Deconstructing Multiple Correspondence Analysis

Jan de Leeuw

First created on March 28, 2022. Last update on May 07, 2022

Abstract

This chapter has two parts. In the first part we review the history of Multiple Correspondence Analysis (MCA) and Reciprocal Averaging Analysis (RAA). Specifically we comment on the 1950 exchange between Burt and Guttman about MCA, and the distinction between scale analysis and factor analysis. In the second part of the chapter we construct an MCA alternative, called Deconstructed Correspondence Analysis (DCA), which is useful in the discussion of "dimensionality", "variance explained", and the "Guttman effect" that were important in the history of the first part.

1 Notation

Let us start by fixing some of the notation used in this paper. We have $i = 1, \cdots n$ observations on each of $j = 1, \cdots, m$ categorical variables. Variable j has k_j categories. We use $k_\star$ for the sum of the k_j, while the maximum number of categories over all variables is k_+. We also define m_s, with $s = 1, \cdots, k_+$, where m_s is the number of variables with $k_j \geq s$. Thus both m_1 and m_2 are always equal to m. Also $\sum_{s=1}^{k_+} m_s = k_\star$. The fact that variables can have a different number of categories is a major notational nuisance. If they all have the same number of categories k then $k_+ = k$, $k_\star = mk$, and all m_s are equal to m.

The data are coded as m indicator matrices G_j, with $\{G_j\}_{ik} = 1$ if and only if object i is in category k of variable j and $\{G_j\}_{ik} = 0$ otherwise. The G_j are $n \times k_j$, zero-one, and columnwise orthogonal (because categories are exclusive). If we concatenate the G_j horizontally we have the $n \times k_\star$ matrix G, which we also call the indicator matrix (in French data analysis it is the "tableau disjonctif complet", in Nishisato (1980) it is the "response-pattern table"). The Burt table ("tableau de Burt"), is the $k_\star \times k_\star$ cross product matrix $C := G'G$. The univariate marginals are in the diagonal matrix $D := \text{diag}(C)$. The normalized Burt table is the matrix $E := m^{\frac{1}{2}} D^{-\frac{1}{2}} C D^{-\frac{1}{2}}$. Symbol $:=$ is used for definitions.

Although we introduced G, C, D and E as partitioned matrices of real numbers, it is also useful to think of them as matrices with matrices as elements. Thus C, for example, is an $m \times m$ matrix with as elements the matrices $C_{j\ell} = n^{-1} G_j' G_\ell$, and G is an $1 \times m$ matrix with as its m elements G_j. Note that because we have divided the cross product by n, all $C_{j\ell}$, and thus all $D_j = C_{jj}$, add up to one.

In the paper we often use the direct sum of matrices. If A and B are matrices, then their direct sum is

$$A \bigoplus B := \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \tag{1}$$

and if A_r are s matrices, then $\bigoplus_{r=1}^{s} A_r$ is block-diagonal with the A_r as diagonal submatrices.

2 Introduction

Multiple Correspondence Analysis (MCA) can be introduced in many different ways.

Mathematically MCA is the Singular Value Decomposition (SVD) of $m^{-\frac{1}{2}}Gy = \sqrt{\lambda}x$ and $m^{-\frac{1}{2}}G'x = \sqrt{\lambda}Dy$, the eigenvalue problem $Ey = \lambda^2 y$ for the normalized Burt table, and the Eigenvalue Decomposition (EVD) of $m^{-1}G'D^{-1}Gx = \lambda^2 x$ for the average projector. Using m in the equations seems superfluous, but it guarantees that $0 \leq \lambda \leq 1$.

Statistically MCA is a scoring method that minimizes the within-individual and maximizes the between-individuals variance, it is a graphical biplot method that minimizes the distances between individuals and the categories of the variables they score in, it is an optimal scaling method that maximizes the largest eigenvalue of the correlation matrix of the transformed variables, and that linearizes the average regression of one variable with all the others. It can also be presented as a special case of Homogeneity Analysis, Correspondence Analysis, and Generalized Canonical Correlation Analysis. See, for example, the review article by Tenenhaus and Young (1985).

It is of some interest to trace the origins of these various MCA formulations, and to relate them to an interesting exchange in the 1950's between two of the giants of psychometrics on whose proverbial shoulders we still stand. In 1950 Sir Cyril Burt published, in his very own British Journal of Statistical Psychology, a great article introducing MCA as a form of factor analysis of qualitative data (Burt (1950)). There are no references in the paper to earlier occurances of MCA in the literature. This prompted Louis Guttman to point out in a subsequent issue of the same journal that the relevant equations were already presented in great detail in Guttman (1941). Guttman assumed Burt had not seen the monograph (Horst (1941)) in which his chapter was published, because of the communication problems during the war, which caused "only a handful of copies to reach Europe" (Guttman (1953)). Although the equations and computations given by both Burt and Guttman were identical, Guttman pointed out differences in interpretation between their two approaches. These differences will especially interest us in the present paper. They were also discussed in Burt's reaction to Guttman's note (Burt (1953)). The three papers are still very readable and instructive, and in the first part of the present paper we'll put them in an historical context.

3 History

3.1 Prehistory

The history of MCA has been reviewed in De Leeuw (1973), Benzécri (1977a), Nishisato (1980) (Section 1.2), Tenenhaus and Young (1985), Gower (1990), and Lebart and Saporta (2014), each from their own tradition and point of view. Although there is agreement on the most important stages in the development of the technique, there are some omissions and some ambiguities. Some of the MCA historians, in their eagerness to produce a long and impressive list of references, do not seem to distinguish multiple from ordinary Correspondence Analysis, one-dimensional from multidimensional analysis, binary data from multicategory data, and data with or without a dependent variable.

What we call "prehistory" is MCA before Guttman (1941). And what we find in the prehistory is almost exclusively Reciprocal Averaging Analysis (RAAA). We define RAAA, in the present paper, starting from the indicator matrix G. Take any set of trial weights for the categories. Then compute the score for the individual by averaging the weights of the categories selected by that individual, and then compute a new set of weights for categories by averaging the scores of the individuals in the categories. And these two reciprocal averaging steps are iterated until convergence, when weights and scores do not change any more (up to a proportionality factor).

In various places it is stated, or at least suggested, that RAA (both the name and the technique) started with Richardson and Kuder (1933). This seems incorrect. The paper has no trace of RAA, although it does document a scale construction using Hollerith sorting and tabulation machines. What seems to be true, however, is that both the RAA name and the technique started at Proctor & Gamble in the early thirties, in an interplay between Richardson and Horst, both P&G employees at the time. This relies on the testimony of Horst (1935), who does indeed attribute the name and basic idea of RAA to Richardson:

> The method which he suggested was based on the hypothesis that the scale value of each statement should be taken as a function of the average score of the men for whom the statement was checked and, further, that the score of each man should be taken as a function of the average scale value of the statements checked for the man.

The definition given by Horst is rather vague, because "a function of" is not specified. It also does even mention the iteration of RAA to convergence (or, as Guttman would say, internal consistency). This iterative extension again seems to be due either to Horst or to Richardson. Horst was certainly involved at the time in the development of very similar techniques for quantitative data (Horst (1936), Edgerton and Kolbe (1936), Wilks (1938)). Both for quantitative and qualitative data these techniques are based on minimizing within-person and maximizing between-person variance, and they all result in computing the leading principal component of some data matrix. Horst (1935), starting from the idea to make linear combinations to maximize between-individual variance, seems to have been the first one to

realize that the equations defining RAA are the same as the equations describing Principal Component Analysis (PCA), and that consequently there are multiple RAA solutions for a given data matrix.

There are some additional hints about the history of RAA in the conference paper of Baker and Hoyt (1972). They also mostly credit Richardson, although they mention he never published a precise description of the technique, and it has been used "informally" without a precise justification ever since. They also mention that the first Hollerith type of computer implementation of RAA was by Mosier in 1942, the first Univac program was by Baker in 1962, and the first FORTRAN program was by Baker and Martin in 1969.

We have not mentioned in our prehistory the work of Fisher (1938), Fisher (1940), and Maung (1941). These contributions, basically contemporaneous with Guttman (1941), clearly introduced the idea of optimal scaling categorical data, of COA of a two-way table, and even of nonlinear transformation of the data to fit a linear (additive) model. And they also came up with the first principal component of a Gramian matrix as a solution, realizing there are multiple solutions to their equations. But, as pointed out by Gower (1990), they do not use MCA as it is currently defined. And, finally, although Hill (1973) seems to have independently come up with the RAA name and technique, its origins are definitely not in ecology.

3.2 Guttman 1941

RAA was used to construct a single one-dimensional scale, but Horst (1935) indicated already its extension to more than one dimension. The first publication of the actual formulas, using the luxuries of modern matrix algebra, was Guttman (1941), ironically in a chapter of a book edited by Horst. This is really where the history of MCA begins, although there are still some notable differences with later practice.

Guttman starts with the indicator matrix G, and then generalizes and systematizes the analysis of variance approach to optimal scaling of indicator matrices. He introduces three criteria of internal consistency: one for the categories (columns), one for the objects (rows), and one for the entire table. All three criteria lead to the same optimal solution, which we now recognize as the first non-trivial dimension of MCA. We now also know, after being exposed to more matrix algebra than was common in the forties and fifties, that this merely restates the fact that for any matrix X the non-zero eigenvalues of $X'X$ and XX' are the same, and moreover equal to the squares of the singular values of X. And the left and right singular vectors of X are the eigenvectors of $X'X$ and XX'.

For our purposes in this paper the following quotation from Guttman's section five is important. When discussing the multiple solutions of the MCA stationary equations he says:

> There is an essential difference, however, between the present problem of quantifying a class of attributes and the problem of "factoring" a set of quantitative variates. The principal axis solution for a set of quantitative variates depends

> on the preliminary units of measurement of those variates. In the present problem, the question of preliminary units does not arise since we limit ourselves to considering the presence or absence of behavior.

Thus Guttman, at least in 1941, shows a certain reluctance to consider the additional dimensions in MCA for data analysis purposes.

In addition to the stationary equations of MCA, Guttman also introduces the chi-square metric. He notes that the rank of C, and thus of E, is that of the indicator G, which is at most $1+\sum_{j=1}^{m}(k_j-1)=k_\star-(m-1)$. Thus C has at least $m-1$ zero eigenvalues, inherited from the linear dependencies in G. In addition E has a trivial eigenpair, independent of the data, with eigenvalue equal to 1. Suppose e has all its $k_\star$ elements equal to $+1$. Then $Ce=mDe$ and thus $Ey=y$, with $y=D^{\frac{1}{2}}e$. If we deflate the eigenvalue problem by removing this trivial solution then the sum of squares of any off-diagonal submatrix of C is the chi-square for independence of that table.

Guttman also points out that the scores and weights linearize both regressions if we interpret the indicator matrix as a discrete bivariate distribution. This follows directly from the interpretation of MCA as a Correspondence Analysis of the indicator matrix, because Correspondence Analysis linearizes regressions in a bivariate table. Of course interpreting the binary indicator matrix G as a bivariate distribution is quite a stretch. Both the chi-square metric and the linearized regressions were discussed earlier by Hirschfeld (1935) in the context of a single bivariate table. Neither Hirschfeld nor Fisher are mentioned in Guttman (1941).

There are no data and examples in Guttman's article. Benzécri (1977a) remarks

> L. Guttman avait défini les facteurs mêmes calculés par l'analyse des correspondances. Il ne les avait toutefois pas calculés; pour la seule raison qu'en 1941 les moyens de calcul requis (ordinateurs) n'existaient pas.

But that is not exactly true. In Horst (1941) the chapter by Guttman is followed by another chapter called "Two Empirical Studies of Weighting Techniques", which does have an empirical application in it. It is unclear who wrote that chapter, but the computations, which were carried out on a combination of tabulating and calculating machines, were programmed by nobody less than Ledyard R. Tucker.

3.3 Burt 1950

Guttman was reluctant to look at additional solutions of the stationary equations (additional "dimensions"), but Burt (1950) had no such qualms. After a discussion of the indicator matrix G and its corresponding cross product C (now known as the Burt table) Burt suggests a PCA of the normalized Burt table, i.e. solving the eigen problem $Ey=m\lambda y$. By the way, Burt discusses PCA as an alternative method of factor analysis, which is not in line with current usage clearly distinguishing PCA and FA.

Most of Burt's references are to previous PCA work with quantitative variables, and much of the paper tries to justify the application of PCA to qualitative data. No references to Guttman, Fisher, Horst, or Hirschfeld. The justifications that Burt presents are from the factor analysis perspective: C is a Gramian matrix, and E is a correlation matrix, and the results of factoring E can lead to useful classifications of the individuals.

In the technical part of the paper Burt discusses the rank, the trivial solutions, and the connection with the chi-squares of the bivariate subtables that we have already mentioned in our Guttman (1941) section.

3.4 Guttman 1953

As we saw in the introduction Guttman (1953) starts his paper with the observation that he already published the MCA equations in 1941. He gives this a positive spin, however.

> It is gratifying to see how Professor Burt has independently arrived at much the same formulation. This convergence of thinking lends credence to the suitability of the approach.

I will now insert a long quote from Guttman (1953), because it emphasizes the difference with Burt, and it is of major relevance to the present paper as well. And Guttman really tells it like it is.

> My own article goes on to point out that, while the principal components here are formally similar to those for quantitative variables, nevertheless their interpretation may be quite different. The interrelations among qualitative items are not linear, nor even algebraic, in general. Similarly, the relation of a qualitative item to a quantitative variable is in general non-algebraic. Since the purpose of principal components - or any other method of factor analysis - is to help reproduce the original data, one must take into account this peculiar feature.
>
> The first principal component can possibly fully reproduce all the qualitative items entirely by itself: the items may be perfect, albeit non-algebraic, functions of this component. *Linear* prediction will not be perfect in this case, but this is not the best prediction technique possible for such data. Therefore, if the first principal component only accounts for a small proportion of the total variance of the data in the ordinary sense, it must be remembered that this ordinary sense implies linear prediction. If the correct, but *non-linear*, prediction technique is used, the whole variation can sometimes be accounted for by but the single component. In such a case, the existence of more than one principal component arises merely from the fact that a linear system is being used to approximate a non-linear one. (Each item is always a perfect linear function of *all* the principal components taken simultaneously.)

This was written after the publication of Guttman (1950), in which the MCA of a perfect scale of binary items is discussed in impressive detail. The additional dimensions in such an analysis are curvilinear functions of the first, in regular cases in fact orthogonal polynomials of a single scale. Specifically, the second dimension is a quadratic, or quadratic-looking, function of the first, which creates the famous "horseshoe" or "arch" (in French: the effect Guttman). Since a horseshoe curves back in at its endpoints that name is often not appropriate, and we will call these non-linearities the Guttman effect. It seems that the second and higher curved dimensions are just mathematical artifacts, and much has been published since 1950 to explain them, interpret them, or to get rid of them (Hill and Gauch (1980)).

In the rest of Guttman (1953) gives an overview of more of his subsequent work on scaling qualitative variables. This leads to material that goes beyond MCA (and thus beyond the scope of our present paper).

3.5 Burt 1953

Burt (1953), in his reply to Guttman (1953), admits there are different objectives involved.

> If, as I gather, he cannot wholly accept my own interpretations, that perhaps is attributable to the fact that our starting-points were rather different. My aim was to factorize such data; his to construct a scale.

This does not answer the question, of course, if it is really advisable to apply PCA to the normalized Burt matrix. It also seems there also are some differences in national foklore.

> In the chapters contributed to *Measurement and Prediction* both Dr. Guttman and Dr. Lazarsfeld draw a sharp distinction between the principles involved in these two cases. Factor analysis, they maintain, has been elaborated solely with reference to data which is quantitative *ab initio*; hence, they suppose, it cannot be suitably applied to qualitative data. On this side of the Atlantic, however, there has always been a tendency to treat the two cases together, and, with this double application in view, to define the relevant functions in such a way that they will (so far as possible) cover both simultaneously. British factorists, without specifying very precisely the assumptions involved, have used much the same procedures for either type of material. Nevertheless, there must of necessity be certain minor differences in the detailed treatment. These were briefly indicated in the paper Dr. Guttman has cited ; but they evidently call for a closer examination. I think in the end it will be found that they are much slighter than might be supposed.

Burt then goes on to treat the case of a perfect scale of binary items, previously analyzed by Guttman (1950). He points out that a PCA of a perfect scale gives (almost) the same results as those given by Guttman, and that consequently his approach of factoring a table works equally well as the approach that constructs a scale. And that the differences between

qualitative and quantitative factoring are indeed "much slighter than might be supposed." Although Burt is correct, he does not discuss where the Guttman effect comes from, and whether it is desirable and/or useful.

3.6 Benzécri 1977

French data analysis ("Analyse des Données") thinks of MCA as a special case of Correspondence Analysis (Le Roux and Rouanet (2010)). Benzécri (1977b) discusses the Correspondence Analysis of the indicator matrix and gives a great deal of credit to Ludovic Lebart. Lebart (1975) and Lebart (1976) are usually mentioned as the first publications to actually use "analyse de correspondences multiples" and "tableau de Burt".

Benzécri also gives Lebart the credit for discovering that a Correspondence Analysis of the indicator G gives the same results as a Correspondence Analysis of the Burt table C, which restates again our familiar matrix result that the singular value decomposition of a matrix gives the same results as the eigen decomposition of the two corresponding cross product matruces.

> L. Lebart en apporta la meilleure justification : les facteurs sur J issus de l'analyse d'un tel tableau I x J ne sont autres (à un coefficient constant près) que ceux issus de l'analyse du véritable tableau de contingence J x J suivant : k(j,j') = nombre des individus i ayant à la fois la modalité j et la modalité j. Dès lors on rejoint le format original pour lequel a été conçue l'analyse des correspondances.

Benzécri also mentions the surprising generality and wide applicability of MCA.

> Le succès maintenant bien compris des analyses de tableaux en 0,1 mis sous forme disjonctive complète invite à rapprocher de cette forme, par un codage approprié, les données les plus diverses.

This generality was later fully exploited in the book by Gifi (1990), which builds a whole system of descriptive multivariate techniques on top of MCA.

3.7 Gifi 1980

Gifi (1990) was mostly written in 1980-1981 as lecture notes for a graduate course in nonlinear multivariate analysis, building on previous work in De Leeuw (1973). Throughout, the main engine of the Gifi approach to multivariate analysis minimizes the meet-loss function

$$\sigma(X; Y_1, \cdots, Y_m) = \sum_{j=1}^{m} \mathbf{SSQ}(X - G_j Y_j) \tag{2}$$

over the $n \times p$ matrices of scores X with $X'X = nI$ and over the $k_j \times p$ matrices of weights Y_j that may or may not satisfy some constraints. Gifi calls this general approach Homogeneity

Anaysis (HA). Loss function (2) was partly inspired by Carroll (1968), who used this least squares loss function in generalized canonical analysis of quantitative variables.

The different forms of multivariate analysis in the Gifi framework arise by imposing additivity, and/or rank, and/or ordinal constraints on the Y_j. See De Leeuw and Mair (2009) for a user's guide to the R package homals, which implements minimization of meet-loss under these various sets of constraints.

If there are no constraints on the Y_j then minimizing (2) computes the p dominant dimensions of an MCA. What makes the loss function (2) interesting in our comparative review of MCA is the distance interpretation and the corresponding geometry of the joint biplot of objects and categories. Gifi minimizes the sum of the squared distances between an object and the categories of the variables that the object scores in. If we make a separate biplot for each variable j it has n objects points and k_j category points. The category points are in the centroid of the object points in that category, and if we connect all those objects with their categorympoints we get k_j star graphs in what Gifi calls the star plot. Minimizing (2) means making the joint plot in such a way that the stars are as small as possible.

Recent versions of the Gifi programs actually compute the proportion of individuals correctly classified if we assign each individual to the category it is closest to (in p dimensions). In this way we can indeed find, like Guttman, that a single component can account for all of the "variance".

There are indications, especially in Gifi (1990), section 3.9, that Gifi is somewhat uncomfortable with the multidimensional scale construction aspects of MCA. They argue that each MCA dimension gives a quantification or transformation of the variables, and thus each MCA dimension can be used to compute a different correlation matrix between the variables. These correlation matrices, of which there are $k_\star - m$, can then all be subjected to a PCA. So the single indicator matrix leads to $k_\star - m$ PCA's. Gifi calls this "data production", and obviously does not like the outcome. Thus, as an alternative to MCA, they suggest using only the first dimension and the corresponding correlation matrix, which is very close to RAA and Guttman (1941).

In the Gifi system the data production dilemma is further addressed in two ways. In the geometric framework based on the loss function (2) a form of nonlinear PCA is defined in which we restrict the $k_j \times p$ category quantifications of a variable to have rank one, i.e. the points representing the categories of a variable are on a line through the origin. Gifi shows that this leads to the usual non-linear PCA techniques (Young, Takane, and De Leeuw (1978), De Leeuw (2006)). The second development to get away from the "data production" in MCA is the "aspect" approach (De Leeuw (1988b), Mair and De Leeuw (2010), De Leeuw, Michailidis, and Wang (1999), De Leeuw (2004)). There we look for a single quantification or transformation of the variables that optimizes any real valued function (aspect) of the resulting correlation matrix. Nonlinear PCA is the special cases in which we maximize the sum of the first p eigenvalues of the correlation matrix, and MCA chooses the scale to maximize the dominant eigenvalue. Other aspects lead to regression, canonical analysis, and structural equation models. In this more recent aspect methodology Guttman's one-dimensional scale construction approach has won out over Burt's multidimensional factoring method.

4 Deconstructing MCA

4.1 Introduction

We are left with the following questions from our history section, and from the Burt-Guttman exchange.

1. What, if anything, is the use of additional dimensions in MCA ?
2. Where does the Guttman effect come from ?
3. Is MCA really just PCA ?
4. How many dimensions of MCA should we keep ?
5. Which "variance" is "explained" by MCA ?
6. How about the "data production" aspects of MCA ?

In De Leeuw (1982) several results are discussed that are of importance in answering these questions, and more generally for the interpretation (and deconstruction) of MCA. Additional, and more extensive, discussion of these same results is in Bekker and De Leeuw (1988) and De Leeuw (1988a)

To compute the MCA eigen decomposition we could use the Jacobi method, which diagonalizes E by using elementary plane rotations. It builds up Y by minimizing the sum of squares of the off-diagonal elements. Thus E is updated by iteratively replacing it by $J_{st}EJ_{st}$, where J_{st} with $s < t$ is a Jacobi rotation, i.e. a matrix that differs from the identity matrix of order $k_\star$ only in elements (s, s) and (t, t), which are equal to u, and in elements (s, t) and (t, s) which are $+v$ and $-v$, where u and v are real numbers with $u^2 + v^2 = 1$. We cycle through all upper-diagonal elements $s < t$ for a single iteration, and continue iterating until the E update is diagonal (within some ϵ).

We shall discuss a different three-step method of diagonalizing E, which, for lack of a better term, we call Deconstructed Correspondence Analysis (DCA). It also works by applying elementary plane rotations to E, but it is different from the Jacobi method because it is not intended to diagonalize any real symmetric matrix. It uses its rotations to eliminate all off-diagonal elements of all m^2 submatrices E_{kl}. If it cannot do this perfectly, it will try to find the best approximate diagonalization. If DCA does diagonalize all submatrices, then some rearranging and additional computation finds the eigenvalues and eigenvectors of E, and thus the MCA. The eigenvectors are, however, ordered differently (not by decreasing eigenvalues), and provide more insight in the innards of MCA. If an exact diagonalization is not possible, the approximate diagonalization often still provides this insight.

We first discuss some theoretical cases in which DCA does give the MCA, and after that some empirical examples in which it only approximates a diagonal matrix. As you will hopefully see, both types of DCA examples show us what MCA as a data analysis technique tries to do, and how the results help in answering the questions arising from the Burt-Guttman exchange.

4.2 Mathematical Examples

4.2.1 Binary Data

Let's start with the case of binary data, i.e. in which all k_j are equal to two. The normalized Burt table $E = m^{-1}D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$ consists of $m \times m$ submatrices $E_{j\ell}$ of dimension 2×2. Suppose the marginals of variable j are p_{j0} and p_{j1}. For each j make the 2×2 table

$$K_j = \begin{bmatrix} +\sqrt{p_{j0}} & +\sqrt{p_{j1}} \\ +\sqrt{p_{j1}} & -\sqrt{p_{j0}} \end{bmatrix}, \tag{3}$$

and suppose K is the direct sum of the K_j, i.e. the block-diagonal matrix with the K_j on the diagonal. Then $F := K'EK$ again has $m \times m$ submatrices of order two. Each of the $F_{j\ell} = K_j'E_{j\ell}K_\ell$ is diagonal, with element $(1,1)$ equal to $+1$ and element $(2,2)$ equal to the point correlation (or phi-coefficient) between binary variables j and ℓ.

This means we can permute rows and columns of F using a permutation matrix P such that $R := P'FP$ is the direct sum of two correlation matrices Γ_0 and Γ_1, both of order m. Γ_0 has all elements equal to $+1$, Γ_1 has its elements equal to the phi-coefficients. We collect the $(1,1)$ elements of all F_{jl}, which are all $+1$, in Γ_0 and the $(2,2)$ elements in Γ_1. Suppose L_0 and L_1 are the normalized eigenvectors of Γ_0 and Γ_1, and L is their direct sum. Then $\tilde{\Lambda} = m^{-1}LRL$ is diagonal, with on the diagonal the eigenvalues of E and with KPL the normalized eigenvectors of E. Thus the eigenvalues of E are those of $m^{-1}\Gamma_0$, i.e. one 1 and $m-1$ zeroes, together with those of $m^{-1}\Gamma_1$. This restates the well-known result, mentioned by both Guttman (1941) and Burt (1950), that an MCA of binary data reduces to a PCA of the phi-coefficients.

4.2.2 Correspondence Analysis

Now let us look at Correspondence Analysis, i.e. MCA with $m = 2$. There is only one single off-diagonal $p \times q$ cross table C_{12} in the Burt matrix. Suppose without loss of generality that $p \geq q$. Define K as the direct sum the left and right singular vectors of E_{12}. Then

$$F = K'EK = \begin{bmatrix} I & \Psi \\ \Psi' & I \end{bmatrix} \tag{4}$$

where Ψ is the $p \times q$ diagonal matrix of singular values of E_{12}, and

$$R = P'FP = \left\{ \bigoplus_{s=1}^{q} \begin{bmatrix} 1 & \psi_s \\ \psi_s & 1 \end{bmatrix} \right\} \bigoplus I, \tag{5}$$

where the identity matrix at the end of equation (5) is of order $p - q$.

Thus the eigenvalues of E are $\frac{1}{2}(1+\gamma_j)$ and $\frac{1}{2}(1-\lambda_j)$ for all j, and DCA indeed diagonalizes E. The relation between the eigen decomposition of E and the singular value decomposition of E_{12} is a classical result in Correspondence Analysis (Benzécri (1977a)), and earlier already in canonical correlation analysis of two sets of variables (Hotelling (1936)).

4.2.3 Multinormal Distribution

Suppose we want to apply MCA to an m-variate standard normal distribution with correlation matrix $R = \{\rho_{k\ell}\}$. Not to a sample, mind you, but to the whole distribution. This means we have to think of the submatrices $C_{j\ell}$ as bivariate standard normal densities ("Gendarme's Hats"), having an infinite number of categories, one for each real number. Just imagine it as a limit of the discrete case (Naouri (1970)).

In this case the columns of the K_j, of which there is a denumerably infinite number, are the Hermite-Chebyshev polynomials $h_0, h_1, \cdots$ on the real line. We know that for the standard bivariate normal $\mathrm{E}_{j\ell}(h_s, h_t) = 0$ is $s \neq t$ and $\mathrm{E}_{j\ell}(h_s, h_s) = \rho_{j\ell}^s$. Thus $F := K'EK$ is an $m \times m$ matrix of diagonal matrices, where each F_{kl} submatrix is of denumerably infinite order and has all the powers of $\rho_{k\ell}$ along the diagonal. Then $R = P'FP$ is the infinite direct sum of elementwise powers of the matrix of correlation coefficients, or

$$R = P'FP = \bigoplus_{s=0}^{\infty} R^{(s)}, \tag{6}$$

and $\tilde{\Lambda} := L'RL$ is diagonal, with first the m eigenvalues of $R^{(0)} = ee'$, then the m eigenvalues of $R^{(1)} = R$, then the m eigenvalues of $R^{(2)} = \{\rho_{jl}^2\}$, and so on to $R^{(\infty)} = I$. Each MCA solution is composed of Hermite-Chebyshev polynomials of the same degree. Again, this restates a known result, already given in De Leeuw (1973).

These results remain true for what Yule called "strained multinormals", i.e. multivariate distributions that can be obtained from the multivariate normal by separate and generally distinct smooth monotone transformations of each of the variables. It also applies to mixtures of multivariate standard normal distributions with different correlation matrices (Sarmanov and Bratoeva (1967)), to Gaussian copulas, as well as to other multivariate distributions whose bivariate marginals have diagonal expansions in systems of orthonormal functions (the so-called Lancaster probabilities, after Lancaster (1958) and Lancaster (1969)).

The multinormal is a perfect example of the Guttman effect, i.e. the eigenvector corresponding with the second largest eigenvalue usually is a quadratic function of the first, the next eigenvector usually is a cubic, and so on. We say "usually", because Gifi (1990), page 382-384, gives a multinormal example in which the first two eigenvectors of an MCA are both linear transformations of the underlying scale (i.e. they both come from Γ_1). However, the Guttman effect is observed approximately in many (if not most) empirical applications of MCA, especially if the categories of the variables have some natural order and if the number of individuals is large enough.

4.2.4 Common Mathematical Structure

What do our three previous examples have in common mathematically ? In all three cases there exist orthonormal K_j and diagonal $\Phi_{j\ell}$ such that $E_{j\ell} = K_j \Phi_{jl} K_\ell'$. Or, in words, the matrices $E_{j\ell}$ in the same row-block of E have their left singular vectors K_j in common, and matrices $E_{j\ell}$ in the same column-block of E have their right singular vectors K_ℓ in common. Equivalently, this requires that for each j the m matrices $E_{j\ell}E_{\ell j}$ commute.

Another way of saying this is that there are weights $y_1, \cdots, y_m$ so that $C_{j\ell}y_\ell = \rho_{j\ell}D_j y_j$, i.e. so that all bivariate regressions are linear (De Leeuw (1988a)). And not only that, we assume that such a set of weights exist for every dimension s, as long as $k_j \geq s$. If $k_j = 2$ then trivially all regressions are linear, because you can always draw a straight line through two points. If $m = 2$ all Correspondence Analysis solutions linearize the regressions in a bivariate table. In the multinormal example the Hermite polynomials provide the linear regressions. Simultaneous linearizability of all bivariate regressions seems like a strong condition, which will never be satisfied for observed Burt matrices. But our empirical examples, analyzed below, suggest it will be approximately satisfied in surprisingly many cases. And, at the very least, assuming simultaneous linearizibility is a far-reaching generalization of assuming multivariate normality.

In all three mathematical examples we used the direct sum of the K_j to diagonalize the $E_{j\ell}$, then use a permutation matrix P to transform $F = K'EK$ into the direct sum $R = P'FP$ of correlation matrices, and then use the direct sum L to diagonalize R to $\tilde{\Lambda} = L'RL$. This means that KPL has the eigenvectors of E, but ordered by decreasing or increasing eigenvalues. It also means that the eigenvectors have a special structure.

First, F is an $m \times m$ matrix of matrices $F_{j\ell}$, which are $k_j \times k_\ell$. If all k_j are equal to, say, k, then R is a $k \times k$ matrix of matrices R_{st}, which are all of order m. If variables have a different number of categories, then R is a $k_+ \times k_+$ matrix of correlation matrices, with R_{st} of order $m_s \times m_t$, where m_s is defined as before as the number of variables with $k_j \geq s$.

KP is an orthonormal $m \times k_+$ matrix of matrices, in which column-block s is the direct sum of the m_s column vectors $K_j e_s$, with e_s unit vector s (equal to zero, except for element s, which is one). Thus $\{KP\}_{js} = K_j e_s e_s'$ and $\{KP\}_{js}L_s = K_j e_s e_s' L_s$. $\{KPL\}_{js}$ is the $k_j \times m_s$ outer product of column s of K_j and row s of L_s. Thus each Γ_s is computed with a single quantification of the variables, and there only $k_+ - 1$ different non-trivial quantifications, instead of the $k_\star - m$ ones from MCA.

That the matrix KPL is blockwise of rank one connects DCA with non-linear PCA, which is after all MCA with rank one restrictions on the category weights. We see that imposing rank one restrictions on MCA force non-linear PCA to choose its solutions from the same Γ_s, thus preventing data production.

5 The Chi-square Metric

In the Correspondence Analysis of a single table it has been known since Hirschfeld (1935) that the sum of squares of the non-trivial singular values is equal to the chi-square (the total inertia) of the table. Athough both Burt and Guttman pay hommage to chi-square in the context of MCA, they do not really wokr through the consequences for MCA. In this section we analyze the total chi square (TCS), which is the sum of all $m(m-1)$ off-diagonal bivariate chi-squares.

De Leeuw (1973), p 32, shows that the TCS is related to the MCA eigenvalues by the simple

equation

$$\sum_{1\leq j\neq \ell\leq m}\sum \mathcal{X}^2_{j\ell} = n\sum_s (m\lambda_s - 1)^2, \tag{7}$$

where the sum on the right is over all $k_\star - m$ nontrivial eigenvalues. sides by two. Equation (7), the MCA decomposition of the TCS, gives us a way to quantify the contribution of each non-trivial eigenvalue.

We now outline the DCA decompositon of the TCS. An identity similar to (7) is

$$\sum_{1\leq j\neq \ell\leq m}\sum \mathcal{X}^2_{j\ell} = \text{tr } E^2 - (K + m(m-1)). \tag{8}$$

Equation (8) does not look particular attractive, until one realizes that the constant subtracted on the right is the number of trivial elements in $F = K'EK$ (and thus in $R = P'K'EPK$) equal to one. There are K elements on the main diagonal, and $m(m-1)$ elements from the off-diagonal elements of the trivial matrix Γ_0.

Thus the TCS can be partitioned using R, which is a $k_+ \times k_+$ matrix of matrices into $(k_+ - 1)^2$ non-trivial components. The most interesting ones are the $k_+ - 1$ sums of squares of the off-diagonal elements of the diagonal submatrices $\Gamma_1, \cdots, \Gamma_{k_+-1}$, which is actually the quantity maximized by DCA. And then there are the $(k_+ - 1)(k_+ - 2)$ sums of squares of the off-diagonal submatrices of R, which is actually what DCA minimizes. The sum of squares of each diagonal block separately is its contribution to the DCA fit, and total contribution to chi-square over all diagonal blocks shows how close DCA is to MCA, i.e. how well DCA diagonalizes E. In the mathematical examples from section 4.2 DCA is just a rearranged MCA, and all of the TCS comes from the diagonal blocks.

6 Computation

So, computationally, DCA works in three steps. All three steps preserve orthonormality, guaranteeing that if DCA diagonalization works we have actually found eigenvalues and eigenvectors.

In the first step we compute the K_j by trying to diagonalize all off-diagonal $E_{j\ell}$. This is is done in the mathematical examples by using known analytical results, but in empirical examples it is done by Jacobi rotations that minimize the sum of squares of all off-diagonal elements of the off-diagonal $K'EK$ (or, equivalently, maximize the sum of squares of the diagonal elements).

Each K_j is $k_j \times k_j$ and square orthonormal. We always set the first column of K_j equal to $n^{-\frac{1}{2}}\sqrt{d_j}$, with d_j the marginals of variable j, to make sure the first column captures the non-zero trivial solution. In the R program this is done by setting the initial K_j to the left singular vectors of row-block j of E and not rotating pairs of indices (s, t) when s or t is one. This usually turns out to be a very good initial solution.

In the second step we permute rows and columns of $F := K'EK$ into direct sum form. The $(1, 1)$ matrix R_{11} in $R := P'K'PKP$, which we also call Γ_0, has the $(1, 1)$ elements of all $F_{j\ell}$,

the $(1,2)$ matrix R_{12} has the $(1,2)$ elements of all $F_{j\ell}$, and so on. Thus, if the first step has diagonalized all off-diagonal $E_{j\ell}$, then all off-diagonal matrices in R are zero. The square symmetric matrices along the diagonal, of which there are k_+, are of order m, or of order m_s if not all k_j are equal. The first two, Γ_0 and Γ_1, are always of order m. Γ_0 takes care of all m trivial solutions and has all its elements equal to one.

Then, in the third step, we diagonalize the matrices along the diagonal of R by computing their eigenvalues and eigenvectors. This gives $\tilde{\Lambda} = L'RL$, which is diagonal if the first step succeeded in diagonalizing all off-diagonal $E_{j\ell}$. All the loss that can make DCA an imperfect diagonalization method is in the first step, computing both P and L does not introduce any additional loss. Note again that the direct sums K and L and the permutation matrix P are all orthonormal, and thus so are KP and KPL.

Finally we compute $Y'KPL$, with Y the MCA solution, to see how close Y and KPL are, and which Γ_s the MCA solutions come from. Note that $Y'KPL$ is also square orthonormal, which implies sums of squares of rows and columns add up to one, and squared elements can be interpreted as proportions of "variance explained".

6.1 The Program

For the empirical examples in the present paper we use the R function DCA(), a further elaboration of the R function jMCA() from De Leeuw and Ferrari (2008). The program, and all the empirical examples with the necessary data manipulations, is included in the supplementary material for this chapter. The program maximizes the percentage of the TCS from the diagonal blocks of the DCA. It is called with arguments

- burt (the burt matrix),
- k (the number of catgories of the variables),
- eps (iteration precision, defaults to 1e-8),
- itmax (maximum number of iterations, defaults to 500),
- verbose (prints DCA fit for all iterations, defaults to TRUE),
- vectors (DCA eigenvectors, if FALSE only DCA eigenvalues, defaults to TRUE).

And it returns a list with

- kek = $K'EK$,
- pkekp = $P'K'EKP$,
- lpkekpl = $L'P'K'EKPL$,
- k = K,
- p = P,
- l = L,
- kp = KP,
- kpl = KPL,
- chisquares = $m(m-1)$ chi-squares
- chipartition = DCA chi-partition,

- chipercentages = chipartition / TCS,
- itel = number of iterations,
- func = optimum value of trace of chipercentages

7 Empirical Examples

We analyzed DCA in our previous examples by relying heavily on specific mathematical properties. There are some expirical examples in the last section of De Leeuw (1982), but with very little detail, and computed with a now tragically defunct APL program. Showing the matrices K, P, L as well as F, R and $\tilde{\Lambda}$ in this chapter would take up too much space, so we concentrate on how well DCA reproduces the MCA eigenvalues. We also discuss which of the correlation matrices in R the first and last MCA vectors of weights (eigenvectors) are associated with, and we give the partitionings of the TCS.

7.1 Burt Data

The data for the example in Burt (1950) were collected by him in Liverpool in or before 1912, and are described in an outrageously politically incorrect paper (Burt (1912)). Burt used the four variables hair-color (fair, red, dark), eye color (light, mixed, brown), head (narrow, wide), and stature (tall, short) for 100 individuals selected from his sample. This is not very interesting as a DCA or MCA example, because the data are so close to binary and thus there is not much room for DCA to work with. We include it for historical reasons.

The Burt table is of order 10, and there are 6 nontrivial eigenvalues. DCA takes one single iteration cycle to convergence to fit 0.9461841756 from the initial SVD solution. Figure 1 plots the sorted MCA and DCA non-trivial eigenvalues. In these plots we always remove the trivial points $(0, 0)$ and $(1, 1)$ because they would anchor the plot and unduly emphasize the closeness of the two solutions.

The matrix R has two diagonal blocks Γ_0 and Γ_1 of order 4, and one block Γ_2 of order 2. Thus the m_s are $(4, 4, 2)$. The first non-trivial MCA solution correlates 1.00 with the first non-trivial DCA solution, which corresponds with the dominant eigenvalue of Γ_1. The second MCA solution correlates 0.73 with the second DCA solution from Γ_1 and 0.37 and 0.57 with the two DCA solutions from Γ_2. The fifth and sixth MCA solutions (the ones with the smallest non-trivial eigenvalues) correlate 0.98 and 0.99 with the remaining two DCA solutions from Γ_1. Thus almost all the variation comes from Γ_1, which is what we expect in an analysis with the k_j equal to $(3, 3, 2, 2)$.

We can further illustrate this with the chi-square partitioning. Of the TCS of 156.68 the diagonal blocks Γ_1 and Γ_2 contribute respectively 148.17 (95%) and 0.08 (0.05%), while the off-diagonal blocks contribute 8.43 (5%).

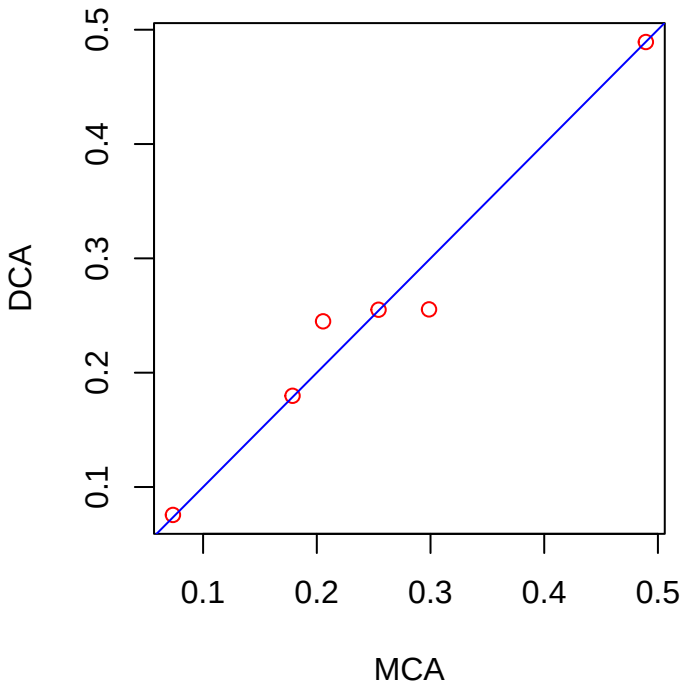

Figure 1: Burt MCA/DCA Eigenvalues

7.2 Johnson Data

In the same year as Burt (1950) an MCA example was published by Johnson (1950). The data were letter grades for forty engineering students on four exam topics. Thus $n = 40$, $m = 4$, and $k_j = 5$ for all j. Johnson took his inspiration from the work of Fisher and Maung, so in a sense he is the bridge between their contributions and MCA. Except for one crucial fact. Johnson decided to require the same numerical letter weights for all four exams. An analysis with that particular constraint is actually a Correspondence Analysis of the sum of the four indicators G_j (Van Buuren and De Leeuw (1992)). And of course Johnson, in the Fisher-Maung tradition, was only interested in the dominant component for the optimal weights.

We have done a proper MCA and DCA of the Johnson data, which clearly have ordered categories. They also have the same number of categories per variables, which is of course common in attitude scales and in many types of questionaires. The Burt matrix has order 20 and consists of 4×4 square submatrices of order five. Matrix R has 5×5 square submatrices of order four. Thus DCA computes the eigenvalues of $\Gamma_0, \cdots, \Gamma_4$. It takes 64 iteration cycles to convergence to a fit of 0.9139239809. The sorted non-trivial MCA and DCA eigenvalues are compared in figure 2.

The correspondence between the first MCA and DCA solutions is straightforward in this case. There is a strong Guttman effect, and thus the first MCA solution is the dominant solution from Γ_1 (correlation with DCA 1.00), the second MCA solution from the largest eigenvalue of Γ_2 (correlation 0.91), and the third from the largest eigenvalue of Γ_2 (correlation 0.98). In the chi-square partitioning the four diagonal blocks take care of 58%, 15%, 11%, and 7%

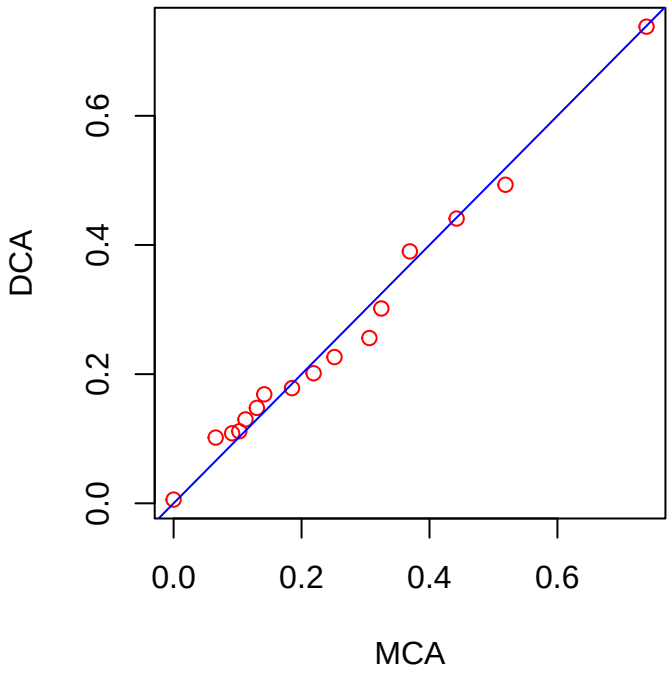

Figure 2: Johnson MCA/DCA Eigenvalues

of the total chi-square, and thus DCA covers 91% of MCA. The complete partitioning is

```
##       [,1]   [,2]   [,3]   [,4]
## [1,] +0.581 +0.003 +0.014 +0.003
## [2,] +0.003 +0.151 +0.006 +0.011
## [3,] +0.014 +0.006 +0.108 +0.006
## [4,] +0.003 +0.011 +0.006 +0.074
```

7.3 GALO Data

The GALO data (Peschar (1975)) are a mainstay Gifi example. The individuals are 1290 sixth grade school children in the city of Groningen, The Netherlands, about to go into secondary education. The four variables are gender (2 categories), IQ (9 categories), teachers advice (7 categories), and socio-economic status (6 categories). The Burt matris is of order 24, and thus there are $K - m = 20$ non-trivial dimensions. Matrix $R = P'FP$ has 9 diagonal correlation blocks, with Γ_0 and Γ_1 of order four, $\Gamma_2, \cdots, \Gamma_5$ of order three, Γ_6 of order two, and Γ_7 and Γ_8 of order one. DCA takes 37 iteration cycles to a fit of 0.8689177854. The 20 sorted non-trivial MCA and DCA eigenvalues are plotted in figure 3.

There is a strong Guttman effect in the GALO data. The first MCA solution correlates 1.00 with the dominant DCA solution from Γ_1, the second MCA solution correlates 0.99 with the dominant DCA solution from Γ_2, and the third correlates 0.86 with the DCA solution from Γ_3. After that correlations become smaller, until we get to the smallest eigenvalues. The

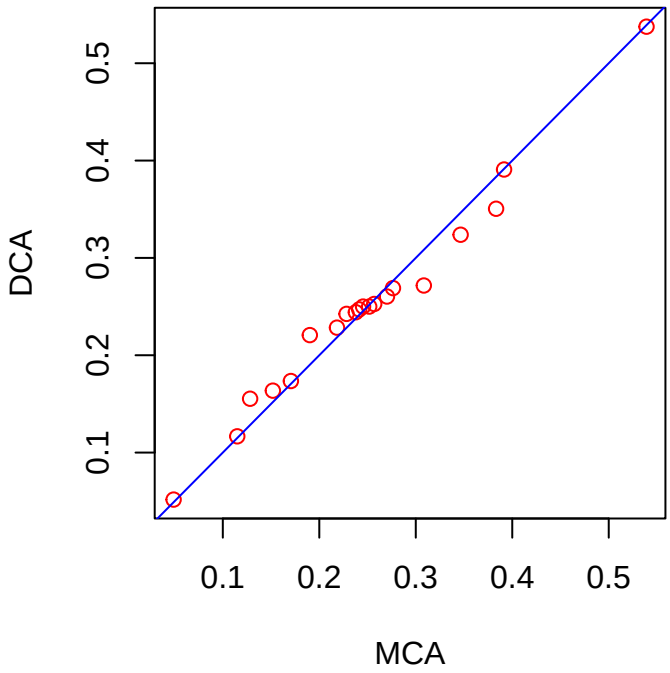

Figure 3: GALO MCA/DCA Eigenvalues

worst MCA solution correlates 0.99 with the smallest eigenvalue of Γ_1, and the next worst correlates 0.98 with the smallest eigenvalue of Γ_2.

The chi-square partioning tells us the diagonal blocks of DCA "explain" 87% of the TCS, with the blocks $\Gamma_1, \cdots, \Gamma_6$ contributing 56%, 16%, 8%, 5%, .5%, and .4%. The complete partitioning is

```
##      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
## [1,] +0.563 +0.002 +0.022 +0.017 +0.015 +0.002 +0.001 +0.000
## [2,] +0.002 +0.164 +0.000 +0.000 +0.001 +0.000 +0.000 +0.000
## [3,] +0.022 +0.000 +0.083 +0.000 +0.001 +0.000 +0.000 +0.000
## [4,] +0.017 +0.000 +0.000 +0.049 +0.002 +0.001 +0.000 +0.000
## [5,] +0.015 +0.001 +0.001 +0.002 +0.006 +0.000 +0.001 +0.000
## [6,] +0.002 +0.000 +0.000 +0.001 +0.000 +0.004 +0.000 +0.000
## [7,] +0.001 +0.000 +0.000 +0.000 +0.001 +0.000 +0.000 +0.000
## [8,] +0.000 +0.000 +0.000 +0.000 +0.000 +0.000 +0.000 +0.000
```

7.4 BFI Data

Our final example is larger, and somewhat closer to actual practice. The bfi data set is taken from the psychTools package (Revelle (2021)). It has 2800 observations on 25 personality self report items. After removing persaons with missing data there are 2436 observations left. Each item has 6 categories, and thus the Burt table is of order 150. Matrix R, excluding Γ_0,

has 5 diagonal blocks of order 25. DCA() takes 54 iterations for a DCA fit of 0.885985659. The sorted non-trivial 125 MCA and DCA eigen values are plotted in figure 4.

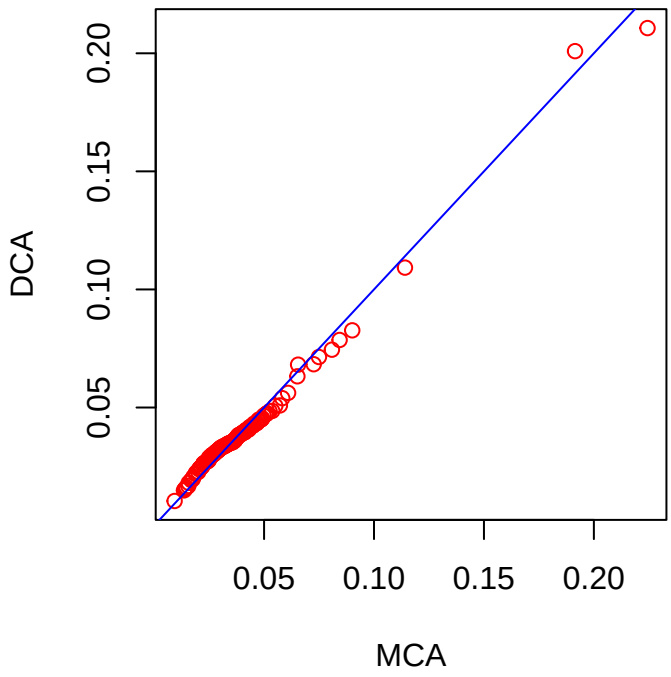

Figure 4: BFI MCA/DCA Eigenvalues

The percentages of the TCS from the non-trivial submatrices of R are

```
##      [,1]   [,2]   [,3]   [,4]   [,5]
## [1,] +0.488 +0.015 +0.006 +0.006 +0.004
## [2,] +0.015 +0.330 +0.005 +0.004 +0.004
## [3,] +0.006 +0.005 +0.039 +0.005 +0.004
## [4,] +0.006 +0.004 +0.005 +0.021 +0.005
## [5,] +0.004 +0.004 +0.004 +0.005 +0.008
```

dim(# Discussion

Our examples, both the mathematical and the empirical ones, show that in a wide variety of circumstances MCA and DCA eigenvalues ore very similar, although DCA uses far fewer degrees of freedom for the diagonalization. This indicates that DCA can be thought of, at least in some circumstances, as a smoothing of MCA. The error is moved to the off-diagonal elements in the submatrices of $R = P'FP$ and the structure is concentrated in the diagonal correlation matrices.

We have also seen that DCA is like MCA, in the sense that it gives very similar solutions, but it is also like non-linear PCA, because it imposes the rank one restrictions on the weights. Thus it is a bridge between the two techniques, and it clarifies their relationship.

DCA also shows where the dominant MCA solutions originate, and indicates quite clearly where the Guttman effect comes from (if it is there). It suggest the Guttman effect, in

a generalized sense, does not necessarily result in polynomials or arcs. As long as there is simultaneous linearization of all bivariate regressions we have that E is orthonormally similar to the direct sum of the Γ_s, and the principal components of the Γ_s will give a generalized Guttman effect.

This allows us to suggest some answer for questions coming from the Burt-Guttman exchange. The principal components of MCA beyond the first in many cases come from the generalized Guttman effect, and should be interpreted as such. Thus the first pricipal component does have a special status, and that justifies singling out RAA and Guttman scaling from the rest of MCA.

DCA also reduces the amount of data production. Instead of $k_\star - m$ non-trivial correlations matrices of order m with their PCA's, we now have $k_+ - 1$ non-trivial correlation matrices of orders given by the m_s. It is still more than one single correlation matrix, as in non-linear PCA and the aspect approach, but the different correlation matrices may actually be related by the Guttman effect.

This also seems the place to point out a neglected aspect of MCA. The smallest non-trivial solution gives a quantification or transformation of the data that maximizes the singularity of the transformed data, i.e. the minimum eigenvalue of the corresponding correlation matrix. We have seen that MCA and DCA often agree strongly in their smallest eigenvalue solutions, and that may indicate that it may be possible to give a scientific interpretation of these solutions. In fact, the smallest DCA and MCA eigenvalues can be used in a regression interpretation in which we consider one or more of the variables as criteria and the others are predictors.

A complaint that many users of MCA have is that, say, the first two components "explain" such a small proportion of the "variance" (by which they mean the trace of E, which is K, the total number of categories, and that, of course, has nothing to do with variance). Equation (7) indicates how to quantify the contributions of the non-trivial eigenvalues. For the BFI data, for example, the first two non-trivial MCA eigenvalue "explain" 0.0831694181 percent of the "variance", but they "explain" 0.6304991891 percent of the TCS. Moreover DCA shows us that we should really relate the eigenvalues to the Γ_s that they come from, and see how much they "explain" of that correlation matrix. And to evaluate their contributions using the TCS and its partitioning described in section 5 of this paper.

References

Baker, F. B., and C. J. Hoyt. 1972. "The Relation of the Method of Reciprocal Averages to Guttman's Internal Consistency Scaling Method." https://eric.ed.gov/?id=ED062397.

Bekker, P., and J. De Leeuw. 1988. "Relation Between Variants of Nonlinear Principal Component Analysis." In *Component and Correspondence Analysis*, edited by J. L. A. Van Rijckevorsel and J. De Leeuw, 1–31. Wiley Series in Probability and Mathematcal Statistics. Chichester, England: Wiley.

Benzécri, J. P. 1977a. "Histoire et Préhistoire de l'Analyse des Données. Part V: l'Analyse des Correspondances." *Les Cahiers de l'Analyse Des Données* 2 (1): 9–40.

———. 1977b. "Sur l'Analyse des Tableaux Binaires Associés à une Correspondance Multiple." *Les Cahiers de l'Analyse Des Données* 2 (1): 55–71.

Burt, C. 1912. "The Inhertitance of Mental Characters." *Eugenics Review* 4: 168–200.

———. 1950. "The Factorial Analysis of Qualitative Data." *British Journal of Statistical Psychology* 3: 166–85.

———. 1953. "Scale Analysis and Factor Analysis." *The British Journal of Statistical Psychology* 6: 5–23.

Carroll, J. D. 1968. "A Generalization of Canonical Correlation Analysis to Three or More Sets of Variables." In *Proceedings of the 76th Annual Convention of the American Psychological Association*, 227–28. Washington, D.C.: American Psychological Association.

De Leeuw, J. 1973. "Canonical Analysis of Categorical Data." PhD thesis, University of Leiden, The Netherlands.

———. 1982. "Nonlinear Principal Component Analysis." In *COMPSTAT 1982*, edited by H. Caussinus, P. Ettinger, and R. Tomassone, 77–86. Vienna, Austria: Physika Verlag.

———. 1988a. "Multivariate Analysis with Linearizable Regressions." *Psychometrika* 53: 437–54.

———. 1988b. "Multivariate Analysis with Optimal Scaling." In *Proceedings of the International Conference on Advances in Multivariate Statistical Analysis*, edited by S. Das Gupta and J. K. Ghosh, 127–60. Calcutta, India: Indian Statistical Institute.

———. 2004. "Least Squares Optimal Scaling of Partially Observed Linear Systems." In *Recent Developments in Structural Equation Models*, edited by K. van Montfort, J. Oud, and A. Satorra. Dordrecht, Netherlands: Kluwer Academic Publishers.

———. 2006. "Nonlinear Principal Component Analysis and Related Techniques." In *Multiple Correspondence Analysis and Related Methods*, edited by M. Greenacre and J. Blasius, 107–33. Boca Raton, FA: Chapman; Hall.

De Leeuw, J., and D. B. Ferrari. 2008. "Using Jacobi Plane Rotations in R." Preprint Series 556. Los Angeles, CA: UCLA Department of Statistics.

De Leeuw, J., and P. Mair. 2009. "Homogeneity Analysis in R: the Package homals." *Journal of Statistical Software* 31 (4): 1–21.

De Leeuw, J., G. Michailidis, and D. Y. Wang. 1999. "Correspondence Analysis Techniques." In *Multivariate Analysis, Design of Experiments, and Survey Sampling*, edited by S. Ghosh, 523–47. Marcel Dekker.

Edgerton, H. A., and L. E. Kolbe. 1936. "The Method of Minimum Variation for the Combination of Criteria." *Psychometrika* 1 (3): 183–87.

Fisher, R. A. 1938. *Statistical Methods for Research Workers*. 6th edition. Oliver & Boyd.

———. 1940. "The Precision of Discriminant Functions." *Annals of Eugenics* 10: 422–29.

Gifi, A. 1990. *Nonlinear Multivariate Analysis*. New York, N.Y.: Wiley.

Gower, J. C. 1990. "Fisher's Optimal Scores and Multiple Correspondence Analysis." *Biometrics* 46: 947–61.

Guttman, L. 1941. "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction." In *The Prediction of Personal Adjustment*, edited by P. Horst, 321–48. New York: Social Science Research Council.

———. 1950. "The Principal Components of Scale Analysis." In *Measurement and Prediction*, edited by S. A. Stouffer and Others. Princeton: Princeton University Press.

———. 1953. "A Note on Sir Cyril Burt's 'Factorial Analysis of Qualitative Data'." *The*

British Journal of Statistical Psychology 6: 1–4.

Hill, M. O. 1973. "Reciprocal Averaging: An Eigenvector Method of Ordination." *Journal of Ecology* 61 (1): 237–49.

Hill, M. O., and H. G. Gauch. 1980. "Detrended Correspondence Analysis: An Improved Ordination Technique." *Vegetatio* 42: 47–58.

Hirschfeld, H. O. 1935. "A Connection between Correlation and Contingency." *Proceedings of the Cambridge Philosophical Society* 31: 520–24.

Horst, P. 1935. "Measuring Complex Attitudes." *Journal of Social Psychology* 6 (3): 369–74.

———. 1936. "Obtaining a Composite Measure from a Number of Different Measures of the Same Attribute." *Psychometrika* 1 (1): 54–60.

———, ed. 1941. *The Prediction of Personal Adjustment*. Social Science Research Council.

Hotelling, H. 1936. "Relations Between Two Sets of Variates." *Biometrika* 28: 321–77.

Johnson, P. O. 1950. "The Quantification of Qualitative Data in Discriminant Analysis." *Journal of the American Statistical Association* 45: 65–76.

Lancaster, H. O. 1958. "The Structure of Bivariate Distributions." *Annals of Mathematical Statistics* 29: 719–36.

———. 1969. *The Chi-Squared Distribution*. Wiley.

Le Roux, B., and H. Rouanet. 2010. *Multiple Correspondence Analysis*. Sage.

Lebart, L. 1975. "L'Orientation du Depouillement de Certaines Enquetes par l'Analyse des Correspondences Multiples." *Consommation*, no. 2: 73–96.

———. 1976. "Sur les Calculs Impliques par la Description de Certains Grands Tableaux." *Annales de l'INSEE*, no. 22-23: 255–71.

Lebart, L., and G. Saporta. 2014. "Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis." In *Visualization and Verbalization of Data*, edited by J. Blasius and M. Greenacre, 31–44. CRC Press.

Mair, P., and J. De Leeuw. 2010. "A General Framework for Multivariate Analysis with Optimal Scaling: The r Package Aspect." *Journal of Statistical Software* 32 (9): 1–23. http://www.stat.ucla.edu/~deleeuw/janspubs/2010/articles/mair_deleeuw_A_10.pdf.

Maung, K. 1941. "Discriminant Analysis of Tocher's Eye Colour Data for Scottish School Children." *Annals of Eugenics* 11: 64–76.

Naouri, J. C. 1970. "Analyse Factorielle des Correspondences Continues." *Publications de l'Institute de Statistique de l'Université de Paris* 19: 1–100.

Nishisato, S. 1980. *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto, Canada: University of Toronto Press.

Peschar, J. L. 1975. *School, Milieu, Beroep*. Groningen, The Netherlands: Tjeek Willink.

Revelle, W. 2021. *psychTools:Tools to Accompany the 'psych' Package for Psychological Research* .

Richardson, M. W., and G. F. Kuder. 1933. "Making a Rating Scale That Measures." *Personnel Journal* 12: 36–40.

Sarmanov, O. V., and Z. N. Bratoeva. 1967. "Probabilistic Properties of Bilinear Expansions of Hermite Polynomials." *Theory of Probability and Its Applications* 12 (32): 470–81.

Tenenhaus, M., and F. W. Young. 1985. "An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data." *Psychometrika* 50: 91–119.

Van Buuren, S., and J. De Leeuw. 1992. "Equality Constraints in Multiple Correspondence Analysis." *Multivariate Behavioral Research* 27: 567–83.

Wilks, S. S. 1938. "Weighting Systems for Linear Functions of Correlated Variables when there is no Dependent Variable." *Psychometrika* 3 (1): 23–40.

Young, F. W., Y. Takane, and J. De Leeuw. 1978. "The Principal Components of Mixed Measurement Level Multivariate Data: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika* 45: 279–81.