Topics in correspondence analysis
van de Velden, M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Chapter 1

Introduction

This book is concerned with the multivariate method of correspondence analysis. Correspondence analysis is primarily a graphical method for inspecting multivariate count data. Typically correspondence analysis is applied to a so-called contingency matrix: a matrix of frequencies of co-occurrences between two categorical variables. However, extensions and adaptations of the method exist that make it possible to apply correspondence analysis to other types of data; for instance, multiple-choice and preference data. In this book several topics in correspondence analysis, both theoretical and applied, will be described.

The name correspondence analysis is derived from the French term “analyse des correspondances”. The method was developed in France in the early 1960s by a group of data analysts led by J.-P. Benzécri. Numerous theoretical results as well as applications were published over the years in the journal Les Cahiers de l'Analyse des Données. In their exposition of the method a large emphasis was placed on geometrical concepts. Since the 1980s the method has considerably gained in popularity outside the French speaking world. The publication of two English textbooks on correspondence analysis, i.e. Lebart et al. (1984), and Greenacre (1984), have been of great importance in this respect.

There exist several methods that are mathematically equivalent to correspondence analysis, e.g. reciprocal averaging, homogeneity analysis, and dual scaling. Differences between the methods are only superficial and the result of underlying rationales. Both Nishisato (1980) and Greenacre (1984) give an extensive historical overview tracing the origins of these methods back to several apparently independent beginnings: Richardson
Chapter 1. Introduction

and Kuder (1933), Hirschfeld (1935), Horst (1935) and Fisher (1940). It is because of
the independent beginnings, as well as the rediscovery of the method over the years by
different authors from different nationalities with different scientific backgrounds, that
there are so many names\(^1\) for what is essentially the same method. In this thesis no
attempt is made to describe any of the mathematically equivalent methods in detail. For
a description of reciprocal averaging, a term first proposed by Horst (1935), we refer
to Hill (1974). (It should be noted, however, that Hill, 1974, refers to the method as
"correspondence analysis: a neglected multivariate method".) Greenacre (1984, p 96-
102) provides a similar exposition of reciprocal averaging and contrasts it with what
is now more commonly regarded as correspondence analysis. Homogeneity analysis is a
method developed in the Netherlands, in particular in Leiden. Gifi (1990) gives a thorough
overview of the method as well as its relationship with various other multivariate methods.
Dual scaling is based on the optimal scaling approach introduced by Guttman (1941). A
comprehensive description of dual scaling can be found in Nishisato (1980, 1994). A
comparison with correspondence analysis can be found in Greenacre (1984).

In addition to the foregoing, it has been shown that correspondence analysis is tightly
connected with several other, well-known multivariate analysis methods. In particular,
canonical correlation analysis (Greenacre, 1984 and Lebart et al., 1984), discriminant
analysis (Lebart et al., 1984), principal component and principal coordinate analysis
(Greenacre, 1984 and Gower and Hand, 1996), factor analysis (Greenacre, 1988 and Boik,
1996), biplots (Greenacre, 1993a and Gower and Hand, 1996) and (weighted) metric
multidimensional scaling (Gifi, 1990 and Cuadras and Fortiana, 1995). Some of these
relationships will, to a certain extent, be exploited in this thesis. For a complete treat-
ment of the relationships, however, we refer the interested reader to the references given
in parentheses.

Applications of correspondence analysis cover a wide range of scientific disciplines and
topics. In a recent article by Volpato and Contarello (1999), for example, correspondence
analysis is used to study social relationships in the extreme situation of the concentration
camp. The authors base their analysis on the famous novel *If This is a Man* by Primo Levi

\(^1\)Nishisato (1980, 1994) gives 12 alternative names for dual scaling.
(1947), written just after his release from Auschwitz. Albers and Keizer (1990) apply correspondence analysis in their study of orthomanual medicine. Greenacre and Vrba (1983) analyze antelope census data in African wildlife areas, and Anna Torres-Lacomba (1999) applies correspondence analysis to beer brands and characteristics associated with these brands. These are just a few examples indicating the multidisciplinary character of the method. Greenacre (1984) provides an extensive overview of applications, sorted by research field. Other sources containing several recent applications of correspondence analysis are Greenacre and Blasius (1994) and Blasius and Greenacre (1998).

Outline

In Chapter 2 we describe two approaches to correspondence analysis. For this purpose we introduce notation that will be employed throughout the thesis. The first approach is based on the close relationship between correspondence analysis and component analysis. In section 2.3 we will, using this relationship, derive equations that are of crucial importance in correspondence analysis. These equations are the mathematical backbone of correspondence analysis and they will frequently be referred to in other chapters. It should be noted that some recent textbooks on multivariate analysis, e.g. Johnson and Wichern (1998), describe correspondence analysis in a similar fashion.

The second approach to correspondence analysis, which is given in section 2.4, stays closer to the usual geometrical approach as described by Greenacre (1984) and Greenacre and Hastie (1987). The rows and columns of the contingency matrix are approximated in a least-squares sense through minimization of the Mahalanobis distance. As we will show, assuming multinomially distributed rows and columns, this approach is identical to the usual correspondence analysis.

Some specific correspondence analysis terms such as chi-squared distance and inertia will also be introduced and clarified in Chapter 2. The relationship between correspondence analysis and a biplot is described in subsection 2.3.1. We conclude the chapter with an example based on experimental data.

Chapter 3 is based on a paper by van de Velden and Neudecker (2001). This chapter deals with a specific eigenvalue property encountered in correspondence analysis. Al-
though the property in question is well-known, a clear algebraic proof was not available. In this chapter an overview of existing proofs is presented along with a selection of new proofs. The existing proofs have in common that they do not solve the problem directly, but instead rely on some method-specific features. Moreover, they are quite lengthy. Proofs that are not related to any particular method but only rely on matrix algebra are given in subsections 3.3.4 and 3.3.5. With the exception of the proof given in subsection 3.3.4 these method-unrelated proofs require rather advanced matrix algebra.

The analysis of preference data, in particular so-called paired comparison data, is the topic of Chapter 4. In marketing research one is often interested in analyzing such data. Consider for example the analysis of consumer preferences with respect to a certain product. As will be shown in this chapter, which is based on van de Velden (2000a) and van de Velden (2000b), correspondence analysis provides a tool for the analysis of preferences that does not require any distributional assumptions. However, as correspondence analysis is typically applied to a contingency matrix, some adaptations need to be made in order to apply the method to preference data.

Preference data can be coded and collected in several ways. Greenacre (1984, 1993b), for example, describes the analysis of so-called rating data. The analysis of paired comparison data, however, is not explicitly treated by Greenacre. Nishisato (1980, 1994) in his description of dual scaling does treat the analysis of paired comparisons. Moreover, he claims that one of the main advantages of dual scaling over correspondence analysis is the wider applicability of dual scaling. In Chapter 4, we introduce two approaches for applying correspondence analysis to paired comparison data. It will be shown that one of these approaches is equivalent to the dual scaling approach. The other approach—which is more in line with the treatment of rating and rank order data in correspondence analysis as described by Benzécri (1973) and Greenacre (1984, 1993b)—gives different results. Using optimal scaling arguments similar to those introduced by Guttman (1946) the differences between the two approaches will be clarified.

Correspondence analysis of data on more than two categorical variables is the topic of Chapter 5. As one often encounters data on more than two categorical variables the analysis of such data is of great practical interest. Multiple-choice data, for example, can
often be considered as data on several categorical variables. The questions then represent the variables, and the answering options are the categories. By coding the data in a specific way, i.e. in the format of a so-called indicator matrix, correspondence analysis can be applied to data on more than two categorical variables. This approach is called multiple correspondence analysis. There are some important problems associated with this approach. In particular, the use of geometrical concepts that are essential in the geometrical approach to correspondence analysis become rather cumbersome. Greenacre (1988) recognizes these problems and proposes an alternative approach which he calls joint correspondence analysis.

Both multiple correspondence analysis and joint correspondence analysis will be treated in this chapter. These methods are closely related to other well-known multivariate methods. In particular, multiple correspondence analysis is closely related to principal component analysis, and joint correspondence analysis is related to (principal) factor analysis. These relationships will be described in detail. Moreover, we will show that in the case of the analysis of two categorical variables, multiple correspondence analysis is closely related to, and joint correspondence analysis is equivalent to correspondence analysis of the corresponding contingency matrix. To illustrate the two approaches an example concerning women’s ‘shopping-for-clothing’ attitudes will be presented.

The relationships between correspondence analysis and principal component and factor analysis will also be of importance in Chapter 6. In that chapter, which is based on joint work with Henk Kiers, we consider a rotation of the correspondence analysis approximation. Rotation in correspondence analysis is particularly useful if the dimensionality of the approximation is greater than two. Similarly to rotation in principal component and factor analysis, the correspondence analysis approximation (or, more precisely, the correspondence analysis biplot) may be rotated to simple structure. There are several issues that arise naturally in this context. Firstly, we need to define simple structure of the correspondence analysis solution. For this purpose we will introduce the well-known varimax criterion. Secondly, as correspondence analysis yields several biplots, an appropriate biplot must be selected. Finally, as rotation in correspondence analysis concerns both the row and column coordinates, we must either choose one set of coordinates that
will be rotated to simple structure, or we must introduce a procedure for rotating both sets simultaneously to simple structure.

In order to interpret both the original and rotated correspondence analysis approximations we will use a decomposition of inertia similar to a common decomposition of variance. Especially in the case where the dimensionality of the correspondence analysis approximation is greater than two it is common to assign labels to axes that facilitate interpretation. These labels are assigned by using the decomposition of inertia. In section 6.7 the decomposition of inertia before and after rotation will be described. To illustrate the usefulness of rotation the chapter concludes with two examples.