CHAPTER TWO

Correspondence Analysis of Archaeological Abundance Matrices

By Jan de Leeuw

Introduction

+++*Correspondence analysis* (CA) is a technique used to analyze data matrices of non-negative numbers. CA is related to *principal component analysis* (PCA) and *multidimensional scaling* (MDS), that is, it is a form of *proximity analysis*. CA is most frequently applied to rectangular tables of frequencies, also known as *cross tables* or *contingency tables*, although applications to binary incidence or presence-absence matrices are also quite common.

The most often used statistical technique for analyzing cross tables computes and tests some measure of *independence* or *homogeneity*, such as chi-square. In the analysis of independence we investigate whether the body of the table is the product of the marginals. Or, if one prefers an asymmetric formulation, if the rows of the table differ only because they have different row totals (and the columns only differ because they have different column totals).

Pearson's chi-square and related measures quantify how different an observed table is from an expected table, based on the row and column totals. Pearson residuals are used to investigate deviations from independence. CA supplements this classical chi-square analysis because it makes both a *decomposition* and a *graphical representation* of the deviations from independence.

History

CA has a complicated history, both in statistics and in archaeology. The prehistory of CA, starting with work by Pearson around 1900 and ending with the reinvention of the technique by Fisher and Guttman around 1940, is discussed in de Leeuw 1983. The technique has been re-reinvented under many different names, in many different countries, and in many scientific disciplines. New reincarnations still continue to appear, although at a slower pace than before, in the data mining and data analysis literature. Beh 2004 is a recent comprehensive bibliographic review.

The history of CA in archaeology is discussed by Baxter (1994:133–39). Although the literature contains some earlier applications of CA to archaeological examples, the credit for the introduction of the technique to archaeologists usually goes to Bølviken and others (1982). Early applications almost without exception came from archaeologists in Continental Europe, under the influence, no doubt, of the French *analyse des données* school and the leadership of Benzécri (1973a, 1973b). A good overview of these various Continental archaeological applications of CA is found in, for example, Müller and Zimmerman 1997.

It is clear from Baxter's discussion that archaeologists in Continental Europe were ahead of archaeologists in Great Britain, who came on board around 1990. Orton (1999:32), one of the deans of quantitative archaeology in Britain, argues that CA was the most important technique introduced into archaeology in the 1980s. From Britain archaeological CA migrated to the United States, where it arrived shortly before 2000. Duff (1996:90) indicates in an influential article from the mid-1990s that CA was "not well established in Americanist literature." And very recently, Smith and Neiman (2007:55) have concurred: "CA has a long history of use by archeologists in continental Europe but its use by Americanist archeologists is both more recent and rare."

There are several possible reasons why CA did not rapidly become popular in archaeology in Britain and the United States. Most importantly, perhaps, archaeological methodologists tend to look to statisticians for guidance, and in statistics CA was not really known until about 1980, despite the work of Hill (1974). Except in France, of course, but French statistics was relatively isolated from that of the mainstream. The dominant multivariate techniques applied in archaeology were MDS and PCA (sometimes in the disguise of factor analysis). The

most influential work in the area in the seventies was Hodson et al. 1971, which concentrated on the MDS techniques of Boneva, Kendall, and Kruskal. These are all forms of proximity analysis, but they differ from CA in various ways.

In a pioneering article, LeBlanc (1975:22) predicted, "Proximity analysis seems to hold a great deal of promise and will in all probability supplant all other seriation methods." If we interpret this prediction narrowly, in terms of the methods that were available in 1975, it turned out to be incorrect, for reasons that are quite obvious in hindsight. Data, in archaeology and elsewhere, come in many different forms. Sometimes we deal with cross tables, sometimes with incidence matrices, and sometimes with multivariate data that describe archaeological objects in terms of a number of qualitative or quantitative variables. There is no reason to expect that a technique designed for one particular type of data will also work, or even be appropriate, for another type of data. A data analysis technique must obviously take the nature of the data into account, and forcing all data into a common "proximity" format may not be an optimal strategy. But the basic advantages of proximity analysis mentioned by LeBlanc (1975:22) are still very much on target: "In the past, the basic goal of seriation has been to order a series of cultural units on the basis of an assumed single underlying variable, usually time. It is now possible to seriate units according to two or more variables by using a form of proximity analysis or MDS. This increases the power of seriation greatly, and among other advantages, it gives a much better idea of the fit of data to one variable (e.g. time alone) than have previous methods."

Because CA was rediscovered and reintroduced in different countries at different times, most authors in the field of archaeology feel obliged to give some sort of introduction to the technique, even in such recent articles such as Poblome and Groenen 2003 and Smith and Neiman 2007. Our discussion of CA differs in some respects from the ones traditionally encountered in archaeology. In other respects it is quite standard. First, and this is actually quite common, we do not present the technique exclusively as a seriation method. Archaeological sites may be similar or dissimilar for many different reasons, and, to quote Kruskal (1971), "time is not the only dimension." Most CA plots are, of course, two-dimensional maps in the plane, which already suggests that more than one dimension may be relevant. Second, we discuss CA

both as an exploratory technique and as a method of fitting a particular statistical model. And finally, we relate the least squares fitting of the CA model to the maximum likelihood fitting of the exponential distance model (EDM). EDM can be considered to be an alternative, and closely related, form of correspondence analysis.

Types and Attributes

LeBlanc (1975) compares *type seriation* and *attribute seriation* (see also Duff 1996). We can discuss this comparison by distinguishing the different types of data that CA can be applied to. In a CA context, attribute seriation corresponds to multiple correspondence analysis (MCA), which is treated in Gifi 1990:chap. 3, and type seriation corresponds to simple CA, treated in Gifi 1990:chap. 8. Or, to translate this into software, attribute seriation corresponds with the R package homals (de Leeuw and Mair 2009a), while type seriation corresponds with the package *anacor* (de Leeuw and Mair 2009b).

LeBlanc (1975:24) carefully distinguishes the terms *attribute*, *type*, *variable*, and *dimension*. Actually, he uses *variable* and *dimension* interchangeably, but *dimension* is probably best reserved for the axes in multidimensional representations of data. A "variable" is then a formally defined aspect of the group of objects in the study. Each variable is measured in terms of a scale, and the mutually exclusive characteristics of the scale are called "attributes." In the book by Gifi (1990), a variable is defined similarly as a mapping of the objects in a study into the categories of a variable. Defining a number of variables on a set of objects creates, in the terminology of the R software system (R Development Core Team 2007), a "data frame." More specific to archaeology is the notion of a "type," which Leblanc (1975:24) defines as "the existence of a non-random association between the attributes of two or more dimensions." Thus, types are aggregations of attributes over different variables, and consequently, they can be counted more easily and are more susceptible to be treated with frequency-based techniques.

This discussion also makes it possible to compare CA with MDS and PCA. In MDS the first step is usually to derive some symmetric matrix of *similarities* between the sites, assemblages, proveniences, or cultural units. Similarities can be defined in many ways, and often the choice of a particular similarity measure is somewhat arbitrary. Moreover, instead of computing similarities between sites, we could also

decide to compute similarities between the variables describing the artifacts found in sites. A commonly used similarity measure between variables is the correlation coefficient. However, it is unclear how the MDS analysis of the sites and the MDS analysis of the variables are related. In PCA we usually start with a correlation matrix between variables and then derive component loadings to describe the variables and component scores to describe the sites. This means PCA can be used to make a joint plot, also known as a biplot (Gower and Hand 1996). Biplots are compelling ways to visualize multidimensional information, and as such they go beyond simple seriation.

One oft-mentioned disadvantage of PCA is that it assumes linear relations between the variables. This disadvantage, however, no longer applies to modern nonlinear versions of PCA, such as those reviewed in de Leeuw 2006. Moreover, nonlinear PCA and MCA are closely related, so closely, in fact, that nonlinear PCA can be carried out with the MCA package homals (de Leeuw and Mair 2009a).

The CA framework of Gifi 1990 gives one single class of techniques to analyze attribute matrices of artifacts by variables, frequency matrices of types by sites, and incidence matrices of types by sites. It is basically, to use a term from Benzécri's analyse des sonnées, all a matter of "codage." One can code both types and sites as attributes of artifacts, and then the type by site frequency table is just the bivariate cross table of those two variables.

One important advantage of CA and MCA over MDS and PCA is that they stay as close as possible to the original data, no matter if the data are frequencies or incidences or variables with attributes. There is no need to first choose a measure of similarity or correlation, and there is no need to aggregate data into correlation or product matrices. It is true that CA can be presented in terms of a particular measure of dissimilarity, the chi-square distance, and we will give such a presentation in this chapter. But it is only one interpretation of the technique, and the chi-square distances have close connections with the familiar chi-squares that can be computed from the frequencies.

Typical Archaeological Applications

We will discuss some of the typical applications of CA to archaeology in more detail to illustrate where the technique may be appropriate and what archaeologists look at.

Bølviken et al. 1982 uses three data sets from the Stone Age in northern Norway. The first one, from Iversfjord, uses thirty-seven lithic types found in fourteen house site assemblages. Because of interpretational difficulties the analysis was repeated after grouping the thirty-seven types into nine tool categories. The joint plot in two dimensions of the house sites and tool categories is interpreted in terms of economic orientation and settlement permanence. The second example is from the Early Stone Age in the Varanger Fjord area—data counts frequencies of sixteen functional tool types in forty-three sites. Two-dimensional plots give a refinement interpreted in terms of earlier qualitative archaeological hypotheses. The analysis was repeated after the tools were grouped into seven classes, which yielded less informative results. In the third example CA was used to establish a chronology. Data came from a farm mount on the island of Helgøy in Troms. Nineteen classes of artifacts were distributed over fifteen excavation layers, carbon dated from the fourteenth to the nineteenth centuries AD. The analysis shows the layers mapped on a two-dimensional, or *arch*, curve. Projections on the curve can be used to reorder the rows and columns of the data matrix, producing a seriation closely corresponding with the one based on carbon dating.

The article by Duff (1996) on micro-seriation compares attribute and type seriation, following LeBlanc (1975). But whereas LeBlanc used multidimensional scaling for the type seriation, Duff used CA. The data are counts of six ceramic types from forty proveniences in Pueblo de las Muertas, in the Zuni (Cíbola) region of New Mexico, from the thirteenth to the fourteenth century AD. The two-dimensional CA solution exhibits a weak arch, with lots of scatter around it, but produces essentially the same ordering of the units as the MDS analysis of Leblanc.

Early on, Clouse (1999) applied CA to Americanist materials and used it to analyze artifacts found in excavations at the military settlement in Fort Snelling, Minnesota. Sites include eight defense buildings, eleven support buildings, and eight habitation buildings. At all sites artifacts were counted and classified into fourteen groups, such as culinary, armament, commerce, and furniture. Separate abundance matrices are given for defense, support, and habitation buildings, and separate CAs are computed. Both joint plots, showing units and artifact groups in two dimensions, and unit plots, which only show the units, are presented. Groupings of the units conform to what is expected on the basis of the military site model but provide more detailed information. Clouse

(1999:105) argues that CA makes expected and unusual features more clearly visible than the numerical summary given by the table.

The excellent paper by Smith and Neiman (2007) aims to compare frequency seriation, in the tradition of Ford (1952), with CA. They use two cases studies. The first case study is from the Gulf Coast area, near the Chattahoochee and Apalachicola Rivers in Alabama, Georgia, and Florida. Data are from the Middle and Late Woodland periods (100 BC to AD 900). Ceramic data were collected at many sites, of which twenty-nine were selected because they had more than eighty painted sherds. The twenty-nine sites were subdivided into eighty-four assemblages, and the sherds were classified into eighteen pottery types. Obviously, the way in which artifacts and proveniences are grouped into the rows and columns of the table is important for the eventual outcome of the technique, and the CA of the eighty-four assemblages shows a very clear arch pattern, with a clear grouping of sites along the curve. "The CA results confirm what the clean seriation solution suggests: there is no significant source of variation in type frequencies other than time" (Smith and Neiman 2007:61). The analysis was repeated after removing some of the later assemblages. This smaller CA was validated (as a seriation method) by plotting CA scores against radiocarbon dates for selected sites.

The second case study in the Smith and Neiman article is from Kolomoki, a well-researched multimound site in southwestern Georgia, and is an intrasite analysis, not an analysis with multiple sites. The CA uses twenty assemblages and nine pottery types. Separate two-dimensional plots for assemblages and types show no arch, but a significant and interpretable second dimension. The CA solution shows effects, for example spatial ones, not detectable by the inherently one-dimensional frequency seriation. The first CA dimension is again validated as time, using radiocarbon data. We will use the same Kolomoki data set as one of our illustrative examples in this chapter.

Seriation

There is an interesting parallel historical development of what could broadly be called "seriation methods" in psychometrics, ecology, and archaeology. The main steps in this development occured in the same order in each field but at different moments in time, not unlike archaeological artifacts in different sites. Let us look at psychometrics first.

Psychometrics

In the 1940s at the war department, Guttman (1944) discovered scalogram analysis, a method to simultaneously order attitude or achievement items (columns) and respondents (rows) with data in a binary data matrix. Initially scales were constructed by trial-and-error methods, in which row and columns of the binary data matrix were permuted to create the "consecutive ones" property. Normally, we look to order rows and columns in such a way that all ones are next to each other. This result was achieved manually with various ingenious devices. At the same time, the theory for principal components–based computations was already available (Guttman 1941, 1950). In fact, Guttman's (1941) paper was the very first to rigorously define MCA, and he proves that the first MCA dimension provides the consecutive-one ordering for error-free data (1950). The monumental book by Coombs (1964) gives a systematic presentation of these heuristic pencil-and-paper techniques applied to the various data matrices in proximity analysis. And although Coombs's conceptual framework is still relevant, the techniques had already been superseded by computerized methods at the time the book appeared.

Archaeology

Guttman's methods were published around 1950, almost simultaneously with Robinson 1951. To discuss this work, we borrow some terminology from Kendall 1969. An incidence matrix of, say, sites by types is a *Petrie matrix* or *P-matrix* if in each column all ones occur consecutively. A non-negative symmetric matrix is a *Robinson matrix* or *R-matrix* if rows and columns are unimodal and attain their maximal values on the diagonal. By *unimodal* we mean that entries increase to a maximum and then decrease again. Similarities between sites whose incidence matrix is a P-matrix often form an R-matrix. Again, there is an interesting connection with psychometrics here. In the original definition of the Spearman model for general intelligence, dating back to 1904, a battery of tests satisfied the model if their correlation matrix was an R-matrix.

The notion of a P-matrix can be generalized to abundance matrices, that is, to any matrix with non-negative entries. An abundance matrix is a *Q-matrix* if its columns are unimodal. That is the same as saying that the columns of the abundance matrix can be represented as a series of *battleship plots*, as defined in Ford 1952 or Smith and Neiman

2007. Many of the original archaeological seriation techniques take an incidence or abundance matrix and permute the sites in such a way that that it becomes a P-matrix or a Q-matrix. The permutation that is found then orders the sites in time, that is, it is a seriation. Ultimately, however, finding optimal permutations, especially for large matrices, is what is known in computer science as NP-hard, which basically means that the optimization problem, although finite, cannot be solved in a practical amount of time, even by the fastest computers (Arlif 1995).

One way around the impractical computations involved with permutations is to use other related definitions of optimality. As we noted, Guttman already proved in 1950 that CA can be used to find the optimal permutation to a P-matrix in the error-free case (for abundance matrices, see also Gifi 1990:chap. 9 or Schriever 1983). In fact, these publications prove more. They also show that in the error-free case, the second dimension of the CA will be a quadratic function of the first—that plotting the sites in the plane will show a quadratic curve.

Kendall (1971) and others later developed the well-known HORSHU program, which applies MDS to similarities derived from abundance matrices and then derives the order from the projection of the sites on the horseshoe or arch. "We view the arch as a relatively benign indicator that the underlying data do, in fact, contain battleship-shaped curves," write Smith and Neiman (2007:60).

Ecology

In ecology the key concept is that of a "gradient." The emphasis in the data analysis is not on time, as in archaeology, but on environmental characteristics. What is called "seriation" in archaeology is called "ordination" in ecology (Gauch 1982). Plant and animal species do well under certain circumstances and do best at some optimum level of, for example, humidity or altitude. Different species need different altitudes and/or different degrees of humidity. The major advantage of ecology, of course, is that environmental gradients such as altitude can be directly measured, unlike in psychometrics, where aptitude and attitude are theoretical constructs, and archaeology, where direct information about the origin in time of an artifact is usually missing. Ecologists use direct gradient analysis and plot frequencies of species as a function of the gradient. In many cases the result is unimodal distributions, that is, the abundance matrix is a Q-matrix.

Initially, as in psychometrics and archaeology, ordination techniques in ecology required pencil-and-paper methods to reorder the rows and columns of the abundance matrix or of derived similarity matrices with a Robinson structure (Whittaker 1978). These methods changed with the advent of the computer, and, like archaeologists and psychometricians, ecologists turned to PCA and MDS for ordination and to a host of measures of resemblance or similarity.

CA was introduced in ecology by Hill (1974) as "reciprocal averaging." Ter Braak (1985) showed how CA was related to the unimodal response model without going into precise mathematical detail. Ecologists initially were worried about the arch because they considered it an artifact without any empirical significance. We now know more precisely where the arched structures come from, and we know that they indicate strong unidimensional effects (see in particular Schriever 1985 or Van Rijckevorsel 1987). We consequently tend to be pleased if we see a clear arch, especially in archaeology, where we have more reason perhaps to expect unidimensionality. (We will discuss the relationship between unimodal response models, in particular the Gaussian model of Ihm and Van Groenewoud [1975], in more detail when we discuss the exponential distance model.)

Abundance Matrices

We now formalize some of the concepts we mentioned in the introduction. Consider an $r \times c$ table N with *counts*. Rows correspond with r *sites*, columns with c *types*. Frequency indicates how often type j was found in site i. Such a matrix with counts N is called an *abundance matrix*. We also define the row sums and column sums of the table. The *grand total* is the sum of all the counts in the table, which we will also abbreviate simply as n.

It should perhaps be mentioned that *presence-absence matrices* or *incidence matrices* are a special case of abundance matrices in which all entries of the table are either zero or one. An entry merely indicates if a type is present in a site or not, which means our discussion of abundance matrices also covers presence-absence matrices.

A more general type of data matrix is also quite common in archaeology. Suppose the observation unit is an artifact such as a pottery sherd, a piece of obsidian, or maybe a fish bone. The units can be described in

terms of a number of variables that can be either qualitative (categorical) or quantitative (numerical). The abundance matrix is a very special case in which only two categorical variables are used to describe the units, namely *site* and *type*.

The abundance data N can be coded as an $n \times 2$ matrix, where n is the grand total of the table, the first column is *site*, and the second is *type*. The table N is then the *cross table*, or the *contingency table*, of the two variables. But clearly, in a more general case, variables such as size, color, weight, or composition could be used as well. For these more general multivariate data we need a technique such as MCA, also known as *homogeneity analysis* (Gifi 1990; Greenacre and Blasius 2006). Since the data analyzed in this book are all of the simpler bivariate contingency table format, we shall not discuss MCA any further. As we mentioned in the introduction, MCA is the perfect technique for attribute-based seriation in the sense of LeBlanc 1975, in which data are not aggregated to types and assemblages or to counts in a cross table.

Examples

Throughout the chapter we shall use two examples to illustrate the concepts of CA. The first example of an abundance matrix comes from a much larger matrix of sherd counts for sites by pottery types. All samples are from surface collections made around 1940 in Jalisco, Mexico, by Kelly (1945).

This example is not a realistic application of CA because it is too small and too simple. The results of CA do not really add anything to what we can easily see by just looking at the table, but this very fact makes the example useful as an illustration of the basic concepts and calculations (table 2-1).

The second example is pottery data from the Kolomoki burial mounts in Georgia (Pluckhahn 2003; Sears 1956), analyzed previously with CA by Smith and Neiman (2007). We have already discussed these data in the introduction; they include twenty assemblages and nine pottery types.

Associated Matrices

With the abundance matrix we can associate several other matrices. First is the matrix P of *proportions*, whose elements are defined by

Table 2-1: Abundance matrix from Kelly data

		TYPE			
	Site	**AutPol**	**MiReBr**	**AuWhRe**	**AltRed**
21	8	14	0	0	22
34	19	35	0	0	54
23	138	6	0	1	145
37	299	11	0	2	312
9	102	12	22	271	407
7	34	14	59	246	520
	600	92	81	520	1293

Source: Data come from Kelly 1945.

Note: The codes for the types, used as column headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware. Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

$$p_{ij} = \frac{n_{ij}}{n_{\square\square}}.$$

The matrix with proportions shows more clearly how the counts are distributed over the cells. Again, the row marginals are $p_{i\square}$, and the column marginals are $p_{\square j}$ (table 2-2).

Independence

We say that the row variable (site) and the column variable (type) are *independent* if $p_{ij} = p_{i\square} \, p_{\square j}$. Independence can be interpreted to mean that the body of the table does not give additional information, that in fact all the information is contained in the marginals. If we know the relative frequencies of the sites and the types, then we can predict perfectly how many of each type will be in each site.

We measure independence by what is called *inertia* in CA, borrowing a term from physics, and define the table Z of *Pearson residuals* with

Table 2-2: Proportions matrix from Kelly data

	TYPE				
	Site	**AutPol**	**MiReBr**	**AuWhRe**	**AltRed**
21	0.006	0.011	0.000	0.000	0.017
34	0.015	0.027	0.000	0.000	0.041
23	0.107	0.005	0.000	0.001	0.112
37	0.231	0.009	0.000	0.002	0.241
9	0.079	0.009	0.017	0.210	0.315
7	0.026	0.011	0.046	0.190	0.273
	0.464	0.071	0.063	0.402	1.000

Source: Data come from Kelly 1945.
Note: The codes for the types, used as column headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware. Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

$$z_{ij} = \frac{p_{ij} \square p_{i\square}p_{\square j}}{\sqrt{p_{i\square}p_{\square j}}}.$$

The elements of Z show the deviation between the observed proportion and the expected proportion on the hypothesis of independence (corrected for the standard error of the proportion). Positive elements indicate that we see more in the corresponding cell than we expect, and negative elements mean that we see less. The *inertia* is defined simply as

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} z_{ij}^2.$$

In the Kelly (1945) example the inertia is 0.9338, and the Pearson residuals are shown in table 2-3.

Table 2-3: Pearson residuals from Kelly data

	TYPE			
Site	**AutPol**	**MiReBr**	**AuWhRe**	**AltRed**
21	-0.02	+0.28	-0.03	-0.08
34	-0.03	+0.44	-0.05	-0.13
23	+0.24	-0.04	-0.08	-0.21
37	+0.35	-0.07	-0.12	-0.31
9	-0.18	-0.09	-0.02	+0.23
7	-0.28	-0.06	+0.22	+0.24

Source: Data come from Kelly 1945.
Note: The codes for the types, used as column headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware. Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

Table 2-4: Z-scores from Kelly data

	TYPE			
Site	**AutPol**	**MiReBr**	**AuWhRe**	**AltRed**
21	-0.69	+9.94	-1.17	-2.97
34	-1.21	+15.90	-1.83	-4.66
23	+8.62	-1.34	-3.01	-7.50
37	+12.81	-2.38	-4.42	-11.02
9	-6.32	-3.15	-0.69	+8.39
7	-10.14	-2.22	+7.84	+8.73

Source: Data come from Kelly 1945.
Note: The codes for the types, used as column headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware. Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

If the data are a random sample, and if types and sites are independent, then X^2 is distributed as a chi-square random variable with $(r - 1)(c - 1) = 15$ degrees of freedom. In our example, X^2 equals 1207.508. Moreover, each Z_{ij} is approximately standard normal; that is, it is what is commonly known as a z-score, and it can be tested for significance in the usual way. The z-scores are listed in table 2-4.

In the Kelly example the total inertia is clearly far too big, the z-scores are mostly hugely significant, and the two variables *site* and *type* are very far from being independent. Of course, in most archaeological applications data are very far from being a random sample because we generally enumerate and classify all the artifacts found in the site. Nevertheless, we can still take inertia as a guideline to indicate how much structure there is in the data or, more precisely, how much structure there is in the data that cannot be predicted from the marginals.

Conditioning on Rows and Columns

In archaeological studies the hypothesis of independence is not the most natural way to look at abundance matrices. Independence is the appropriate concept if the contingency table results from a random sample from a discrete bivariate distribution, that is, if we sample both sites and types. Usually, however, sites are not sampled. They are fixed either by design or by geographical circumstances.

What really interests us is a comparison of the distribution of types in the different sites that we have selected. Thus, we are mainly interested in comparing the rows of the abundance matrix because each row defines a distribution over types. Fortunately, the hypothesis of homogeneity of rows is mathematically equivalent to the hypothesis of independence. We can most easily see this equivalence by normalizing the rows—dividing each row by its row sum.

To keep our treatment symmetrical, we also consider the case (less common in archaeology) in which it may be interesting or appropriate to also compare the columns. Using the row and column sums, we can normalize the frequency table (or, equivalently, the table with proportions) by dividing the entries of the table by their row or column marginals. This process defines two new tables, the first one conditioned by rows, the second conditioned by columns. The elements are defined by

$$p_{j|i} = \frac{n_{ij}}{n_{i\square}} = \frac{p_{ij}}{p_{i\square}}, \quad \text{and}$$
$$p_{i|j} = \frac{n_{ij}}{n_{\square j}} = \frac{p_{ij}}{p_{\square j}}.$$

The hypothesis of independence can now be written in the two equivalent forms

$$p_{j|i} = p_{\square j} \quad \text{and}$$
$$p_{i|j} = p_{i*},$$

which we can call *homogeneity of rows* and *homogeneity of columns*. Homogeneity of rows says that the probability distribution of types is the same for all sites. Homogeneity of columns says that the probability distribution of sites is the same for all types, which in our context seems a less natural way of expressing the same basic mathematical fact.

Table 2-5 shows the distribution of types over each of the sites and within the last row the distribution of types over all sites, that is, the p_{*j}. We have homogeneity if and only if all rows of the table, including the last row, are the same. Table 2-6 shows the distribution of sites over each of the types and within the last column the distribution of sites over all types, that is, the p_{i*}. We have homogeneity if and only if all columns of the table, including the last column, are the same.

We can define appropriate measures of homogeneity of the rows and columns. These are again called *inertias* in CA, and one inertia exists for each row and each column. They are defined by

$$X^2_{i*} = \sum_{j=1}^{c} \frac{(p_{j|i} - p_{*j})^2}{p_{*j}} \quad \text{and}$$
$$X^2_{*j} = \sum_{i=1}^{r} \frac{(p_{i|j} - p_{i*})^2}{p_{i*}}$$

Rows with a large inertia differ from the average row, that is, the

vector of column marginal proportions. And columns with a large inertia differ from the average column.

Previously, we have defined the *total inertia*. Because of the simple relationship

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} := \sum_{i=1}^{r} p_{i*}\chi^2_{i*} = \sum_{j=!}^{c} p_{*j}\chi^2_{*j}$$

the total inertia is the weighted sum of the row and column inertias.

Under the hypothesis of random sampling from sites and homogeneity of rows, the $n\chi^2_{i*}$ are distributed as chi-squares with *c* - 1 degrees of freedom. If we have random sampling and homogeneity of columns, the $n\chi^2_{*j}$ are distributed as chi-squares with *r* - 1 degrees of freedom.

Table 2-5: Conditioning on the rows in Kelly data

	TYPE				
Site	AutPol	MiReBr	AuWhRe	AltRed	χ^2_{i*}
21	0.36	0.64	0.00	0.00	4.98
34	0.35	0.65	0.00	0.00	0.04
23	0.95	0.04	0.00	0.01	0.11
37	0.96	0.03	0.00	0.01	0.24
9	0.25	0.03	0.05	0.52	0.31
7	0.10	0.04	0.17	0.47	0.27
p_{*j}	0.46	0.07	0.06	0.40	0.93

Source: Data come from Kelly 1945.

Note: The codes for the types, used as column headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware. Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

Table 2-6: Conditioning on the columns in Kelly data

	TYPE				
Site	**AutPol**	MiReBr	AuWhRe	AltRed	p_{i*}
21	0.01	0.15	0.00	0.00	0.02
34	0.03	0.38	0.00	0.00	0.04
23	0.23	0.07	0.00	0.00	0.11
37	0.50	0.12	0.00	0.00	0.24
9	0.17	0.13	0.27	0.52	0.31
7	0.06	0.15	0.73	0.47	0.27
ιX^2_{*j}	0.64	4.06	1.18	0.68	0.93

Source: Data come from Kelly 1945.

Note: The codes for the types, used as column headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware. Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

Exploratory Correspondence Analysis

The basic purpose of exploratory CA is to make a *map of the types* and a *map of the sites*. By a "map" we mean a low-dimensional geometric representation. If we choose dimensionality equal to two, for instance, a map of the types consists of c points in the plane, with one point corresponding to each type. If we choose dimensionality three, then a map of the sites consists of r points in three-dimensional space. Sometimes even a one-dimensional map, which puts all sites on a straight line, is already enough to present the essential information in the table.

The location of the points in the map is not arbitrary, of course. If we make a two-dimensional map of the types, for example, we want the distances between the c points in the plane to be approximately equal to the distances between the c columns of the abundance matrix N. And similarly for the map of the sites and the rows of N.

Distance on the map is defined in the usual way, "as the crow flies." In other words, it is ordinary Euclidean distance. But distance between

columns of the abundance matrix depends on weights that take into account the statistical stability of the cell counts. Specifically, in CA we use *chi-square distances* (Gifi 1990; in de Leeuw and Mair 2009a we use *Benzécri distances* instead). The squared chi-square distance between row *i* and row *k* of table *N* is given by

$$d_{ik}^2 = \sum_{j=1}^{m} \frac{(p_{j|i} - p_{j|k})^2}{p_{\cdot j}},$$

and the squared chi-square distance between column *j* and column *l* of table *N* is

$$d_{jl}^2 = \sum_{i=1}^{n} \frac{(p_{i|j} - p_{i|l})^2}{p_{i\cdot}}.$$

We give the squared chi-square distances for the rows and columns in the Kelly example in tables 2-7 and 2-8.

If we look more closely at table 2-7 we can already predict what CA will do. If we want a geometric representation in which the distances approximate the chi-square distances, then it is pretty clear how such a representation would look. The chi-square distances between sites 21 and 34 and between sites 23 and 37 are almost zero. Thus, in a map sites 21 and 34 will coincide, and sites 23 and 37 will also coincide. Sites 9 and 7 are close as well, and 21/34 is about equally distant from the two groups 7/9 and 23/37. A two-dimensional map will thus look like an isosceles triangle with the three groups of sites at the edges. The shorter side is somewhere around $\sqrt{2}$ or $\sqrt{3}$, and the two longer sides are around $\sqrt{6}$. We also see that it will in general be impossible to map the distance information on a straight line because in that case we would have to let 7/9 coincide with 23/37. In this small example we can easily see what a map would look like, but in a larger example, such as the Kolomoki one, this map becomes much more complicated. That is why we have CA, which approximates the chi-square distances to the Euclidean distances in a precise way on the map.

In CA we approximate chi-square distances *from below*. Let me explain this concept. In any CA map of the sites, for instance, we will

Table 2-7: Squared Benzécri distances, rows (sites)

	21	**34**	**23**	**37**	**9**	**7**
21	0.000					
34	0.002	0.000				
23	5.721	5.950	0.000			
37	5.841	6.072	0.001	0.000		
9	6.353	6.550	2.188	2.208	0.000	
7	6.812	6.999	3.207	3.233	0.259	0.000

Source: Data come from Kelly 1945.
Note: Site 21 (Cofradía No. 1) and Site 34 (Hacienda Nueva) are included in the Cofradía Complex (early); Site 23 (Cofradía No. 3) and Site 37 (Amilpa) are included in the Mylpa Complex (intermediate); and Site 9 (Altillos) and Site 7 (Mezquitlan) are included in the Autlan Complex (late).

Table 2-8: Squared Benzécri distances, columns (types)

	AutPol	**MiReBr**	**AuWhRe**	**AltRed**
AutPol	0.000			
MiReBr	4.921	0.000		
AuWhRe	3.221	6.203	0.000	
AltRed	2.539	5.780	0.436	0.000

Source: Data come from Kelly 1945.
Note: The codes for the types, used as column and row headers, are AutPol for Autlan Polychrome, MiReBr for Miscellaneous Red on Brown/Buff, AuWhRe for Autlan White on Red, and AltRed for Altillos Red Ware.

always have $d_{ik} \square d_{ik}$, where d_{ik} is Euclidean distance between points *i* and *k* on the map. More precisely, CA constructs a sequence of maps: the first one has only one dimension, the second has two, and so on. The final map has $t = \min(r - 1, c - 1)$ dimensions, that is, three in the Kelly example and eight in the Kolomoki example. The maps are *nested* in the sense that the projection on the first dimension of all the maps is identical to the one-dimensional map, the projection on the plane of the

first two dimensions of all maps with at least two dimensions is equal to the two-dimensional map, and so on. If $d_{ik}^{(s)}$ represents the distances on the *s*-dimensional map, with $1 \leq s \leq t$, then

$$d_{ik}^{(1)} \leq d_{ik}^{(2)} \leq \ldots \leq d_{ik}^{(t)} = d_{ik}.$$

Thus, the *t*-dimensional map has distances exactly equal to the chi-square distances. Maps in fewer dimensions approximate the distances, and the approximation becomes better for each of the distances when the dimensionality increases. Approximation is from below because map distances are always smaller than chi-square distances, no matter what the dimensionality of the map is. Of course the same reasoning applies to chi-square distances between columns and the CA map for types.

The map does not only approximate chi-square distances between sites or types, it also approximates the inertias of the sites and the types. In the sites map, for instance, the inertia is approximated (from below, as usual) by the distance of the site to the origin of the map. Or, equivalently, by the length of the vector corresponding with the site. This means that a site that differs little from the average site, and thus has a small inertia, will be close to the origin of the map. And sites that are different from the others will be tend to be in the periphery of the map. As a consequence, the center of the map, the area near the origin, can quite easily be cluttered with sites that are similar to the average site.

A CA program (we use *anacor* by de Leeuw and Mair [2009b]) typically takes the abundance matrix and the desired dimensionality of the map as its arguments. It then outputs coordinates for the maps of the row objects (sites) and the column objects (types). In addition it can provide a variety of plots, and it provides a *decomposition of the inertia*. This type of decomposition is familiar from PCA. Consider the weighted squared length of the projections of the site points on the first dimension, on the second dimension, and so on. This decomposes the total inertia of the vectors into a component due to the first dimension, the second dimension, and so on. By dividing the components by the total, we can say that a certain percentage of the inertia is "explained" by the first dimension, another smaller percentage by the second dimension, and so on. Ultimately, there are $t = \min(r - 1, c - 1)$ dimensions, and each of them takes care of a certain decreasing percentage of the total inertia.

CA can also make *joint maps* or *biplots* in which we basically take the site plot and the type plot and put them on top of each other. We then have a plot in which types will tend to be close to sites in which they occur more frequently than one would expect on the basis of the marginals. We say "tend to" because there is no chi-square distance defined between a site and a type, and thus there is no approximation in some well-defined mathematical sense. The CA program *anacor* basically lets the user make four choices for the joint plot. The first one is to put the two Benzécri plots on top of each other. Distances between sites and distances between types approximate chi-square distances, but distances between sites and types have no simple relation to the data. The second option, which is called Goodman scaling in the program, is to adjust the length of the site and type vectors in such a way that their inner product approximates the Pearson residual. Unfortunately, this result invalidates the interpretation of site and type distances as approximations of chi-square distances. The last two options use the *centroid principle*. We can take the Benzécri map for the sites and then plot the types by taking weighted averages (centroids) of the sites using the frequencies of the types in those sites as weights. This process produces a joint plot in which site distances approximate chi-square distances. The locations of the types in the plot again only differ in vector length from the locations in the Benzécri type plot. Type distances cannot be interpreted as approximating chi-square distances between types anymore, but they do have a clear geometric interpretation as weighted averages of site points. By symmetry, there is a second centroid principle in which we use the Benzécri type plot and then plot the sites as weighted averages of types.

The centroid principle can also be used to fit passive sites or types into the plots. Suppose an additional site, not used in the analysis, is excavated, and the objects are classified using the same typology as the one used in the analysis. The type scores from the analysis can be used to compute the score for this new additional site just by calculating the average CA score of the site on each of the dimensions. In the same way we could use the site scores to add additional types to the analysis, for example if we decided to split one original type into two new types. Of course, the alternative is to repeat the CA with the additional sites and types, which then would actively determine the overall CA solution.

The Kelly Example

Let us illustrate exploratory CA with the small Kelly example. The two-dimensional maps for sites and types from CA are shown in figure 2-1.

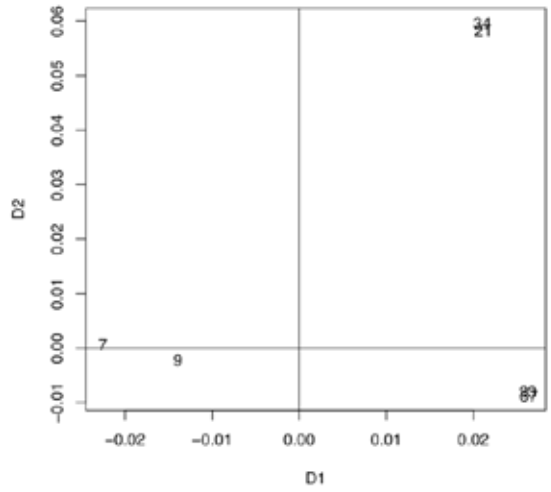
As expected, in the sites map we see the three clusters of points at the vertices of a triangle, and as we know, the one-dimensional map is simply the projection of all points on the horizontal axis.

In figure 2-2a we see the approximation of the chi-square distances between sites in one dimension and in Figure 2-2b in two dimensions. Chi-square distances are on the horizontal axis, Euclidean map distances on the vertical axis. Approximation from below means that all points are below the 45-degree line of perfect fit. But as we can see, fit in two dimensions is already almost perfect. In one dimension some of the larger chi-square distances, in particular those between 21/34 and 23/37 are seriously underestimated.

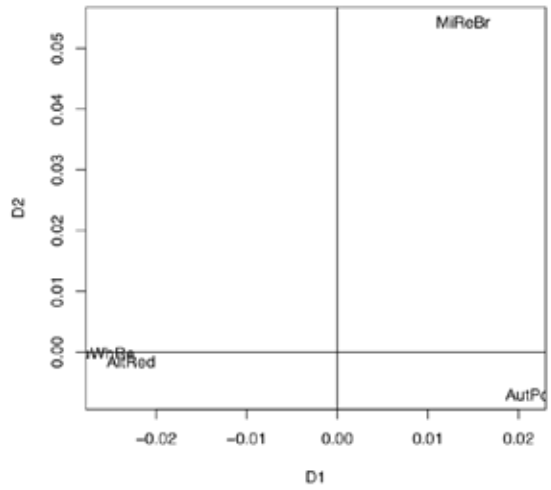
We finally show the chi-square decomposition for the Kelly example (table 2-9). Not surprisingly, the two first dimensions account for 97 percent of the total inertia, and the third dimension is of very little importance.

The Kolomoki Example

We now apply CA to the Kolomoki data, our more realistic example. The chi-square decomposition is given in table 2-10. Two dimensions account for 80 percent of the inertia, three dimensions for almost 90 percent. The CA maps for the types in two and three dimensions are given in figures 2-3 and 2-4. Again, the two-dimensional map is just the projection of the three-dimensional map onto the horizontal plane (except for a possible rotation). Note that the points in the two-dimensional maps are the centers of ellipses of varying sizes. These ellipses represent 95 percent confidence regions for the points. Confidence region computations, which are shown in de Leeuw and Mair 2009b, are based on the assumption that the abundances are a large random sample from a population. As with chi-squares, this assumption may not be appropriate in archaeological examples, but also as with chi-squares, the sizes of the ellipses do give a useful representation of variability. Without the random sampling assumption they measure how the location of the points in the plot changes with small perturbations of the data. We see larger ellipses for outlying points, which generally



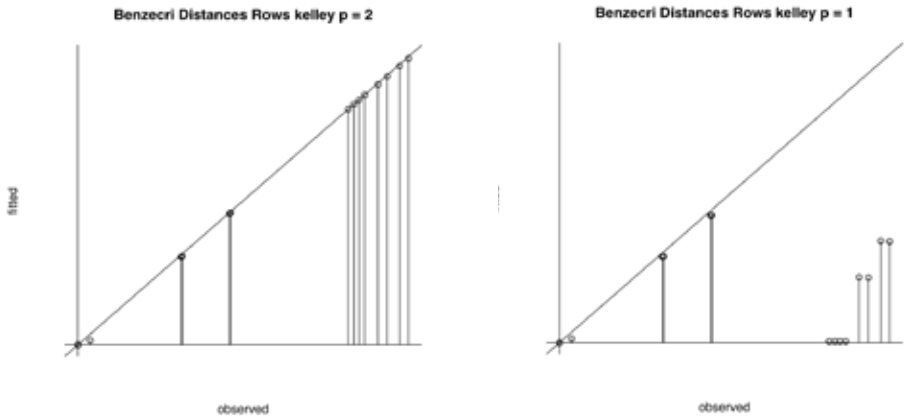
(a) Sites



(b) Types

Figure 2-1. A two-dimensional CA map of the Kelly data.

(a) One Dimension



(b) Two Dimensions

Figure 2-2. An approximation of the Benzécri distances for the Kelly data.

Table 2-9: Chi-square decomposition of Kelly data

	X^2	%	Cum %
1	787.9	0.65	0.65
2	390.0	0.32	0.97
3	29.6	0.03	1.00
Total	**1207.5**		

Source: Data come from Kelly 1945.

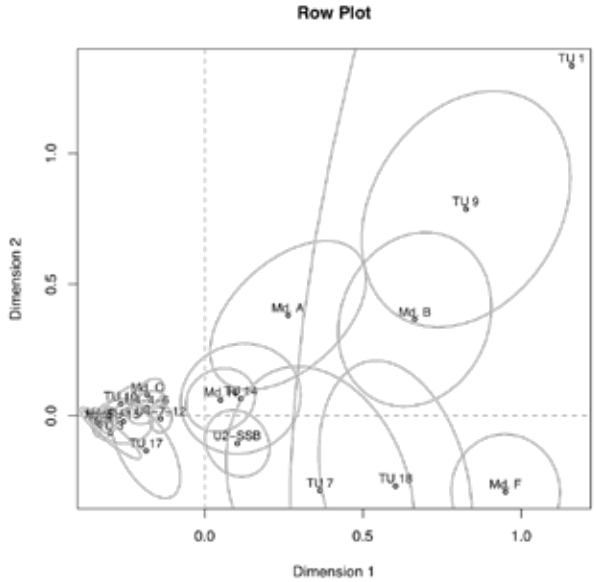
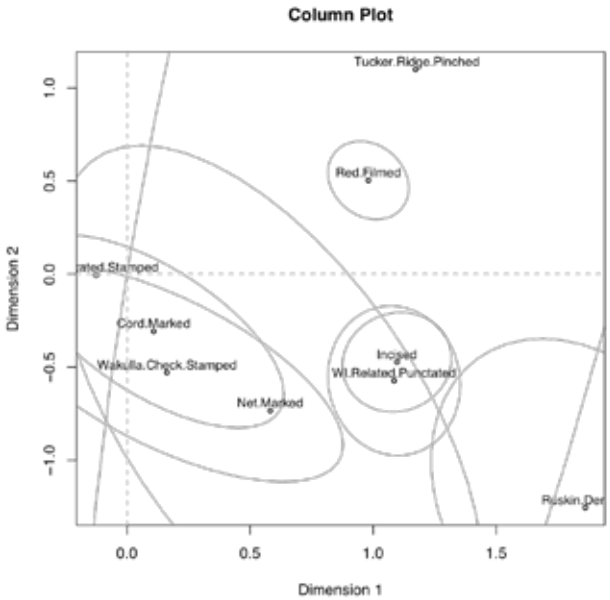


Figure 2-3. CA maps of the Kolomoki data.



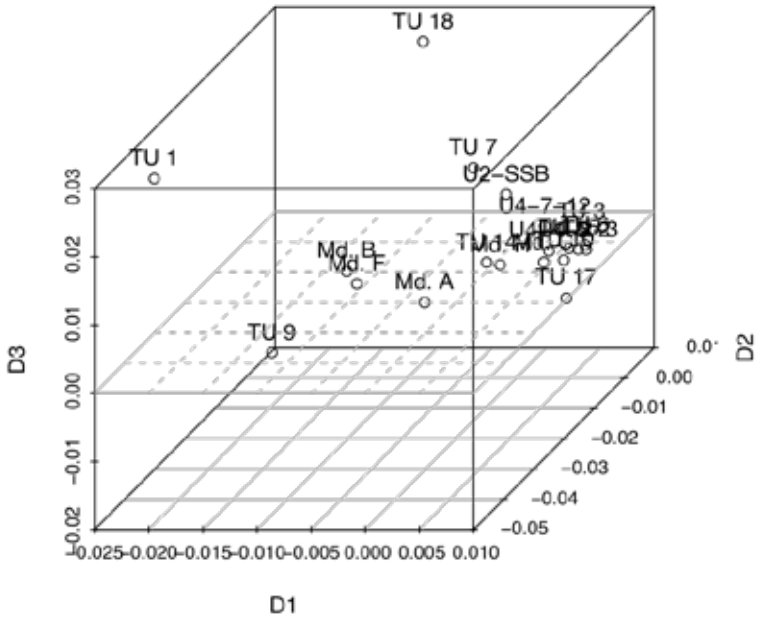
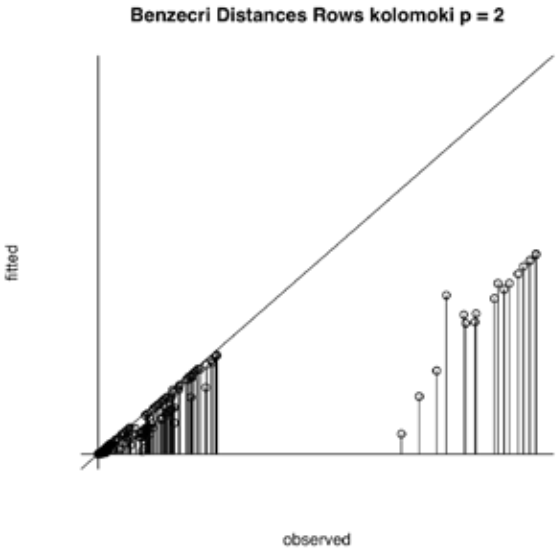


Figure 2-4. A three-dimensional map of the Kolomoki data.

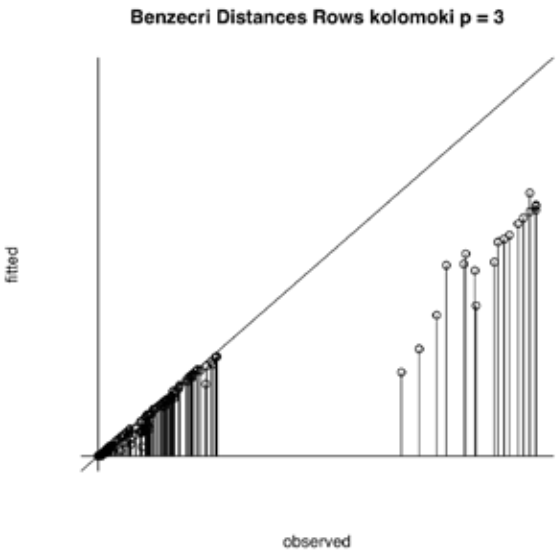
correspond with smaller abundances, and we see examples of overlapping ellipses for sites or types that cannot really be distinguished.

For the interpretation of the two-dimensional Kolomoki results, we refer to the experts Smith and Neiman (2007). The third dimension does not add much (only 9 percent of the total inertia), but it does allow us to better approximate some of the larger chi-square distances. In particular, the third dimension emphasizes the differences between the outliers T9 and T1/T18.

If we continue to add dimensions, we will probably see each new dimension take care of a group of the large chi-square distances, which are still seriously underestimated in three dimensions. See the Benzécri plots in figure 2-5.



(a) Two Dimensions



(b) Three Dimensions

Figure 2-5. An approximation of the Benzécri distances for the Kolomoki data.

Table 2-10: Chi-square decomposition of Kolomoki data

	X^2	%	Cum %
1	1018.8	0.63	0.63
2	261.6	0.16	0.79
3	144.7	0.09	0.88
4	128.0	0.08	0.96
5	38.6	0.02	0.98
6	17.9	0.01	0.99
7	9.0	0.01	1.00
8	3.8	0.00	1.00
Total	**1622.5**		

Source: Data come from Kelly 1945.

Variations of Correspondence Analysis

Several variations of CA are also used. We have not applied them in our example, but we briefly mention them for completeness. One may wonder, for example, if approximation from below is such a good idea. It seems obvious that a better approximation of the chi-square distances is possible if we allow some of the map distances to overestimate and others to underestimate. This idea is exploited in de Leeuw and Meulman 1986. The idea, basically, is to compute chi-square distances first and then apply multidimensional scaling to these distances.

A second question is whether there are suitable alternatives to the chi-square distances. Remember that chi-square distances are used because we correct the proportions for their standard errors, on the assumption of independence. Chi-square distances have a natural connection to chi-square, to the weighted sum of squares, and thus to Euclidean distance. Alternative methods for weighting the proportions are indeed possible, as in the spherical CA of Domingues and Volle (1980), but generally the connection with Euclidean geometry becomes less transparent.

And finally, we can get away from the interpretation of abundance matrices in terms of relative frequencies. Instead, we can think of them

as *compositional data*. Each row is a vector of proportions adding up to one, but the proportions may come from a chemical analysis of samples and may not come from counts. Compositional data are very common in chemometrics and the earth sciences and are also quite common in archaeology. Variations of principal component analysis for compositional data similar to, but not identical with, CA are discussed in the monograph by Aitchison (2003).

Exponential Distance Models

In ecology (Ihm and Van Groenewoud 1975; Ter Braak 1985), and to some extent in archaeology, much attention has been paid to the Gaussian ordination model (GOM). The model says that for site *i* and species *j* the expected value of the abundance is

$$E(f_{ij}) = a_i b_j \exp\left(-\frac{1}{2}\left(\frac{x_i - y_j}{s_i}\right)^2\right).$$

Thus, sites and types can be scaled on a common one-dimensional scale. Abundance f_{ij} is, except for the marginal row and column effects a_i and b_j, related to the distance between the scale value of site *i* and the scale value of type *j*. More precisely, a type will be abundant in sites whose scale value is close to the type's scale value, and it will be largest if type and site coincide on the scale. Rows of the abundance matrix will be unimodal: they have a single peak and then level off in both directions. Or, to use Kendall's terminology, they are Q-matrices. Again, except for the marginal effects, the same thing is true for the columns. Thus, if the model fits we can reorder the sites and types in such a way that both rows and columns of the abundance matrix are unimodal.

The GOM can be generalized easily to more than one dimension.

$$E(f_{ij}) = a_i b_j \exp\left(-\frac{1}{2}\sum_{s=1}^{p}(x_{is} - y_{js})^2\right).$$

For obvious reasons we call this the exponential distance model

(EDM). The EDM is unimodal in a more general geometrical sense. The response curves in the plane, if $p = 2$, have a single peak and level off in all directions. There are many ways in which the EDM can be fitted to abundance matrices. Most of them are based on multinomial maximum likelihood, and thus they naturally come with large-sample significance tests and confidence regions. Not surprisingly, contributions have been made by both psychometricians and ecologists. For a recently proposed technique, and a good overview of earlier work, see De Rooij and Heiser 2005.

We can simplify the EDM, by expanding the square and collecting terms, to the equivalent form

$$E(f_{ij}) = a_i b_j \exp(\sum_{s=1}^{p} x_{is} y_{js}).$$

This equation shows how we expand the abundances into the product of marginal effects and an interaction term, which is an inner product of row and column effects and is actually quite close to CA. In the social sciences this is often referred to as the row-column or RC-model. For small arguments we have $\exp(x) \approx 1 + x$ and consequently

$$E(f_{ij}) \approx a_i b_j (1 + \sum_{s=1}^{p} x_{is} y_{js}).$$

This model is fitted by CA, using weighted least squares. Thus, we see that CA can be interpreted as a convenient and inexpensive approximation to EDM but also as a model in its own right in which the multiplicative (exponential) interactions are replaced by additive ones. Besides this relationship, of course, both EDM and CA can be discussed as data reduction and data representation methods, without necessarily referring to a statistical model.

The two-dimensional Kolomoki EDM solution is given in figure 2-6. We will not give an interpretation of the result but merely point out that there are some differences from the CA solution. The grouping of sites and types is approximately the same, but the EDM solution displays less of an arch, and this is typical.

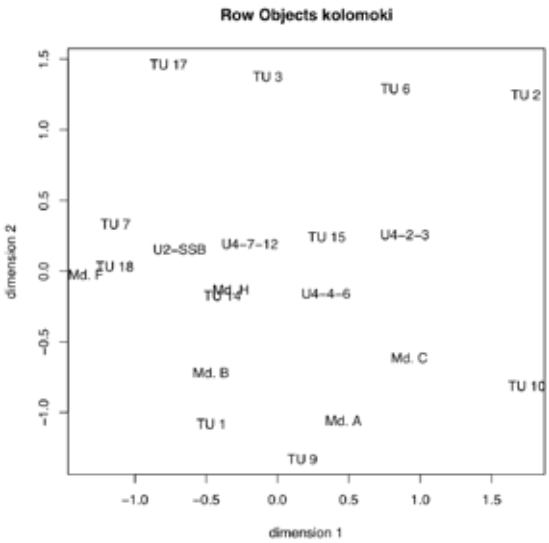
Discussion

This chapter could be called "The Many Faces of Correspondence Analysis," and in it we have tried to provide various interpretational frameworks to look at CA plots in terms of distances, centroids, association models, and chi-square. It also shows how the same models and techniques appear in many different disciplines, often under different names, and that combining ideas from these disciplines gives us additional possibilities for interpretation.

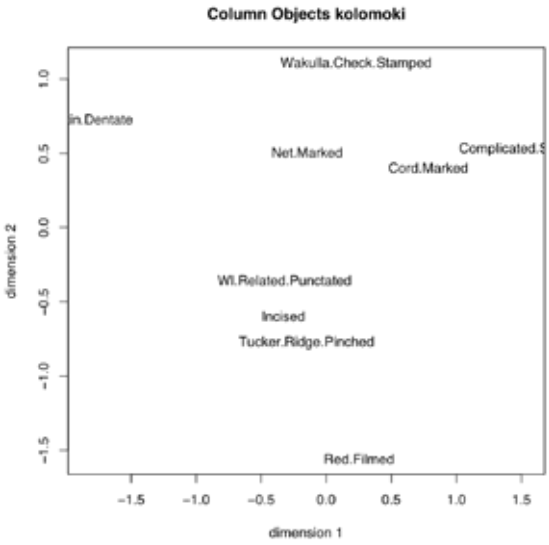
We have also discussed the EDM model in its various disguises as the GOM or the RC-model. It can be used to embed a form of CA into a maximum likelihood framework and to shift the emphasis from multivariate exploration to model testing.

Archaeologists not familiar with CA can use this chapter to look at previous examples in their discipline and to think in a different way about abundance and incidence matrices. We have tried to emphasize the continuity between CA and previous seriation methods used in archaeology.

As we have indicated, convenient, free R packages are available for CA. We mentioned homals and *anacor*, but in de Leeuw and Mair 2009b other available packages are discussed as well. All standard statistical systems, such as SAS, SPSS, and Stata, also have CA methods as either built-ins or add-ons.



(a) Rows



(b) Columns

Figure 2-6. EDM maps of the Kolomoki data.