Simple and Canonical Correspondence Analysis Using the R Package anacor

Jan de Leeuw
University of California, Los Angeles

Patrick Mair
Wirtschaftsuniversität Wien

Abstract

This paper presents the R package `anacor` for the computation of simple and canonical correspondence analysis with missing values. The canonical correspondence analysis is specified in a rather general way by imposing covariates on the rows and/or the columns of the two-dimensional frequency table. The package allows for scaling methods such as standard, Benzécri, centroid, and Goodman scaling. In addition, along with well-known two- and three-dimensional joint plots including confidence ellipsoids, it offers alternative plotting possibilities in terms of transformation plots, Benzécri plots, and regression plots.

Keywords: anacor, simple correspondence analysis, canonical correspondence analysis, R.

1. Introduction

Correspondence Analysis (CA; Benzécri 1973) is a multivariate descriptive method based on a data matrix with non-negative elements and related to principal component analysis (PCA). Basically, CA can be computed for any kind of data but typically it is applied to frequencies formed by categorical data. Being an exploratoy tool for data analysis, CA emphasizes two- and three-dimensional graphical representation of the results.

In this paper we revise briefly mathematical foundations of simple CA and canonical CA in terms of singular value decomposition (SVD). The main focus is on the computational implementation in R (R Development Core Team 2007), on scaling methods based on Benzécri distances, centroid principles, and Fischer-Maung decomposition and on the elaboration of corresponding graphical. More details about CA, various extensions and related methods can be found in Greenacre (1984), Gifi (1990) and Greenacre and Blasius (2006) and numerous practical issues are discussed in Greenacre (2007).

Recently, several R packages have been implemented and updated, respectively. The **ca** package by Nenadić and Greenacre (2006) allows for the computation of simple CA using SVD on standardized residuals. Multiple CA is carried out in terms of SVD on either the indicator matrix or the Burt matrix. Joint CA, which can be regarded as variant of multiple CA excluding the diagonal cross tabulations when establishing the Burt matrix, can be performed as well as subset CA. The package provides two- and three-dimensional plots of standard and principal coordinates with various scaling options.

The **ade4** package (Chessel, Dofour, and Thioulouse 2004; Dray, Dofour, and Chessel 2007) which has been developed within an ecological context, allows for multiple CA, canonical CA, disriminant CA, fuzzy CA and other extensions. Another related package is **vegan** (Dixon 2003), also developed within the field of ecology, which allows for constrained and partially constrained CA as well. Another related package is **homals** (de Leeuw and Mair 2008) which fits models of the Gifi-family (homogeneity analysis aka multiple CA, nonlinear PCA, nonlinear canonical correlation analysis). Additional CA-related packages and functions in R can be found in Mair and Hatzinger (2007).

The **anacor** package we present offers, compared to the packages above, additional possibilites for

scaling the scores in simple CA and canonical CA, additional graphical features, and allows for missing values which are imputed using the Nora's algorithm (Nora 1975).

2. Simple Correspondence Analysis

2.1. Basic Principles of Simple CA

The input unit of analysis is a bivariate frequency table F having n rows ($i = 1, \ldots, n$) and m columns ($j = 1, \ldots, m$). Thus the f_{ij} are non-negative integers. Without loss of generality we suppose that $n \geq m$. The row marginals $f_{i\bullet}$ are collected in a $n \times n$ diagonal matrix D and the column marginals $f_{\bullet j}$ in a $m \times m$ diagonal matrix E. Suppose u_n and u_m are vectors of lengths n and m with all elements equal to 1. It follows that the grand total can be written as $\mathfrak{n} = u_n' F u_m$.

Suppose we want to find row scores and column scores such that the correlation in the bivariate table F is as large as possible. This means maximizing $\lambda(x, y) = \mathfrak{n}^{-1} x' F y$ over the row score vector x and column score vector y. These vectors are *centered* by means of

$$u_n' D x = 0, \tag{1a}$$
$$u_m' E y = 0, \tag{1b}$$

and *normalized* on the grand mean by

$$x' D x = \mathfrak{n}, \tag{1c}$$
$$y' E y = \mathfrak{n}. \tag{1d}$$

Such vectors, i.e. both centered and normalized, are called *standardized*. The optimal x and y must satisfy the centering and normalization conditions in (1), as well as the stationary equations

$$F y = \xi_x D x + \mu_x D u, \tag{2a}$$
$$F' x = \xi_y E y + \mu_y E u, \tag{2b}$$

where $(\xi_x, \xi_y, \mu_x.\mu_y)$ are Lagrange multipliers. By using the side constraints (1) we find that the Lagrange multipliers must satisfy $\xi_x = \xi_y = \sigma(x, y)$ and $\mu_x = \mu_y = 0$. Thus we can solve the simpler system

$$F y = \sigma D x, \tag{3a}$$
$$F' x = \sigma E y, \tag{3b}$$

together with the side conditions in (1). The system in (3) is a *singular value problem*. We find the stationary values of σ as the singular values of

$$Z = D^{-\frac{1}{2}} F E^{-\frac{1}{2}}. \tag{4}$$

Since $m \leq n$, we have the singular value decomposition $Z = P \Sigma Q'$. P is $n \times n$ and composed of the left singular vectors; Q is $m \times m$ and composed of the right singular vectors. Both matrices are orthonormal, i.e. $P'P = Q'Q = I$. Σ is the diagonal matrix containing the $\min(n, m) = m$ singular values in descending order.

The m solutions of the stationary equations (3) can be collected in

$$X = \sqrt{\mathfrak{n}} D^{-\frac{1}{2}} P, \tag{5a}$$
$$Y = \sqrt{\mathfrak{n}} E^{-\frac{1}{2}} Q, \tag{5b}$$

where X is the $n \times m$ of row scores and Y is $m \times m$. Except for the case of multiple singular values, the solutions are uniquely determined. If (x, y, σ) solves (3) we shall call it a *singular triple*,

while the two vectors (x, y) are a *singular pair*. In total there are $s = 0, \ldots, m-1$ singular triples (x_s, y_s, σ_s) where x_s and y_s are the columns of X and Y respectively.

We still have to verify if the m columns of X and Y satisfy the standardization conditions in (1). First, $X'DX = \mathfrak{n}P'P = \mathfrak{n}I$ and $Y'EY = \mathfrak{n}Q'Q = \mathfrak{n}I$, which means both X and Y are normalized. In fact we have *orthonormality*, i.e. if (x_s, y_s, σ_s) and $(x_{s'}, y_{s'}, \sigma_{s'})$ are different singular triples, then $x_s'Dx_{s'} = 0$ and $y_s'Ey_{s'} = 0$.

To investigate centering, we observe that $(u_n, u_m, 1)$ is a singular triple, which is often called the *trivial solution*, because it does not depend on the data. All other singular triples (x_s, y_s, λ_s) with $\sigma_s < 1$ are consequently orthogonal to the trivial one, i.e. satisfy $u_n'Dx = 0$ and $u_m'Ey = 0$. If there are other singular triples $(x_s, y_s, 1)$ with perfect correlation, then x_s and y_s can always be chosen to be orthogonal to u_n and u_m as well. It follows that all singular triples define stationary values of σ, except for $(u_n, u_m, 1)$ which does not satisfy the centering conditions.

The squared singular values σ^2 correspond to the eigenvalues λ of $Z'Z$ and ZZ', respectively. Let us denote the corresponding diagonal matrix of eigenvalues by Λ. In classical CA terminology (see e.g. Greenacre 1984) these eigenvalues are referred to as *principal inertias*. By ignoring λ_0 based on the trivial triple $(x_0, y_0, 1)$, the Pearson decomposition can be established by means of

$$\mathfrak{n}\sum_{s=1}^{m-1}\sigma_s^2 = \mathfrak{n}\sum_{s=1}^{m-1}\lambda_s = \mathfrak{n}(\mathbf{tr}\ Z'Z - 1) = \chi^2(F). \tag{6}$$

$\chi^2(F)$ is called *total inertia* and corresponds to the Pearson chi-square statistic for independence of the table F with $df = (n-1)(m-1)$. The single composites are the contributions of each dimension to the total inertia. Correspondingly, for each dimension a percentage reflecting the contribution of dimension s to the total intertia can be computed. The larger the eigenvalue, the larger the contribution. In practical applications, a "good" CA solution is characterized by large eigenvalues for the first few dimensions.

2.2. Methods of Scaling in Simple CA

The basic plot in CA is the *joint plot* which draws parts of X and Y jointly in a low-dimensional Euclidean space. Note that instead of joint plot sometimes the term *CA map* is used. Both symmetric and asymmetric CA maps can be drawn with the **ca** package and corresponding descriptions are given in Nenadić and Greenacre (2006).

We provide additional methods for scaling X and Y which lead to different interpretations of the distances in the joint plot. Ideally we want the dominant geometric features of the plot (distances, angles, projections) to correspond with aspects of the data. So let us look at various ways of plotting row-points and column-points in p dimensions using the truncated solutions X_p which is $n \times p$, and Y_p which is $m \times p$.

In the simplest case we can use the standardized solution of X_p and Y_p without any additional rescaling and plot the coordinates into a device. This corresponds to a symmetric CA map and the coordinates are referred to as *standard coordinates*.

An additional option of scaling is based on *Benzécri distances*, also known as *chi-square distances*. A Benzécri distance between two rows i and k is defined by

$$\delta_{ik}^2(F) = \sum_{j=1}^{m}\left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{kj}}{f_{i\bullet}}\right)^2 / f_{\bullet j}. \tag{7}$$

If we use e_i and e_k for unit vectors of length n, then

$$\begin{aligned}\delta_{ik}^2(F) &= (e_i - e_k)'D^{-1}FE^{-1}F'D^{-1}(e_i - e_k) = \\ &= (e_i - e_k)'D^{-\frac{1}{2}}ZZ'D^{-\frac{1}{2}}(e_i - e_k) = \\ &= (e_i - e_k)'D^{-\frac{1}{2}}P\Sigma^2P'D^{-\frac{1}{2}}(e_i - e_k) = \\ &= (e_i - e_k)'X\Sigma^2X'(e_i - e_k).\end{aligned}$$

Thus, the Benzécri distances between the rows of F are equal to the Euclidean distances between the rows of $X\Sigma$. Again, X_p is the row scores submatrix and Σ_p the submatrix containing the first $p \leq m-1$ singular values. Based on these matrices fitted Benzécri distances can be computed. It follows that

$$d_{ik}(X_1\Sigma_1) \leq d_{ik}(X_2\Sigma_2) \leq \cdots \leq d_{ik}(X_{m-1}\Sigma_{m-1}) = \delta_{ik}(F). \tag{8}$$

In the same way the Euclidean distances between the rows of $Y\Sigma$ approximate the Benzécri distances between the columns of F. In CA terminology this type of coordinates is sometimes referred to as *principal coordinates* of rows and columns. Based on these distances we can compute a Benzécri root mean squared error (RSME) for the rows and columns separately (see also de Leeuw and Meulman 1986). For the rows it can be expressed as

$$RMSE = \sqrt{\frac{1}{n(n-1)} \sum_i \sum_k (\delta_{ik}(Z) - \delta_{ik}(X_p))^2}. \tag{9}$$

A third way to scale the scores is based on the *centroid principle*. The row centroids (averages) expressed by means of the column scores are $\overline{X}(Y) = D^{-1}FY$. In the same way, the column centroids are given by $\overline{Y}(X) = E^{-1}F'X$. These equations will be used in Section 5.1 to produce the regression plot. Using this notation, the stationary equations can be rewritten as

$$\overline{X}(Y) = X\Sigma, \tag{10a}$$

$$\overline{Y}(X) = Y\Sigma. \tag{10b}$$

This shows that for each singular triple (x, y, σ) the regression of y on x and the regression of x on y are both linear, go through the origin, and have slopes λ and λ^{-1}. Depending on whether X and/or Y are centered, the distances between the points in the joint plot can be interpreted as follows. Suppose that we plot the standard scores of X_p together with $\overline{Y}(X_p)$. Distances between column points approximate Benzécri distances and distances between row points and column points can be interpreted in terms of the centroid principle. Observe that in this scaling the column points will be inside the convex hull of the row points, and if the singular values are small, column points will form a much smaller cloud than row-points.

The same applies if we reverse the role of X_p and Y_p. If we plot $\overline{Y}(X_p)$ and $\overline{X}(Y_p)$ in the plane, then distances between row points in the plane approximate Benzécri distances between rows and distances between column points in the plane approximate Benzécri distances between columns. Unfortunately, distances between row points and column points do not correspond directly to simple properties of the data.

A further possibility of scaling is *Goodman scaling* which starts with the *Fisher-Maung decomposition*. Straightforwardly, $Z = P\Sigma Q'$ can be rewritten as $D^{-\frac{1}{2}}FE^{-\frac{1}{2}} = D^{\frac{1}{2}}X\Sigma Y'E^{\frac{1}{2}}$. It follows that $F = DX\Sigma Y'E$. Now we plot the row-points $X_p\Sigma_p^{\frac{1}{2}}$ and the column points $Y_p\Sigma_p^{\frac{1}{2}}$. The scalar product of the two sets of points approximates $X\Sigma Y'$, which is the matrix of Pearson residuals

$$\frac{Nf_{ij}}{f_{i\bullet}f_{\bullet j}} - 1. \tag{11}$$

For this Goodman scaling there does not seem to be an obvious interpretation in terms of distances. This is somewhat unfortunate because people find distances much easier to understand and work with than scalar products.

It goes without saying that if the singular values in Σ_p are close to one, the four different joint plots will be similar. Generally, plots based on the symmetric Benzécri and Goodman scalings will tend to be similar, but the asymmetric centroid scalings can lead to quite different plots.

3. Canonical Correspondence Analysis

3.1. Basic Principles of Canonical CA

Ter Braak (1986) presented canonical CA within an ecological context having the situation where the whole dataset consists of two sets: data on the occurrence or abundance of a number of species, and data on a number of environmental variables measured which may help to explain the interpretation of the scaled solution. In other words, they are incorporated as effects in the CA computation in order to examine their influence on the scores.

To give a few examples outside ecology, in behavioral sciences such environmental variables could be various types schools, in medical sciences different hospitals etc. Thus, from this particular point of view canonical CA reflects multilevel situations in some sense; from a general point of view it reflects any type of effects on the rows and/or columns of the table. We introduce canonical CA from the general perspective of having covariates A on the row margins $f_{i\bullet}$ and/or covariates B on the column margins $f_{\bullet j}$. Hence, canonical CA can be derived by means of a linear regression of A and B on the row scores X and the column scores Y, i.e.

$$X = AU, \tag{12a}$$
$$Y = BV, \tag{12b}$$

where A and B are known matrices of dimensions $n \times a$ and $m \times b$, respectively, and U and V are weights. We suppose, without loss of generality, that A and B are of full column rank. We also suppose that u_n is in the column-space of A and u_m is in the column-space of B. Note that ordinary CA is a special case of canonical CA in which both A and B are equal to the identity.

By using basically the same derivation as in the previous section, we find the singular value problem

$$(A'FB)V = (A'DA)U\Sigma, \tag{13a}$$
$$(B'F'A)U = (B'EB)V\Sigma. \tag{13b}$$

Analogous to Section 2.1, X and Y, expressed by means of (12), satisfy the standardization conditions $U'A'DAU = \mathfrak{n}I$ and $V'B'EBV = \mathfrak{n}I$. If $u_n = Ag$ and $u_m = Bh$, then (g, h) defines a solution to (13) with $\sigma = 1$. Thus we still have the dominant trivial solution which makes sure that all other singular pairs are centered.

The problem that we have to solve is the SVD on Z which for canonical CA can be expressed as

$$Z = (A'DA)^{-\frac{1}{2}}A'FB(B'EB)^{-\frac{1}{2}} \tag{14}$$

using the inverse of the symmetric square roots of $A'DA$ and $B'EB$. Suppose again that $Z = P\Sigma Q'$ is the singular value decomposition of Z. Then $U = (A'DA)^{-\frac{1}{2}}P$ and $V = (B'EB)^{-\frac{1}{2}}Q$ are the optimal solutions for the weights in our maximum correlation problem, and the corresponding scores are $X = A(A'DA)^{-\frac{1}{2}}P$ and $Y = B(B'EB)^{-\frac{1}{2}}Q$. Both X and Y are normalized, orthogonal, and, except for the dominant solution, centered. Again, X and Y are the standard coordinates which can be rescaled by means of the principles described in Section 3.2.

If we assume, for convenience, that u_n is the first column of A and u_m is the first column of B, then the elements of the first row and column of Z are zero, except for element z_{11}, which is equal to one. The other $(a-1)(b-1)$ elements of Z are, under the hypothesis of independence, asymptotically independent $N(0,1)$ distributed. Thus

$$\mathfrak{n}\sum_{s=1}^{p}\sigma_s^2 = \mathfrak{n}\sum_{s=1}^{p}\lambda_s = \mathfrak{n}(\mathbf{tr}\ Z'Z - 1) = \chi^2\left(F_{A,B}\right), \tag{15}$$

which is asymptotically a chi-square with $df = (a-1)(b-1)$. Hence, in canonical CA we compute a canonical partition of the components of chi-square corresponding with orthogonal contrasts A and B.

3.2. Methods of Scaling in Canonical CA

In this section the same methods of rescaling of row and column scores used for simple CA, are applied to canonical CA. Again, we start with Benzécri distances $\delta^2_{ik}(F_{AB})$ between two rows i and k and using unit vectors e_i and e_k of length n:

$$\begin{aligned}\delta^2_{ik}(F_{AB}) &= (e_i - e_k)'(A'DA)^{-1}A'FB(B'EB)^{-1}B'F'A(A'DA)^{-1}(e_i - e_k) = \\ &= (e_i - e_k)'(A'DA)^{-\frac{1}{2}}ZZ'(A'DA)^{-\frac{1}{2}}(e_i - e_k) = \\ &= (e_i - e_k)'(A'DA)^{-\frac{1}{2}}P\Sigma^2P'(A'DA)^{-\frac{1}{2}}(e_i - e_k) = \\ &= (e_i - e_k)'X\Sigma^2X'(e_i - e_k) = \\ &= (e_i - e_k)'AU\Sigma^2U'A'(e_i - e_k).\end{aligned}$$

Analogous to (8) the monotonicity property holds for the distances for the first p singular values in terms of the row scores submatrix X_p and the singular value submatrix Σ_p. The Benzécri distances for the columns can be derived in an analogous manner.

For the centroid principle we rewrite the stationary equations in (13) as follows (cf. Equation 10):

$$A(A'DA)^{-1}A'DX^\star = X\Sigma, \tag{16a}$$
$$B(B'EB)^{-1}B'EY^\star = Y\Sigma, \tag{16b}$$

where

$$X^\star = D^{-1}FY = \overline{X}(Y), \tag{16c}$$
$$Y^\star = E^{-1}F'X = \overline{Y}(X), \tag{16d}$$

and of course $X = AU$ and $Y = BV$. We see that the columns of X are proportional to the projections in the metric D of $X^\star$ on the space spanned by the columns of A. The same applies to the column scores Y. Note that if we solve the linear regression regression problem of minimizing

$$\mathbf{tr}\ (X^\star - AT)'D(X^\star - AT) \tag{17}$$

then the minimizer is $T = (A'DA)^{-1}A'DX^\star$. As a solution of the stationary equations is follows that $T = U\Sigma$. Ter Braak (1986) calls T the *canonical coefficients*. In our more general setup there are also canonical coefficients for the columns, which are the regression coefficients when regressing $Y^\star$ on B.

Within the context of canonical CA there are various matrices of correlation coefficients that can be computed to give *canonical loadings*. For the rows, we can correlate $X, X^\star$, and A. Now $X'DX^\star = X'FY = \Sigma$. We know that $X'DX = I$, but generally $X^\star$ is not normalized, and thus the correlations are not equal to Σ. In fact, using the Loewner order, $(X^\star)'DX^\star = Y'F'D^{-1}FY \lesssim Y'EY = I$ and, since $\Lambda = (X^\star)'DA'(A'DA)^{-1}AD$ by 16a, also $(X^\star)'DX^\star \gtrsim \Lambda$. If the columns of A are centered and normalized, the correlations become Σ. For the columns, the situation is the same for $Y, Y^\star$ and B.

The Fisher-Maung decomposition is merely a rewriting of the singular value decomposition. The most obvious generalization in the constrained case uses

$$(A'DA)^{-\frac{1}{2}}A'FB(B'EB)^{-\frac{1}{2}} = P\Sigma Q', \tag{18a}$$

or

$$(A'DA)^{-1}A'FB(B'EB)^{-1} = U\Sigma V', \tag{18b}$$

or

$$A'FB = (A'DA)U\Sigma V'(B'EB) = A'(DX\Sigma Y'E)B. \tag{18c}$$

This can be written as $A'RB = 0$ with

$$r_{ij} = f_{ij} - f_{i\bullet}f_{\bullet j}(1 + \sum_{s=1}^{c-1}\sigma_s x_{is}y_{js}), \tag{19}$$

where $c = min(a, b)$.

Note that the joint plots pertaining to the different scaling methods are again based on the p-dimensional solution with the corresponding row scores X_p based on the linear combination of matrix A, and the corresponding column scores Y_p based on the linear combination of matrix B.

4. Additional Topics

4.1. Confidence Ellipsoids Using the Delta Method

The core computation in the **anacor** package is the SVD on $Z = P\Sigma Q'$. As a result we get the $n \times n$ matrix P of left singular vectors, the $m \times m$ matrix Q of the right singular vectors, the diagonal matrix Σ of order m containing the singular values, and, correspondingly, the eigenvalue matrix Λ. Based on these results the $n \times p$ row score matrix X_p and the $m \times p$ column score matrix are computed (standard scores). At this point an important issue is the replication stability of the results in terms of confidence ellipsoids around the standard scores in the joint plot.

A general formal framework to examine stability in multivariate methods is given in Gifi (1990, Chapter 12). The starting point of the replication stability is the well-known *delta method*. Let us assume that we have a sequence of multivariate random variables $\mathbf{x}_\mathfrak{n}$, it follows that $\sqrt{\mathfrak{n}}(\mathbf{x}_\mathfrak{n} - \mu) \xrightarrow{D} N(\mathbf{0}, \Sigma)$. If we apply a transformation $\phi(\mathbf{x}_n)$ the delta method states that $\sqrt{\mathfrak{n}}(\phi\mathbf{x}_\mathfrak{n} - \phi(\mu)) = \sqrt{\mathfrak{n}}\nabla\phi(\mu)(\mathbf{x}_\mathfrak{n} - \mu) \xrightarrow{D} N(\mathbf{0}, \nabla\phi(\mu)'\Sigma\nabla\phi(\mu))$. In simple words: The delta method provides the transformed variance-covariance (VC) matrix which is based on the gradient of ϕ evaluated at μ.

To apply this method for CA we have to embed our observations $p_{ij} = f_{ij}/\mathfrak{n}$ into a sequence of random variables, i.e. a sequence of multinomial distributed random variables with cell probabilities π_{ij}. Asymptotic theory states that $\sqrt{\mathfrak{n}}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{D} N(\mathbf{0}, \Pi - \boldsymbol{\pi}\boldsymbol{\pi}')$ where $\mathbf{p}$ and $\boldsymbol{\pi}$ are the vectors of relative frequencies and probabilities and Π is the diagonal matrix with the elements of $\boldsymbol{\pi}$ on the diagonal.

The SVD system of transformations ϕ we use is $ZQ = P\Sigma$ and $Z'P = Q\Sigma$ with $P'P = Q'Q = I$. Expressions for the partial derivatives $\partial\phi/\partial p_{ij}$ as well as other related derivatives are given in de Leeuw (2008).

4.2. Reconstitution Algorithm for Incomplete Tables

As an additional feature of the **anacor** package, incomplete tables are allowed. The algorithm we use was proposed by Nora (1975) and revised by de Leeuw and van der Heiden (1988). This algorithm should not be mistaken for the CA reconstitution formula which allows for the reconstruction of the data matrix from the scores. Nora's algorithm is rather based on the complementary use of CA and log-linear models (see van der Heiden and de Leeuw 1985) and provides a decomposition of the residuals from independence. We will describe briefly the *reconstitution of order 0* which is implemented in **anacor**.

We start at iteration $l = 0$ by setting the missing values in F to zero. The corresponding table which will be iteratively updated is denoted by $F^{(0)}$. Correspondingly, the row margins are $f_{i\bullet}^{(0)}$, the column margins $f_{\bullet j}^{(0)}$ and the grand mean $f_{\bullet\bullet}^{(0)}$. The elements of the new table $F^{(1)}$ are computed under independence. Pertaining to iteration l, this corresponds to

$$f_{ij}^{(l+1)} = \frac{f_{i\bullet}^{(l)} f_{\bullet j}^{(l)}}{f_{\bullet\bullet}^{(l)}}. \tag{20}$$

Within each iteration a measure for the change in the frequencies is computed, i.e. $H^{(l)} = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}^{(0)} \log f_{ij}^{(l)}$. The iteration stops if $|H^{(l)} - H^{(l-1)}| < \epsilon$. After reaching convergence, we set $F := F^{(l)}$ and we proceed with the computations from Section 2 and Section 3, respectively.

5. Applications of Simple and Canonical CA

5.1. Plotting options in anacor

The basic function in the package is `anacor` which performs simple or canonical CA with different scaling options. The `NA` cells in the table will be imputed using the reconstitution algorithm. The results are stored in an object of class `"anacor"`. For objects of these class a `print.anacor` and a `summary.anacor` method is provided. Two-dimensional plots can be produced with `plot.anacor`, static 3-D plots with `plot3dstatic.anacor` and dynamic `rgl`-plots with `plot3d.anacor`. The type of the plot can be specified by the argument `plot.type`:

- `"jointplot"`: Plots row and column scores into the same device (also available as 3-D).
- `"rowplot"`, `"colplot"`: Plots the row/column scores into separate devices (also available as 3-D).
- `"graphplot"`: This plot type is an unlabeled version of the joint plot where the points are connected by lines. Options are provided (i.e. `wlines`) to steer the line thickness indicating the connection strength.
- `"regplot"`: First, the unscaled solution is plotted. A frequency grid for the row categories (x-axis) and column categories (y-axis) is produced. The regression line is based on the category weighted means of the relative frequencies. More precise, the black line on the column-wise means (x-axis) and the column category on the y-axis, the red line is based on the row categories (x-axis) and the row-wise means on the y-axis. In a second device the scaled solution is plotted. The frequency grid is determined by the row scores (x-axis) and the column scores(y-axis). Now, instead of the row/column categories, the column scores (black line y-axis) and the row scores (red line x-axis) are used (see centroid principle in Section 2.2).
- `"transplot"`: The transformation plot plots the initial row/column categories against the scaled row/column scores.
- `"benzplot"`: The Benzécri plot shows the observed distances against the fitted Benzécri distances; assumed that the row and/or columns in the CA result are Benzécri scaled. For the rows the observed distances are based on $D^{-\frac{1}{2}}ZZ'D^{-\frac{1}{2}}$ and the fitted distances on $X_p\Sigma_p^2X_p'$; for the columns on $E^{-\frac{1}{2}}ZZ'E^{-\frac{1}{2}}$ and Y_pY_p', respectively.

In addition, **anacor** offers various CA utility functions: `expandFrame()` expands a data frame into a indicator supermatrix, `burtTable()` establishes the Burt matrix, and `mkIndiList()` returns a list of codings with options for crisp indicators, numerical versions, and fuzzy coding.

5.2. Applications of Simple CA

We start with an application of simple CA on Tocher's eye color data (Maung 1941) collected on 5387 Scottish school children. This frequency table consists of the eye color in the rows (blue, light, medium, dark) and the hair color in the columns (fair, red, medium, dark, black).

```
> library(anacor)
> data(tocher)
> res <- anacor(tocher, scaling = c("standard", "centroid"))
> res
```

```
CA fit:
Sum of eigenvalues:  0.2293315

Total chi-square value: 1240.039

Chi-Square decomposition:
               Chisq Proportion Cumulative Proportion
Component 1 1073.331      0.866                0.866
Component 2  162.077      0.131                0.996
Component 3    4.630      0.004                1.000
```

```
> plot(res, plot.type = "jointplot", ylim = c(-1.5, 1.5))
> plot(res, plot.type = "graphplot", wlines = 5)
```

For this two-dimensional solution we use asymmetric scaling by having standard coordinates for the rows and principal coordinates for the columns. As graphical representation methods the joint plot including 95% confidence ellipsoids and the graph plot are chosen (see Figure 1).

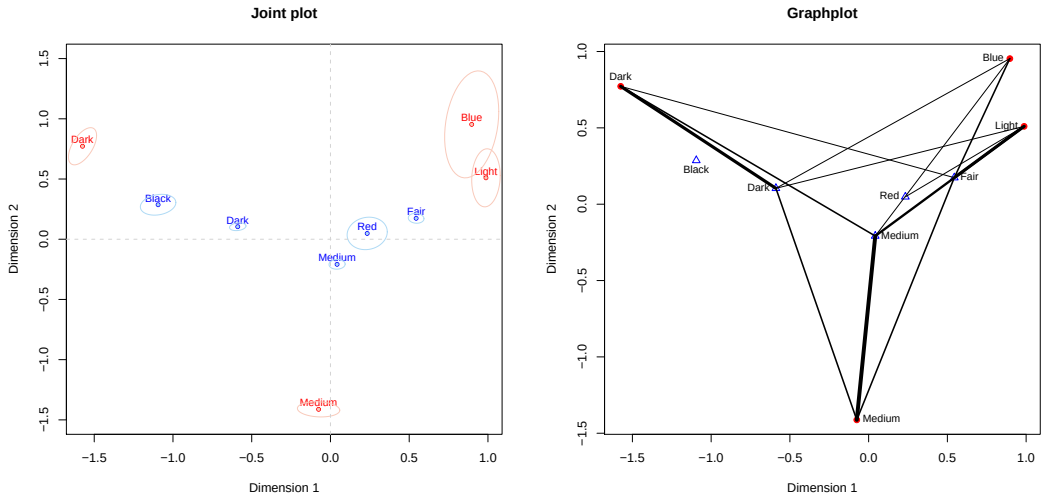

Figure 1: Joint Plot and Graph Plot for Tocher Dataset.

As mentioned above the coordinates of the points in both plots are the same. Note that the column scores (blue points) in the joint plot are scaled around their centroid. The row scores (red points) are not rescaled. In the graph plot the columns scores are represented by blue triangles and the row scores by red points. The thickness of the connecting lines reflect the frequency of the table or, in other words, the strength of the connection. The distances within row/column categories can be interpreted and we see that black/dark hair as well as fair/red hair are quite close to each other. The same applies to blue/light eyes. The distances between single row and column categories can not be interpreted.

We can run a χ^2-test of independence

```
> chisq.test(tocher)

        Pearson's Chi-squared test

data:  tocher
X-squared = 1240.039, df = 12, p-value < 2.2e-16
```

and see that it is highly significant. Looking at the χ^2-decomposition of the CA result we see that the first component accounts for 88.6% of the total χ^2-value (i.e. inertia).

In a second example we show two CA solutions for the Bitterling dataset (Wiepkema 1961) which concerns the reproductive behavior of male bitterlings. The data are derived from 13 sequences using a moving time-window of size two (time 1 in rows, time 2 in columns) and are organized in a 14×14 table with the following categories: jerking (jk), turning beats (tu), head butting (hb), chasing (chs), fleeing (ft), quivering (qu), leading (le), head down posture (hdp), skimming (sk), snapping (sn), chafing (chf), and finflickering (ffl).

We fit a two-dimensional and a five-dimensional CA solution using Benzécri scaling. With two dimensions we explain 53.2% of the total inerita (sum of eigenvalues is 1.33) and with five dimensions we explain 85.8% (sum of eigenvalues is 2.15).

```
> data(bitterling)
> res1 <- anacor(bitterling, ndim = 2, scaling = c("Benzecri",
+     "Benzecri"))
> res2 <- anacor(bitterling, ndim = 5, scaling = c("Benzecri",
+     "Benzecri"))
> res2

CA fit:
Sum of eigenvalues:  2.147791
Benzecri RMSE rows:  0.0002484621
Benzecri RMSE columns:  0.000225833

Total chi-square value: 14589.07

Chi-Square decomposition:
                Chisq Proportion Cumulative Proportion
Component 1  4026.287      0.276                0.276
Component 2  3730.218      0.256                0.532
Component 3  1996.814      0.137                0.669
Component 4  1635.673      0.112                0.781
Component 5  1145.514      0.079                0.859
Component 6   904.313      0.062                0.921
Component 7   832.702      0.057                0.978
Component 8   284.566      0.020                0.998
Component 9    31.421      0.002                1.000
Component 10    1.357      0.000                1.000
Component 11    0.206      0.000                1.000

> plot(res1, plot.type = "benzplot", main = "Benzecri Distances (2D)")
> plot(res2, plot.type = "benzplot", main = "Benzecri Distances (5D)")
```

The improvement of the five-dimensional solution with respect to the two-dimensional one is reflected by the Benzécri plots in Figure 2. For a perfect (saturated) solution the points would lie on the diagonal. This plot can be used as an overall goodness-of-fit plot or, alternatively, single distances can be interpreted.

The data for the next example were collected by Glass (1954). In this 7×7 table the occupational status of fathers (rows) and sons (columns) of 3497 British families were cross-classified. The categories are professional and high administrative (PROF), managerial and executive (EXEC), higher supervisory (HSUP), lower supervisory (LSUP), skilled manual and routine non-manual (SKIL), semi-skilled manual (SEMI), and unskilled manual (UNSK).

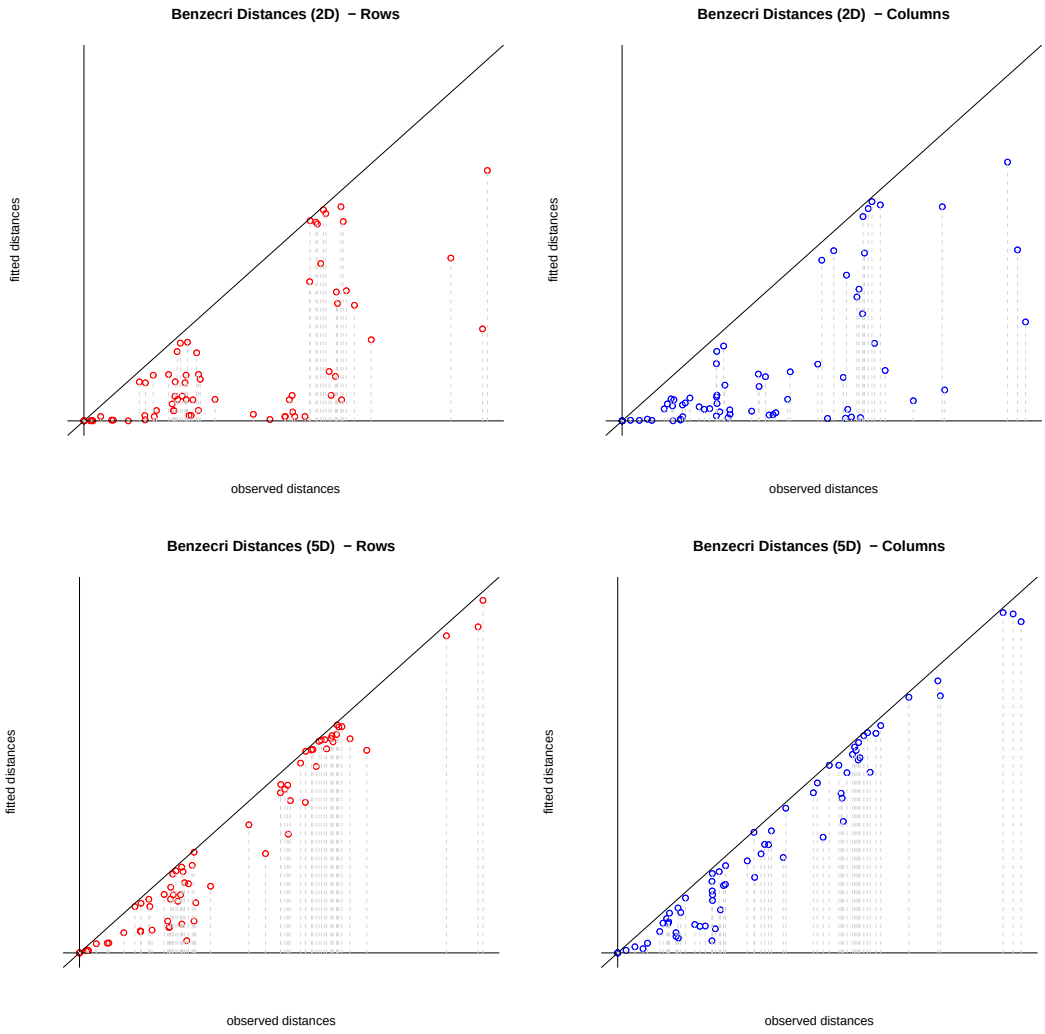

Figure 2: Benzécri Plots for Bitterling Data.

```
> data(glass)
> res <- anacor(glass)
> plot(res, plot.type = "regplot", xlab = "fathers occupation",
+     ylab = "sons occupation")
```

Figure 3 represents regression plots for the first CA dimension. On the left hand side we show the unscaled solution. The father's occupation is on the abscissae and the occupation of the sons on the ordinate. The grid represents the (transposed) table with the corresponding frequencies. Let us focus on the red line first: The coordinates in x-direction correspond to single row categories aka father's occupation. Now, for each father occupation (i.e. conditional) the category-weighted average of the (relative) frequencies is computed. The weights range from 1 to m. The corresponding points are connected and we see that the son's occupation increases monotonically conditional on the father's occupation. The same applies to the black line. Conditional on each son's occupation the relative frequencies are weighted from 1 to n. The average values are plotted in x-direction and are again monotonically increasing. The monotonicity is not surprising since the categories (professions) are ordered in the table (from PROF down to UNSK) and the variables are highly

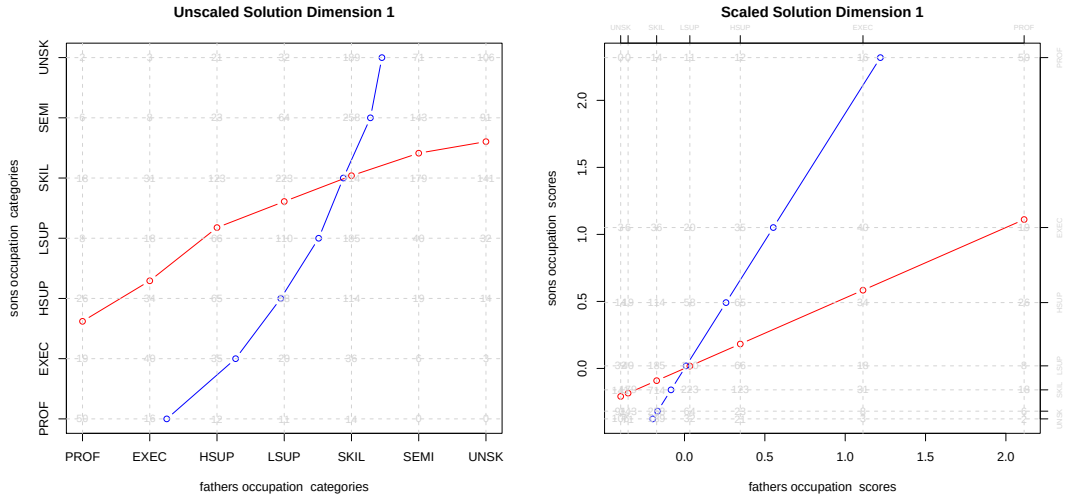

Figure 3: Regression Plots for Glass Data.

dependent ($\chi^2 = 1361.742$, $df = 36$, $p < 0.000$).

On the right hand side of Figure 3 we find the scaled solution. The first obvious characteristic is that the grid components are not equidistant anymore due to the category scaling. The ordering of the professions in terms of the scaled values is given on the top and the right, respectively. Compared to the unscaled solution they are reversed. By means of these grid margins we see that the differences between PROF, EXEC, and HSUP are considerably large compared to lower profession levels such as UNSK, SEMI, and SKIL. The regression lines are computed in an analogous fashion than in the unscaled solution; with the exception that the category scores are taken as weights. The red line is composed of the weighted averages conditional on the row scores on the abscissae, the black line by the weighted averages conditional on the columns scores on the ordinate. This leads to two linear "regressions" with the row/column scores as predictors.

As a final interpretation we see that there is a positive relationship between the intra-familiar occupations: The higher the father's occupation level, the higher the son's occupation level. More detailed, if the father occupies one of the three highest levels, the son is (on the average) in the level below. For the three lowest levels we have the opposite case: On the average the son is in the next higher level.

5.3. Canonical CA on Maxwell Data

A hypothetical dataset by Maxwell (1961) is used to demonstrate his method of discriminant analysis. We will use it to illustrate canonical CA. The data consist of three criterion groups (columns), i.e. schizophrenic, manic-depressive and anxiety state; and four binary predictor variables each indicating either presence or absence of a certain symptom. The four symptoms are anxiety suspicion, schizophrenic type of thought disorders, and delusions of guilt. These four binary variables were factorially combined to form 16 distinct patterns of symptoms (predictor patterns), and each of these patterns is identified with a row of the table. In total we have a cross-classification of 620 patients according to the 16 patterns of symptoms and the three criterion groups.

We fit a symmetric (Goodman scaled) two-dimensional solution and get an amount of explained inertia of 87.2%.

```
> data(maxwell)
> res <- anacor(maxwell$table, row.covariates = maxwell$row.covariates,
```

```
+     scaling = c("Goodman", "Goodman"))
> res

CA fit:
Sum of eigenvalues:  0.6553413

Total chi-square value: 406.312

Chi-Square decomposition:
              Chisq Proportion Cumulative Proportion
Component 1 302.568      0.650                0.650
Component 2 103.743      0.223                0.872

> plot(res, plot.type = "colplot", xlim = c(-1.5, 1), arrows = TRUE,
+     conf = NULL)
> plot(res, plot.type = "transplot", legpos = "topright")
```

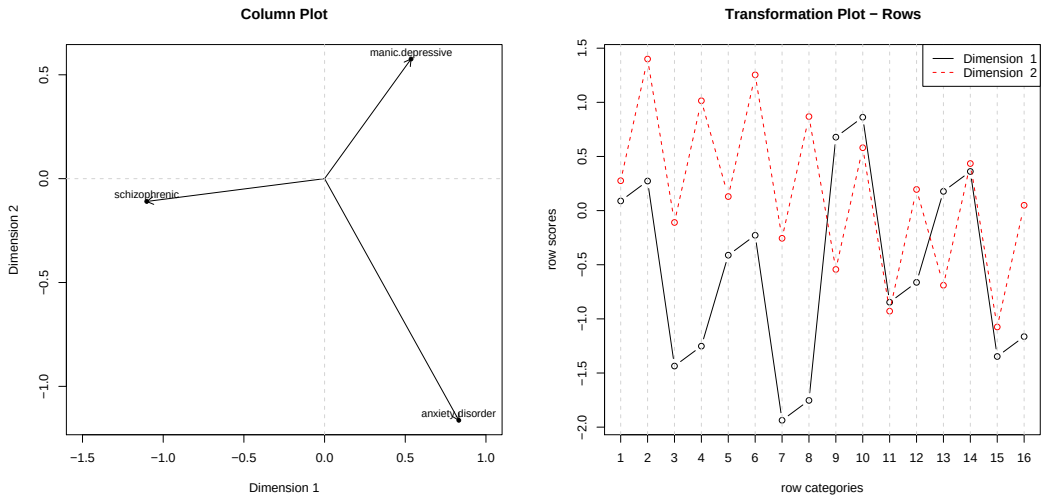

Figure 4: Column Scores and Transformation Plot for Maxwell Data.

The plot of the column scores on the left hand side of Figure 4 shows that the mental diseases go into somewhat different directions and thus they are not really related to each other. The transformation plot on the right hand side shows interesting patterns. For the first dimension a cyclic behavior over the predictors is identifiable. The scores (y-axis) for pairs of points 1-2, 3-4, 5-6, etc. do not change much within these pairs. Note that these pairs are contrasted by the (fourth) predictor "delusions of guilt". Between these pairs some obvious differences in the scores are noticeable. These between-pairs-differences are contrasted by the (third) predictor "thought disorders": 1-2 has 0, 3-4 has 1, 5-6 has 0 etc. Therefore, the first dimension mainly reflects thought disorders.

The second dimension shows an alternating behavior. Referring to the pair notation above, it reflects within-pair-differences based on "delusions of guilt". In addition a slight downward trend due to "anxiety" (first predictor) can be observed.

6. Discussion

The **anacor** package provides additional features which are not offered by other CA packages

on CRAN. These features are additional scaling methods for simple and canonical CA, missing data, and graphical representations such as regression plots, Benzécri plots, transformation plots, and graphplots. The included utilities make it possible to switch from the data format used in **anacor** to the data format used in **homals**, and this gives the user a great deal of flexibility. The confidence ellipsoids from CA are a powerful tool to visualize the dispersions of the row and column projections in the plane.

References

Benzécri JP (1973). *Analyse des Données*. Dunod, Paris, France.

Chessel D, Dofour AB, Thioulouse J (2004). "The ade4 package I: One-table methods." *R News*, **4/1**, 5–10.

de Leeuw J (2008). "Derivatives of generalized eigen systems, with applications." *Journal of Statistical Software*, **forthcoming**.

de Leeuw J, Mair P (2008). "Homogeneity Analysis in R: The package homals." *Journal of Statistical Software*, **forthcoming**.

de Leeuw J, Meulman J (1986). "Principal component analysis and restricted multidimensional scaling." In W Gaul, M Schrader (eds.), "Classification as a tool of research," pp. 83–96. North Holland Publishing Company, Amsterdam.

de Leeuw J, van der Heiden P (1988). "Correspondence analysis of incomplete contingency tables." *Psychometrika*, **53**, 223–233.

Dixon P (2003). "VEGAN: A package of R functions for community ecology." *Journal of Vegetation Science*, **14**, 927–930.

Dray S, Dofour AB, Chessel D (2007). "The ade4 package II: Two-table and K-table methods." *R News*, **7/2**, 47–52.

Gifi A (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester, England.

Glass DV (1954). *Social Mobility in Britain*. Free Press, Glencoe.

Greenacre M (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London, England.

Greenacre M (2007). *Correspondence Analysis in Practice*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.

Greenacre M, Blasius J (2006). *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, Boca Raton, FL.

Mair P, Hatzinger R (2007). "Psychometrics Task View." *R News*, **forthcoming**.

Maung K (1941). "Discriminant analysis of Tocher's eye colour data for Scottish school children." *Annals of Eugenics*, **11**, 64–67.

Maxwell AE (1961). "Canonical variate analysis when the variables are dichotomous." *Educational and Psychological Measurement*, **21**, 259–271.

Nenadić O, Greenacre M (2006). "Correspondence analysis in R, with two- and three-dimensional Graphics: The **ca** package." *Journal of Statistical Software*, **20(3)**, 1–13.

Nora C (1975). *Une méthode de reconstitution et d'analyse de données incomplètes [A method for reconstruction and for the analysis of incomplete data]*. Unpublished Thèse d'Etat, Université P. et M. Curie, Paris VI.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ter Braak CJF (1986). "Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradients." *Ecology*, **67**, 1167–1179.

van der Heiden P, de Leeuw J (1985). "Correspondence analysis used complementary to loglinear analysis." *Psychometrika*, **50**, 429–447.

Wiepkema PR (1961). "An ethological analysis of the reproductive behavior of the bitterling (rhodeus amarus bloch)." *Archives Neerlandais Zoologique*, **14**, 103–199.