapter 12

Suggestions for Further Research

A doctoral thesis is never finished, though from a practical point of view, a book can not grow without limit. In this last chapter we therefore consider a few suggestions for further research that are inspired by the work presented in previous chapters.

12.1 Canonical Correlation Analysis

It may have come as a surprise that canonical correlation analysis (CCR) has not been applied to the survey data. There is an entire book by Gittins (1985) dedicated to applications of CCR to ecological data. First, it should be noted that CCR has not become such a popular multivariate methods as CA, PCA, or, in an ecological context, CCA. This probably has to do with the fact that all these method produce fancy biplots which are relatively easy to interpret. On the contrary, the output of a CCR is highly numerical. From a more practical point of view, the computations become difficult because there is a singularity problem, and standard statistical packages complain. In this survey there are many more species than sites, which makes one of the variance-covariance matrices used in CCR singular. Mardia (1979), referring to Rao (1971), notes these problems can be solved by the use of generalized inverses, but no applications are given. It would be interesting to try (and interpret) CCR, with a properly implemented generalized inverse. A program that can be modified for this purpose is given in appendix A.3.

12.2 Redundancy Analysis

It has been noted in chapter 4 that many species decrease in abundance as a function of the environmental variables, and that only a few species show a unimodal response. If the decrease is close to linear, then redundancy analysis might be a promising alternative for the analysis of the survey data. Redun-

dancy analysis corresponds to principal component analysis with linear constraints (Ter Braak and Prentice, 1988).

12.3 Data Fusion Problems

In section 2.1 it has been mentioned that the station network has been reduced over the years in order to save expenses (cf. section 2.1). Priority has been given to the chemical data: for all three consecutive years the chemical information is present for about 40 stations, but in 1991 and 1992 biological variables have been measured for about 12 stations only. Thus, there exists a large hole in the biological data for these years. The estimation of the missing biological data for 1991 and 1992 constitutes a data fusion problem. Different methods can be conceived to estimate the missing biological data, like imputations by nearest neighbour methods, or prediction with previously estimated regression equations.

12.4 PLS regression

The number of samples in the survey data is small with respect to the number of variables, in particular for the data from 1991 and 1992. If we would like to investigate, by ordinary multiple regression, how the abundance of one particular species depends on the abundance of others, then this in not possible, due to the singularity of the cross-product matrix of the predictors, $\mathbf{X}'\mathbf{X}$. PLS (Partial Least Squares) regression (Tenenhaus, 1998) was specially designed for the situation where the number of predictors exceeds the number of observations, and could be applied in these circumstances.

Appendix A

Some Computer programs

A.1 Estimation of a Zero-Inflated Poisson

```
*! version 1.0 Jan Graffelman
program define poi2zero
   version 5.0
   local varlist "required existing"
   parse "'*'"
   parse "'varlist'",parse(" ")

   gen i = 1
   replace i = 0 if `1' == 0
   quietly summ i, detail
   local te = 1.0 - _result(3)
   local mu = ln(`te'/(1-`te'))
   quietly summ `1' if i, detail
   local la = _result(3)
   local lla = ln(`la')
   matrix b0 = (`lla',`mu')
   matrix colnames b0 = lla1:_cons mu:_cons

   ml begin
   ml function cas2cont
   ml method lf
   eq lla1 : `1'
   eq mu :
   ml model b = lla1 mu, depv(10) from(b0)
   ml sample mysamp
   ml max f v
   ml post mixtlf, title(Poisson With Zeros:lf method)
   ml mlout mixtlf
   capture drop i
   local la1 = exp([lla1][_cons])
   local te1 = 1/(exp(-[mu][_cons])+1)
   local ste1 = `te1'*(1-`te1')*[mu]_se[_cons]
   local sla1 = `la1'*[lla1]_se[_cons]
   local te1ll = `te1'-invnorm(0.975)*`ste1'
   local te1ul = `te1'+invnorm(0.975)*`ste1'
   local la1ll = `la1'-invnorm(0.975)*`sla1'
   local la1ul = `la1'+invnorm(0.975)*`sla1'
   #delimit ;
   disp in green _col(1) "------------------------------------------------------------------------------" ;
   disp in yellow "$S_eqnm1" _col(12) in yellow %10.6f `la1' _col(23) in yellow %10.6f `sla1'
   _col(57) in yellow %10.6f `la1ll' _col(70) in yellow %10.6f `la1ul' ;
   disp in green _col(1) "------------------------------------------------------------------------------" ;
   disp in yellow "$S_eqnm2" _col(12) in yellow %10.6f `te1' _col(23) in yellow %10.6f `ste1'
   _col(57) in yellow %10.6f `te1ll' _col(70) in yellow %10.6f `te1ul' ;
   disp in green _col(1) "------------------------------------------------------------------------------" ;
   #delimit cr
end


program define cas2cont
   local lnf      "`1'"
   local lla1     "`2'"
   local mu       "`3'"
   #delimit ;
   quietly replace `lnf' = cond(i,-ln(1+exp(`mu'))+`lla1'+$S_mldepn
                           - exp(`lla1') - lnfact($S_mldepn),
                           ln(1-exp(-`mu'-exp(`lla1'))) -
                           ln(exp(-`mu') + 1)) ;
   #delimit cr
end
```

A.2 Canonical Correspondence Analysis

```
function y = cca(N,Z,verbose)
%
% Examples:
%
%    cca(N,Z,verbose)
%    cca(N,Z,1)
%    cca(N,Z)
%
%
% Parameters:
%
%    N:          The I x J (species by sites) abundance matrix.
%    Z:          The Q x J (variables by sites) environmental data matrix.
%    verbose:    a number determining how much output is shown.
%                   0 - no numerical output
%                   1 - show everything (profiles, principal & standard coordinates, etc).
%
% Jan Graffelman
% Universitat Pompeu Fabra
% Last change: September 1999
%

[I J] = size(N);        % I species, J sites, Q variables.
[J Q] = size(Z);

n = sum(sum(N));        % gran total
P = N/n;                % percentage table
r = sum(P')';           % row masses (average column profile)
c = sum(P)';            % column masses (average row profile)
Dr = diag(r);           % row masses in diagonal form
Dc = diag(c);           % column masses in diagonal form
RP = inv(Dr)*P;         % row profiles
Corr = corrcoef(Z);     % correlation between the variables.

OZ = Z;

% centre Z on weighted mean

Z = wcen(Z,c);

% standardize by dividing by weighted standard deviation.

S = Z'*Dc*Z;
wvar=diag(S);
Z = Z*inv(sqrt(diag(wvar)));

% weighted correlation matrix

S = Z'*Dc*Z;

% weighted averages of species

WA = (inv(Dr)*P - ones(length(r),1)*c')*Z;

[U1,D1,W] = gensvd(WA,Dr,pinv(S));

k = rank(D1);
D1 = D1(1:k,1:k);
U1 = U1(:,1:k);
W = W(:,1:k);

% Calculate coordinates

F1 = U1*D1;
H = W*D1;

PHI = F1*inv(D1);
OMEGA = W;

% Calculate Weights.

B = pinv(S)*W;

% Site points (standard coordinates)

SP = Z*B;

% site points in principal coordinates:

G1 = SP*D1;

%
% Analyse the unrestricted dimensions
%

R = (inv(Dr)*P - ones(length(r),1)*c') - U1*D1*B'*Z'*Dc;

% rank remaining dimensions

rr = rank(R);
rr = (J-1)-Q;

[Uu Du Vu] = gensvd(R,Dr,inv(Dc));

Uu = Uu(:,1:rr);
Vu = Vu(:,1:rr);
```

```
Du = Du(1:rr,1:rr);

% Coordinates in unrestricted dimensions

Fu = Uu*Du;
PHIu = Uu;
Gu = inv(Dc)*Vu*Du;
GAMu = inv(Dc)*Vu;

% Site scores in restricted dimensions that are NOT LC of
% the environmental variables (Ter Braak's sample scores)
% Note: results do not coincide with canoco for the unrestricted
% dimensions.

D = diag([diag(D1) ; diag(Du)]');
TBS = inv(Dc)*P'*[PHI PHIu]*inv(D);

% Total CCA inertia

CCA_IN = trace(D1*D1)+trace(Du*Du);

% CCA inertia in restricted space

RES_CCA_IN = trace(D1*D1);

% CCA inertia in remaining dimensions

UNRES_CCA_IN = trace(Du*Du);

% Principal Inertias: absolute, relative and cumulative

INABS = [diag(D1*D1); diag(Du*Du)]';
INREL = INABS/sum(INABS)*100;
INCUM = cumsum(INREL);

% inertia contributions

G = [G1 Gu];
DECsi = Dc*G.*G;

F = [F1 Fu];
DECsp = Dr*F.*F;

% contributions species to axes

sptoax = DECsp*inv(diag(INABS));

% contributions axes to species

axtosp = inv(diag(sum(DECsp')))*DECsp;

% contributions spites to axes

sitoax = DECsi*inv(diag(INABS));

% contributions axes to sites

axtosi = inv(diag(sum(DECsi')))*DECsi;

% Ter Braak inter set correlations

INTER = wcorrm([Z TBS],c);
INTER = INTER(1:Q,(Q+1):2*Q);

if verbose == 1

   fprintf(1,'Numerical Output CCA\n\n')
   disp('Abundance Matrix:')
   disp(N)

   disp('Chemicial Data (variables x sites): ')
   disp(NZ)

   disp('Chemicial Data (variables x sites) (centered):')
   disp(Z)

   disp('Rank Abundance Matrix:')
   disp(rank(N))

   disp('Rank Environmental data Matrix:');
   disp(rank(Z))

   disp('Correspondence Matrix:')
   disp(P)

   disp('Species profiles:')
   disp(RP)

   disp('Column masses:')
   disp(c)

   disp('Row masses:')
   disp(r)

   disp('Total Inertia for CCA')
   disp(CCA_IN)

   disp('Inertias in restricted and unrestricted dimensions')
   disp([CCA_IN RES_CCA_IN UNRES_CCA_IN])
```

```
   disp([100 RES_CCA_IN/CCA_IN*100 UNRES_CCA_IN/CCA_IN*100])

   disp('Principal Inertias')
   disp([INABS; INREL; INCUM])

   disp('Species Inertias')
   disp(sum(DECsp'))

   disp('Contributions of species to axes')
   disp(sptoax)

   disp('Contributions axes to species')
   disp(axtosp)

   disp('Site Inertias')
   disp(sum(DECsi'))

   disp('Contributions of sites to axes')
   disp(sitoax)

   disp('Contributions axes to sites')
   disp(axtosi)

   disp('Principal coordinates of the Species: ');
   disp([F1 Fu])

   disp('Principal coordinates of the Variables: ')
   disp(H)

   disp('Standard coordinates of the Species: ');
   disp(PHI)

   disp('Standard coordinates of the Variables: ')
   disp(OMEGA)

   disp('Standard coordinates of the Sites: ')
   disp(SP)

   disp('Ter Braak sample scores (not LC): ')
   disp(TBS)

   disp('Principal coordinates of the sites: ')
   disp([G1 Gu])

   disp('Weighted averages of the Chemical Variables:');
   disp(c'*OZ)

   disp('Ordinary Correlations between the variables:');
   disp(Corr)

   disp('Weighted Correlations between the variables:');
   disp(wcorrm(Z,c))

   disp('Ter Braak Inter set Correlations: (env. var x axes)')
   disp(INTER)

   disp('Weighted averages of the Species for the chemical variables');
   disp('using abundances as weights:');

   disp(RP*Z)

   disp('Weights for the environmental Variables:')
   disp(B)

end
```

A.3 Canonical Correlation Analysis

```
function [U, V] = canocorr(X,Y,verbose)
%
% Examples:
%
%    canocorr(X,Y,verbose)
%
% Parameters:
%
%    X:          first data matrix
%    Y:          second data matrix
%    verbose:    0 - be silent (default) 1 - show numerical output.
%    U:          canonical variates X-variables
%    V:          canonical variates Y-variables
%
% The program CANOCORR performs Canonical Correlation Analysis
%
% Jan Graffelman
% University Pompeu Fabra
% Last change 17 september 1999

if exist('verbose') == 0
   verbose = 0;
end

[n,p] = size(X);
[n,q] = size(Y);

Xc = sd(X);
```

```
Yc = sd(Y);

S11 = 1/(n-1)*Xc'*Xc;
S22 = 1/(n-1)*Yc'*Yc;
S12 = 1/(n-1)*Xc'*Yc;

[v,d] = eig(S11);
rr = rank(S11);
v = v(:,1:rr);
d = d(1:rr,1:rr);
S11mh = v*pinv(sqrt(d))*v';

[v,d] = eig(S22);
rr = rank(S22);
v = v(:,1:rr);
d = d(1:rr,1:rr);
S22mh = v*pinv(sqrt(d))*v';

% Computational scheme: singular value decomposition

K = S11mh*S12*S22mh;
[uu,dd,vv] = svd(K,0);
dim = min([p q]);
dd = dd(1:dim,1:dim);
uu = uu(:,1:dim);
vv = vv(:,1:dim);

% Canonical Weights

A = S11mh*uu;
B = S22mh*vv;

% Canonical Variates

U = Xc*A;
V = Yc*B;

% compute canonical loadings (correlations with original variables)

R = corrcoef([X U]);
Rx = R(1:p,(p+1):(p+dim));

R = corrcoef([Y V]);
Ry = R(1:q,(q+1):(q+dim));

% variance explained by canonical variates

Vex = 1/p*diag(Rx'*Rx);
Vey = 1/q*diag(Ry'*Ry);

% redundancy coefficients (amount of variance in Y-set accounted
% for by the X-set).

Redygx= dd*dd*Vey;
Redxgy= dd*dd*Vex;

% Cross loadings

R = corrcoef([Y U]);
CrossYU = R(1:q,(q+1):(q+dim));

R = corrcoef([X V]);
CrossXV = R(1:p,(p+1):(p+dim));

Wilks = det(eye(dim,dim)-dd*dd);
Chi = -1*((n-1) - 0.5*(p+q+1))*log(Wilks);
pval = 1-chi2cdf(Chi,p*q);

if verbose == 1
   disp('Canonical Weights (Coefficients of LC) variables x variates')
   A
   B
   disp('Canonical Variates')
   U
   V
   disp('Canonical Correlations')
   diag(dd) % same as cov([U V])
   disp('Correlations with X-Variables (Canonical loadings: Xvariables x Xvariates)')
   Rx
   disp('Correlations with Y-Variables (Canonical loadings: Yvariables x Yvariates)')
   Ry
   disp('Correlations with X-Variables (Cross loadings: Xvariables x Yvariates)')
   CrossXV
   disp('Correlations with Y-Variables (Cross loadings: Yvariables x Xvariates)')
   CrossYU
   disp('% variance explained by canonical x-variates:')
   [(1:dim)' Vex]
   disp('% variance explained by canonical y-variates:')
   [(1:dim)' Vey]
   disp('Reduncancy coefficient (amount of of variance in criterion set accounted for by predictor set)')
   Redygx
   disp('Reduncancy coefficient (amount of of variance in predictor set accounted for by criterion set)')
   Redxgy
   disp('Significance of first canonical variate:')
   fprintf(1,'Wilks lambda: %6.4f Chi^2: %10.4f p-value: %6.4f\n',[Wilks Chi pval])
end
```

A.4 Correspondence Analysis

```
function y = sca(x,verbose,plottype)
%
% Examples:
%
%    sca(x)
%    sca(x,verbose,plottype)
%
%
% Parameters:
%
%    x:          a raw I x J contingency table
%    verbose:    a number determining how much output is shown.
%                  0 - no numerical output
%                  1 - show everything (profiles, principal & standard coordinates, etc).
%    plottype:   allows to specify which plot is generated. There are
%                4 possibilities:
%                  0 - No graphical output.
%                  1 - Symmetric map (default)
%                  2 - Asymmetric map of the rows.
%                  3 - Asymmetric map of the columns.
%
% The program SCA performs simple correspondence analysis.
%
% Jan Graffelman
% University Pompeu Fabra
% Last change 14 february 1996

[I,J] = size(x);

if exist('verbose') == 0
   verbose = 0;
end

if exist('plottype') == 0
   plottype = 0;
end

%
% Preparation of the data.
%

N = sum(sum(x));                % grand total
P = x/N;                        % table of percentages, the correspondence table
[I J] = size(P);                % get dimensions of the IxJ table
r = (sum(P'))';                 % row masses (average column profile)
c = (sum(P))';                  % column masses (average row profile)
Dr = diag(r);
Dc = diag(c);
RP = inv(Dr)*P;                 % row profiles
CP = P*inv(Dc);                 % column profiles

A = inv(sqrt(Dr))*(P - r*c')*inv(sqrt(Dc)); % standardized residuals.

%
% Chisquare calculations:
%

CHI = sum(sum(A.*A))*N;
chicon = N*A.*A;

%
% SVD and calculation of coordinates.
%

[U,D,V] = svd(A,0);

k = rank(D);
D = D([1:k], [1:k]);
U = U(:,[1:k]);         % basis for the rows
V = V(:,[1:k]);         % basis for the columns

%
% principal coordinates:
%

F = inv(sqrt(Dr))*U*D;
G = inv(sqrt(Dc))*V*D;

%
% standard coordinates:
%

PHI = inv(sqrt(Dr))*U;
GAM = inv(sqrt(Dc))*V;

%
% Inertia and inertia contributions:
%

PRIN_IN  = D*D;

% total inertia:

IN_TOT = sum(diag(PRIN_IN));

% Percentage of explained dispersion for each dimension:
```

```
IN_DIM = (diag(P&IN_IN)/IN_TOT)'*100;

% Cumulative percentage of explained dispersion:

IN_DIM_CUM = cumsum(IN_DIM);

% Decompostion of Principal Inertias for each row:

&OW_DEC = Dr*F.*F;

% &ow inertias:

&OW_IN&RT = sum(&OW_DEC');

% &ow inertias relative to total inertia.

&OW_IN&RT_&ELTOTOT = &OW_IN&RT/sum(&OW_IN&RT);

% Correlations of row profiles and axes:

SQU&R_&OW_CO&R = inv(diag(&OW_IN&RT))*&OW_DEC;

&OW_CO&R = sign(F).*sqrt(SQU&R_&OW_CO&R);

% Contributions of principal axis to the rows (or
% quality of the rows for each principal axis)

% Contributions of axis to rows.

COM_AXTO&OW = SQU&R_&OW_CO&R;

% Contribution of rows to axis.

COM_&OWTOAX = &OW_DEC*inv(P&IN_IN);

% Quality in two dimensions

QU&_&OW = COM_AXTO&OW(:,1:2);
QU&_&OW = sum(QU&_&OW');

%
% &ow similar things for the columns
%

% decomposition of inertia for each column.

COL_DEC = Dc*G.*G;

% Contributions of the columns to principal inertias

COM_COLTOAX = COL_DEC*inv(P&IN_IN);

% Column inertias

COL_IN&RT = sum(COL_DEC');

% Column inertias relative to total inertia.

COL_IN&RT_&ELTOTOT = COL_IN&RT/sum(COL_IN&RT);

% Correlations of column profiles and axes:

SQU&R_COL_CO&R = inv(diag(COL_IN&RT))*COL_DEC;

COL_CO&R = sign(G).*sqrt(SQU&R_COL_CO&R);

% Contributions of principal axis to columns (or
% quality of the columns for each principal axis)

% Contributions of axis to columns

COM_AXTOCOL = SQU&R_COL_CO&R;

% Contributions of columns to axis

COM_COLTOAX = COL_DEC*inv(P&IN_IN);

% Quality in two dimensions

QU&_COL = COM_AXTOCOL(:,1:2);
QU&_COL = sum(QU&_COL');

%
% &ow show all numerical output
%

if verbose == 0
    :
elseif verbose == 1
   disp('Grand total:')
   disp(N);
   disp('Correspondence Matrix:');
   disp(P);
   disp('&ow masses:');
   disp(r);
   disp('Column masses:');
   disp(c);
   disp('&ow Profiles:')
```

```
   disp(RP);
   disp('Column Profiles:')
   disp(CP);
   disp('Standardized residuals:');
   disp(A);
   disp('Chi-square:');
   disp(CHI);
   disp('Chi-square contributions:');
   disp(chicon);
   disp('Singular Values:');
   disp(D);
   disp('Left singular vectors:');
   disp(U);
   disp('Right singular vectors:');
   disp(V);
   disp('Principal coordinates of the rows:');
   disp(F);
   disp('Principal coordinates of the columns:');
   disp(G);
   disp('Standard coordinates of the rows:');
   disp(PHI);
   disp('Standard coordinates of the columns:');
   disp(GAH);
   disp('Total Inertia:');
   disp(IN_TOT);
   disp('Inertias for each dimension:');
   disp(diag(PRIN_IN));
   disp('Percentage of explained dispersion for each dimension:')
   disp(IN_DIM);
   disp('Cumulative percentage of explained dispersion:')
   disp(IN_DIM_CUM);
   disp('Decomposition of inertia for each row:')
   disp(ROW_DEC)
   disp('Contribution of each row on the inertia of each dimension:');
   disp(COR_ROWTOAX);
   disp('Contribution of each principal axis to the rows:')
   disp(COR_AXTOROW);
   disp('Correlations of row profiles and axes:')
   disp(ROW_CORR)
   disp('Row inertias:')
   disp(ROW_INERT)
   disp('Row inertias relative to total:')
   disp(ROW_INERT_RELTOTOT)
   disp('Quality of the rows for each principal axis:')
   disp(COR_ROWTOAX)
   disp('Quality of the rows in two dimensions:')
   disp(QUA_ROW')
   disp('Decomposition of inertia for each column:')
   disp(COL_DEC)
   disp('Contributions of the columns to principal inertias:')
   disp(COR_COLTOAX)
   disp('Correlations between columns and principal axes:')
   disp(COL_CORR)
   disp('Column inertias:')
   disp(COL_INERT)
   disp('Column inertias relative to total:')
   disp(COL_INERT_RELTOTOT)
   disp('Quality of the columns for each principal axis')
   disp(COR_COLTOAX)
   disp('Quality of the columns in two dimensions')
   disp(QUA_COL)
else
   error('unknown value for parameter verbose')
end

%
% Now show the graphical output
%

if plottype == 0
   ;
elseif plottype == 1
   plot(F(:,1),F(:,2),'.',G(:,1),G(:,2),'o');
   ax([F(:,1:2); G(:,1:2)]);
   title('Symmetric Map');
elseif plottype == 2
   plot(F(:,1),F(:,2),'.',GAH(:,1),GAH(:,2),'o');
   ax([F(:,1:2); GAH(:,1:2)]);
   title('Asymmetric Map of the Rows');
elseif plottype == 3
   plot(G(:,1),G(:,2),'.',PHI(:,1),PHI(:,2),'o');
   ax([G(:,1:2); PHI(:,1:2)]);
   title('Asymmetric Map of the Columns');
else
   error('Unkown value for parameter plottype');
end;
```

Bibliography

Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit techniques.* Marcel Dekker, New York.

Barbolla, R. and Sanz, P. (1998). *Álgebra lineal y teoría de matrices.* Prentice Hall, Madrid.

Benzécri, J. P. (1973). *Analyse des Données.* Dunod, Paris.

Birks, H. J. B. and Austin, H. A. (1992). *An Annotated Bibliography of Canonical Correspondence Analysis and Related Constrained Ordination Methods 1986-1991.* Botanical Institute, Bergen, Norway.

Böckenholt, U. and Takane, Y. (1994). Linear constraints in correspondence analysis. In Greenacre, M. and Blasius, J., editor, *Correspondence Analysis in the Social Sciences*, pages 112–127. Academic Press.

Dargie, T. C. D. (1984). On the integrated interpretation of indirect site ordinations: a case study using semi-arid vegetation in southeastern spain. *Vegetatio*, 55:37–55.

Digby, P. G. N. and Kempton, R. A. (1987). *Multivariate Analysis of Ecological Communities.* Chapman and Hall.

Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis Methods and Applications.* John Wiley & Sons.

Dobson, A. J. (1991). *An introduction to generalized linear models. 2nd ed.* Chapman & Hall, New York.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis.* John Wiley & Sons, New York.

Dunn, G. (1989). *Design and Analysis of Reliability Studies.* Oxford University Press, Oxford.

Feller, W. (1971). *An Introduction to probability theory and its applications.* Wiley cop., New York.

Fieler, R. and Greenacre, M. (1994). Evaluation and development of statistical methods, main report. Technical report, AkvaPlan, Tromsø.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* John Wiley & Sons, New York.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Gabriel, K. R. and Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, 9(5):469–485.

Gifi, A. (1981). *Nonlinear Multivariate Analysis.* John Wiley.

Gittins, R. (1985). *Canonical Analysis.* Springer Verlag.

Goldstein, H. (1987). *Multilevel models in Educational and Social Research.* Oxford University Press, New York.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3):325–338.

Gower, J. C. (1971). Statistical methods of comparing different multivariate analysies of the same data. In Hodson, F. R., Kendall, D. G., and Tautu, P., editors, *Mathematics in the Archaeological and Historical Sciences*, pages 138–149, Edinburgh. Edinburgh University Press.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40:33–51.

Graffelman, J. (1997). Using environmental information in the correspondence analysis of abundance data. In *VI Conferencia española de biometria*, pages 109–110, Córdoba. Sociedad Española de Biometria.

Graffelman, J. (1998). A fundamental matrix result on scaling in multivariate analysis. *Econometric Theory*, 14(5):693–694.

Graffelman, J. (1999a). A distance-based approach to canonical correspondence analysis. In *VII Conferencia española de biometria*, pages 165–168, Palma de Mallorca. Sociedad Española de Biometria.

Graffelman, J. (1999b). The justification of multidimensional scaling under euclidean conditions. *Econometric Theory*, 15(6):908–909.

Graffelman, J. (1999c). Use of the moore-penrose inverse in canonical correspondence analysis. *Econometric Theory*, 15(5):777.

Graffelman, J., Fugger, E. F., Keyvanfar, K., and Schulman, J. D. (1999). Human live birth and sperm sex ratio compared. *Human Reproduction*, 14:2917–2920.

Graffelman, J. and Hoekstra, R. F. (2000). A statistical analysis of the effect of warfare on the human secondary sex ratio. *Human Biology*, 72(3). In press.

Graffelman, J. and van de Velden, M. (1999). Upper bounds for the eigenvalues of the product of a symmetric idempotent and a non-negative definite matrix. *Econometric Theory*, 15(4):631.

Graybill, F. A. (1983). *Matrices with Applications in Statistics*. Wadsworth and Brooks/Cole.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.

Greenacre, M. J. (1993a). Biplots in correspondence analysis. *Journal of Applied Statistics*, 20(2):251–269.

Greenacre, M. J. (1993b). *Correspondence Analysis in Practice*. Academic Press.

Greenacre, M. J. and Blasius, J. (1994). *Correspondence Analysis in the Social Sciences*. Academic Press.

Hamilton, L. C. (1992). *Regression with graphics a second course in applied statistics*. Brooks/Cole.

Hamilton, L. C. (1998). *Statistics with Stata 5*. Duxbury Press.

Haughton, D. (1997). Packages for estimating finite mixtures: A review. *The American Statistician*, 51:194–205.

Hilbe, J. (1994). sg16.5: Negative binomial regression. *Stata Technical Bulletin Reprints*, 3:84–88.

Hill, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Applied Statistics*, 23(3):340–354.

Johnson, K. W. and Altman, N. S. (1999). Canonical correspondence analysis as an approximation to gaussian ordination. *Environmetrics*, 10(1):39–52.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Jongman, R. H. G., Ter Braak, C. J. F., and van Tongeren, O. F. R. (1987). *Data analysis in community and landscape ecology*. Pudoc Wageningen.

Magnus, J. R. and Neudecker, H. (1994). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley.

Manly, B. F. J. (1989). *Multivariate statistical methods : a primer*. Chapman and Hall, London.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London, second edition.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press London.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall.

McCune, B. (1997). Influence of noisy environmental data on canonical correspondence analysis. *Ecology*, 78(8):2617–2623.

Michailidis, G. and de Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science*, 13(4):307–336.

Neudecker, H., Satorra, A., and van de Velden, M. (1997). A fundamental matrix result on scaling in multivariate analysis. *Econometric Theory*, 13(6):890.

Neudecker, H., Satorra, A., and van de Velden, M. (1999). A fundamental matrix result on scaling in multivariate analysis. *Econometric Theory*, 15(4):637.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. Toronto Univ. Press.

Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika*, 61(4):559–599.

Noreen, E. W. (1989). *Computer Intensive Methods for Testing Hypotheses. An Introduction*. New York, John Wiley.

Palmer, M. W. (1993). Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology*, 74(8):2215–2230.

Pettitt, A. N. and Stephens, M. A. (1977). The kolmogorov-smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19:205–210.

Puntanen, S. and Styan, G. P. H. (1998). A fundamental matrix result on scaling in multivariate analysis. *Econometric Theory*, 14(5):693–694.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya (series A)*, 26:329–358.

Rao, C. R. and Mitra, S. K. (1971). *Generalized inverse of matrices and its applications*. John Wiley, New York.

Reymen, R. and Jöreskog, K. G. (1993). *Applied Factor Analysis in the Natural Sciencies*. Cambridge University Press.

Rice, J. A. (1995). *Mathematical statistics and data analysis*. Duxbury cop., Belmont (Calif.).

Saporta, G. (1990). *Probabilités Analyse des Données et Statistique*. Éditions technip, Paris.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley and Sons.

Sørensen, J. B. (1999). Zero-inflated poisson and negative binomial regression models. *Stata Technical Bulletin Reprints*, 46:194–199.

Takane, Y. (1991). Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika*, 56(4):667–684.

Tenenhaus, M. (1998). *La Régression PLS*. Éditions Technip, Paris.

Ter Braak, C. J. F. (1985). Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics*, 41:859–873.

Ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179.

Ter Braak, C. J. F. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69:69–77.

Ter Braak, C. J. F. (1988). *Canoco - a Fortran program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1)*. Agricultural Mathematics Group, Wageningen, The Netherlands.

Ter Braak, C. J. F. (1994). Canonical community ordination. part I. basic theory and linear methods. *Écoscience*, 1(2):127–140.

Ter Braak, C. J. F. and Prentice, I. C. (1988). A theory of gradient analysis. *Advances in Ecological Research*, 18:271–317.

Ter Braak, C. J. F. and Šmilauer, P. (1998). *CANOCO Reference Manual and User's Guide to Canoco for Windows, version 4.0*. Centre for Biometry, Wageningen, The Netherlands.

van de Velden, M., Satorra, A., and Neudecker, H. (1999). The justification of multidimensional scaling under euclidean conditions. *Econometric Theory*, 15:153.

van den Wollenberg, A. (1977). Redundancy analysis, an alternative for canonical correlation analysis. *Psychometrika*, 42(2):207–219.

van der Heijden, P. G. M. (1987). *Correspondence analysis of longitudinal categorical data*. DSWO Press, Leiden.

Index