The Gifi Package for Categorical Multivariate Analysis in R

Patrick Mair
Harvard University

Jan de Leeuw
University of California, Los Angeles

Abstract

This package vignette is an update and extension of the paper by published in the Journal of Statistical Software. Homogeneity analysis combines the idea of maximizing the correlations between variables of a multivariate data set with that of optimal scaling. In this article we present methodological and practical issues of the R package **homals** which performs homogeneity analysis and various extensions. By setting rank constraints nonlinear principal component analysis can be performed. The variables can be partitioned into sets such that homogeneity analysis is extended to nonlinear canonical correlation analysis or to predictive models which emulate discriminant analysis and regression models. For each model the scale level of the variables can be taken into account by setting level constraints. All algorithms allow for missing values.

Keywords: Gifi methods, optimal scaling, homogeneity analysis, correspondence analysis, nonlinear principal component analysis, nonlinear canonical correlation analysis, homals, R.

```
## Loading required package:  Gifi
## Loading required package:  ape
```

1. Introduction

In recent years correspondence analysis (CA) has become a popular descriptive statistical method to analyze categorical data (Benzécri 1973; Greenacre 1984; Gifi 1990; Greenacre and Blasius 2006). Due to the fact that the visualization capabilities of statistical software have increased during this time, researchers of many areas apply CA and map objects and variables (and their respective categories) onto a common metric plane.

Currently, R (R Development Core Team 2015) offers a variety of routines to compute CA and related models. An overview of functions and packages is given in Mair and Hatzinger (2007). The package **ca** (Nenadic and Greenacre 2006) is a comprehensive tool to perform simple and multiple CA. Various two- and three-dimensional plot options are provided.

In this paper we present the R package **homals**, starting from the simple homogeneity analysis, which corresponds to a multiple CA, and providing several extensions. Gifi (1990) points out that homogeneity analysis can be used in a *strict* and a *broad* sense. In a strict sense homogeneity analysis is used for the analysis of strictly categorical data, with a particular loss function and a particular algorithm for finding the optimal solution. In a broad sense homogeneity analysis refers to a class of criteria for analyzing multivariate data in general,

sharing the characteristic aim of optimizing the homogeneity of variables under various forms of manipulation and simplification (Gifi 1990, p. 81). This view of homogeneity analysis will be used in this article since **homals** allows for such general computations. Furthermore, the two-dimensional as well as three-dimensional plotting devices offered by R are used to develop a variety of customizable visualization techniques. More detailed methodological descriptions can be found in Gifi (1990) and some of them are revisited in Michailidis and de Leeuw (1998).

2. Homogeneity analysis

In this section we will focus on the underlying methodological aspects of **homals**. Starting with the formulation of the loss function, the classical alternating least squares algorithm is presented in brief and the relation to CA is shown. Based on simple homogeneity analysis we elaborate various extensions such as nonlinear canonical analysis and nonlinear principal component analysis. A less formal introduction to Gifi methods can be found in Mair and de Leeuw (2009).

2.1. Establishing the loss function

Homogeneity analysis is based on the criterion of minimizing the departure from homogeneity. This departure is measured by a loss function. To write the corresponding basic equations the following definitions are needed. For $i = 1, \ldots, n$ objects, data on m (categorical) variables are collected where each of the $j = 1, \ldots, m$ variable takes on k_j different values (their *levels* or *categories*). We code them using $n \times k_j$ binary *indicator matrices* G_j, i.e., a matrix of dummy variables for each variable. The whole set of indicator matrices can be collected in a block matrix

$$G \stackrel{\Delta}{=} \begin{bmatrix} G_1 & \vdots & G_2 & \vdots & \cdots & \vdots & G_m \end{bmatrix}. \tag{1}$$

In this paper we derive the loss function including the option for missing values. For a simpler (i.e. no missings) introduction the reader is referred to Michailidis and de Leeuw (1998, p. 307–314). In the indicator matrix missing observations are coded as complete zero rows; if object i is missing on variable j, then row i of G_j is 0. Otherwise the row sum becomes 1 since the category entries are disjoint. This corresponds to the first missing option presented in Gifi (*missing data passive* 1990, p. 74). Other possibilities would be to add an additional column to the indicator matrix for each variable with missing data or to add as many additional columns as there are missing data for the j-th variable. However, our approach is to define the binary diagonal matrix M_j if dimension $n \times n$ for each variable j. The diagonal element (i, i) is equal to 0 if object i has a missing value on variable j and equal to 1 otherwise. Based on M_j we can define $M_\star$ as the sum of the M_j's and $M_\bullet$ as their average.

For convenience we introduce

$$D_j \stackrel{\Delta}{=} G_j' M_j G_j = G_j' G_j, \tag{2}$$

as the $k_j \times k_j$ diagonal matrix with the (marginal) frequencies of variable j in its main diagonal.

Now let X be the unknown $n \times p$ matrix containing the coordinates (*object scores*) of the object projections into $\mathbb{R}^p$. Furthermore, let Y_j be the unknown $k_j \times p$ matrix containing the

coordinates of the category projections into the same p-dimensional space (*category quantifications*). The problem of finding these solutions can be formulated by means of the following loss function to be minimized:

$$\sigma(X; Y_1, \ldots, Y_m) \triangleq \sum_{j=1}^{m} \mathbf{tr}(X - G_j Y_j)' M_j (X - G_j Y_j) \tag{3}$$

We use the normalization $\mathbf{u}' M_\bullet X = 0$ and $X' M_\bullet X = I$ in order to avoid the trivial solution $X = 0$ and $Y_j = 0$. The first restriction centers the graph plot (see Section 4) around the origin whereas the second restriction makes the columns of the object score matrix orthogonal. Multiplying the scores by $\sqrt{n/m}$ gives a mean of 0 and a variance of 1 (i.e., they are z-scores). Note that from an analytical point of view the loss function represents the sum-of-squares of $(X - G_j Y_j)$ which obviously involves the object scores and the category quantifications. Thus, we minimize simultaneously over X and Y_j. We give a graphical interpretation of the loss function in the following section.

2.2. Geometry of the loss function

In the **homals** package we use homogeneity analysis as a graphical method to explore multivariate data sets. The *joint plot* mapping object scores and the category quantifications in a joint space, can be considered as the classical or standard homals plot. The category points are the centers of gravity of the object points that share the same category. The larger the spread between category points the better a variable discriminates and thus the smaller the contribution to the loss. The closeness of two objects in the plot is related to the similarity of their response patterns. A "perfect" solution, i.e., without any loss at all, would imply that all object points coincide with their category points.

Moreover, we can think of G as the adjacency matrix of a bipartite graph in which the n objects and the k_j categories ($j = 1, \ldots, m$) are the vertices. In the corresponding *graph plot* an object and a category are connected by an edge if the object is in the corresponding category. The loss in (3) pertains to the sum of squares of the line lengths in the graph plot. Producing a *star plot*, i.e., connecting the object scores with their category centroid, the loss corresponds to the sum over variables of the sum of squared line lengths. More detailed plot descriptions are given in Section 4.

2.3. Minimizing the loss function

Typically, the minimization problem is solved by the iterative *alternating least squares algorithm* (ALS; sometimes quoted as *reciprocal averaging algorithm*). At iteration $t = 0$ we start with arbitrary object scores $X^{(0)}$. Each iteration t consists of three steps:

1. Update category quantifications: $Y_j^{(t)} = D_j^{-1} G_j' X^{(t)}$ for $j = 1, \ldots, m$
2. Update object scores: $\tilde{X}^{(t)} = M_\star^{-1} \sum_{j=1}^{m} G_j Y_j^{(t)}$
3. Normalization: $X^{(t+1)} = M_\star^{-\frac{1}{2}} \mathbf{orth}(M_\star^{-\frac{1}{2}} \tilde{X}^{(t)})$

Note that matrix multiplications using indicator matrices can be implemented efficiently as cumulating the sums of rows over X and Y.

Here **orth** is some technique which computes an orthonormal basis for the column space of a matrix. We can use QR decomposition, modified Gram-Schmidt, or the singular value decomposition (SVD). In the **homals** package the left singular vectors of $\tilde{X}^{(t)}$, here denoted as **lsvec**, are used.

To simplify, let P_j denote the orthogonal projector on the subspace spanned by the columns of G_j, i.e., $P_j = G_j D_j^{-1} G_j'$. Correspondingly, the sum over the m projectors is

$$P_\star = \sum_{j=1}^{m} P_j = \sum_{j=1}^{m} G_j D_j^{-1} G_j'. \tag{4}$$

Again, $P_\bullet$ denotes the average. By means of the **lsvec** notation and including $P_\bullet$ we can describe a complete iteration step as

$$X^{(t+1)} = \mathbf{lsvec}(\tilde{X}^{(t)}) = \mathbf{lsvec}(M_\bullet^{-1} P_\bullet X^{(t)}). \tag{5}$$

In each iteration t we compute the value of the loss function to monitor convergence. Note that Formula (5) is not suitable for computation, because it replaces computation with sparse indicator matrices by computations with a dense average projector.

Computing the homals solution in this way is the same as performing a CA on G. Usually, multiple CA solves the generalized eigenproblem for the Burt matrix $C = G'G$ and its diagonal D (Greenacre 1984; Greenacre and Blasius 2006). Thus, we can put the problem in Equation 3 into a SVD context (de Leeuw, Michailidis, and Wang 1999). Using the block matrix notation, we have to solve the generalized singular value problem of the form

$$GY = M_\star X \Lambda, \tag{6}$$
$$G'X = DY\Lambda, \tag{7}$$

or equivalently one of the two generalized eigenvalue problems

$$GD^{-1}G'X = M_\star X \Lambda^2, \tag{8}$$
$$G'M_\star^{-1} GY = DY\Lambda^2. \tag{9}$$

Here the eigenvalues Λ^2 are the ratios along each dimension of the average between-category variance and the average total variance. Also $X'P_jX$ is the between-category dispersion for variable j. Further elaborations can be found in Michailidis and de Leeuw (1998).

Compared to the classical SVD approach, the ALS algorithm only computes the first p dimensions of the solution. This leads to an increase in computational efficiency. Moreover, by capitalizing on sparseness of G, the **homals** package is able to handle large data sets.

The goodness-of-fit of a solution can be examined by means of a screeplot of the eigenvalues. The contribution of each variable to the final solution can be examined by means of discrimination measures defined by $||G_jY_j||^2/n$ (see Meulman 1996).

3. Extensions of homogeneity analysis

Gifi (1990) provides various extensions of homogeneity analysis and elaborates connections to other multivariate methods. The package **homals** allows for imposing restrictions on the

variable ranks and levels as well as defining sets of variables. These options offer a wide spectrum of additional possibilities for multivariate data analysis beyond classical homogeneity analysis (cf. broad sense view in the Introduction).

3.1. Nonlinear (categorical) principal component analysis

Having a $n \times m$ data matrix with metric variables, principal components analysis (PCA) is a common technique to reduce the dimensionality of the data set, i.e., to project the variables into a subspace $\mathbb{R}^p$ where $p \ll m$. The Eckart-Young theorem states that this classical form of *linear* PCA can be formulated by means of a loss function. Its minimization leads to a $n \times p$ matrix of *component scores* and an $m \times p$ matrix of *component loadings*.

However, having nonmetric variables, nonlinear PCA (NLPCA) can be used. The term "nonlinear" pertains to nonlinear transformations of the observed variables (de Leeuw 2006). In Gifi terminology, NLPCA can be defined as homogeneity analysis with restrictions on the quantification matrix Y_j. Let us denote $r_j \leq p$ as the parameter for the imposed restriction on variable j. If no restrictions are imposed, as e.g. for a simple homals solution, $r_j = k_j - 1$ iff $k_j \leq p$, and $r_j = p$ otherwise.

We start our explanations with the simple case for $r_j = 1$ for all j. In this case we say that all variables are *single* and the rank restrictions are imposed by

$$Y_j = \mathbf{z}_j \mathbf{a}_j', \tag{10}$$

where $\mathbf{z}_j$ is a vector of length k_j with category quantifications and $\mathbf{a}_j$ a vector of length p with weights. Thus, each quantification matrix is restricted to rank 1, which allows for the existence of object scores with a single category quantification.

3.2. Multiple quantifications

It is not necessarily needed that we restrict the rank of the score matrix to 1. Our **homals** implementation allows for multiple rank restrictions. We can simply extend Equation 10 to the general case

$$Y_j = Z_j A_j' \tag{11}$$

where again $1 \leq r_j \leq \min(k_j - 1, p)$, Z_j is $k_j \times r_j$ and A_j is $p \times r_j$. We require, without loss of generality, that $Z_j' D_j Z_j = I$. Thus, we have the situation of *multiple quantifications* which implies imposing an additional constraint each time PCA is carried out.

To establish the loss function for the rank constrained version we write $r_\star$ for the sum of the r_j and $r_\bullet$ for their average. The block matrix G of dummy variables now becomes

$$Q \stackrel{\Delta}{=} \begin{bmatrix} G_1 Z_1 & \vdots & G_2 Z_2 & \vdots & \cdots & \vdots & G_m Z_m \end{bmatrix}. \tag{12}$$

Gathering the A_j's in a block matrix as well, the $p \times r_\star$ matrix

$$A \stackrel{\Delta}{=} \begin{bmatrix} A_1 & \vdots & A_2 & \vdots & \cdots & \vdots & A_m \end{bmatrix} \tag{13}$$

results. Then, Equation 3 becomes

$$\begin{aligned}\sigma(X;Z;A)=\sum_{j=1}^m\mathbf{tr}\ (X-G_jZ_jA_j')'M_j(X-G_jZ_jA_j')=\\ =m\mathbf{tr}\ X'M_\star X-2\mathbf{tr}\ X'QA+\mathbf{tr}\ A'A=\\ =mp+\mathbf{tr}\ (Q-XA)'(Q-XA)-\mathbf{tr}\ Q'Q=\\ =\mathbf{tr}(Q-XA)'(Q-XA)+m(p-r_\bullet)\end{aligned}\tag{14}$$

This shows that $\sigma(X;Y_1,\cdots,Y_m)\geq m(p-r_\bullet)$ and the loss is equal to this lower bound if we can choose the Z_j such that Q is of rank p. In fact, by minimizing (14) over X and A we see that

$$\sigma(Z)\overset{\Delta}{=}\min_{X,A}\sigma(X;Z;A)=\sum_{s=p+1}^{r_\star}\lambda_s^2(Z)+m(p-r_\bullet),\tag{15}$$

where the λ_s are the ordered singular values. A corresponding example in terms of a *lossplot* is given in Section 4.

3.3. Level constraints: Optimal scaling

From a general point of view, *optimal scaling* attempts to do two things simultaneously: The transformation of the data by a transformation appropriate for the scale level (i.e. level constraints), and the fit of a model to the transformed data to account for the data. Thus it is a simultaneous process of data transformation and data representation (Takane 2005). In this paper we will take into account the scale level of the variables in terms of restrictions within Z_j. To do this, the starting point is to split up Equation 14 into two separate terms. Using $\hat{Y}_j=D_j^{-1}G_j'X$ this leads to

$$\begin{aligned}&\textstyle\sum_{j=1}^m\mathbf{tr}(X-G_jY_j)'M_j(X-G_jY_j)\\ =&\textstyle\sum_{j=1}^m\mathbf{tr}(X-G_j(\hat{Y}_j+(Y_j-\hat{Y}_j)))'M_j(X-G_j(\hat{Y}_j+(Y_j-\hat{Y}_j)))\\ =&\textstyle\sum_{j=1}^m\mathbf{tr}(X-G_j\hat{Y}_j)'M_j(X-G_j\hat{Y}_j)+\sum_{j=1}^m\mathbf{tr}(Y_j-\hat{Y}_j)'D_j(Y_j-\hat{Y}_j).\end{aligned}\tag{16}$$

Obviously, the rank restrictions $Y_j=Z_jA_j'$ affect only the second term and hence, we will proceed on our explanations by regarding this particular term only, i.e.,

$$\sigma(Z;A)=\sum_{j=1}^m\mathbf{tr}(Z_jA_j'-\hat{Y}_j)'D_j(Z_jA_j'-\hat{Y}_j).\tag{17}$$

Now, level constraints for nominal, ordinal, polynomial, and numerical variables can be imposed on Z_j in the following manner. For nominal variables, all columns in Z_j are unrestricted. Equation 17 is minimized under the conditions $\mathbf{u}'D_jZ_j=0$ and $Z_j'D_jZ_j=I$. The stationary equations are

$$A_j=Y_j'D_jZ_j,\tag{18a}$$

$$Y_jA_j=Z_jW+\mathbf{uh}',\tag{18b}$$

with W as a symmetric matrix of Langrange multipliers. Solving, we find

$$\mathbf{h} = \frac{1}{\mathbf{u}'D_j\mathbf{u}} A_j'Y_j'D_j\mathbf{u} = \mathbf{0}, \tag{19}$$

and thus, letting $\overline{Z}_j \triangleq D_j^{1/2}Z_j$ and $\overline{Y}_j \triangleq D_j^{1/2}Y_j$, it follows that

$$\overline{Y}_j\overline{Y}_j'\overline{Z}_j = \overline{Z}_jW. \tag{20}$$

If $\overline{Y}_j = K\Lambda L'$ is the SVD of $\overline{Y}_j$, then we see that $\overline{Z}_j = K_rO$ with O as an arbitrary rotation matrix and K_r as the singular vectors corresponding with the r largest singular values. Thus, $Z_j = D_j^{-1/2}K_rO$, and $A_j = \overline{Y}_j'\overline{Z}_j = L_r\Lambda_rO$. Moreover, $Z_jA_j' = D_j^{-1/2}K_r\Lambda_rL_r'$.

Having ordinal variables, the first column of Z_j is constrained to be either increasing or decreasing, the rest is free. Again (17) has to be minimized under the condition $Z_j'D_jZ_j = I$ (and optionally additional conditions on Z_j). If we minimize over A_j, we can also solve the problem $\mathbf{tr}(Z_j'D_jY_jY_j'D_jZ_j)$ with $Z_j'D_jZ_j = I$.

For polynomial constraints the matrix Z_j are the first r_j orthogonal polynomials. Thus all p columns of Y_j are polynomials of degree r_j. In the case of numerical variables, the first column in Z_j denoted by $\mathbf{z}_{j0}$ is fixed and linear with the category numbers, the rest is free. Hence, the loss function in (17) changes to

$$\sigma(Z, A) = \sum_{j=1}^{m} \mathbf{tr}(Z_jA_j' + \mathbf{z}_{j0}\mathbf{a}_{j0}' - \hat{Y}_j)'D_j(Z_jA_j' + \mathbf{z}_{j0}\mathbf{a}_{j0}' - \hat{Y}_j). \tag{21}$$

Since column $\mathbf{z}_{j0}$ is fixed, Z_j is a $k_j \times (r_j - 1)$ matrix and A_j, with $\mathbf{a}_{j0}$ as the first column, is $p \times (r_j - 1)$. In order to minimize (21), $\mathbf{z}_{j0}'D_jZ_j = 0$ is required as minimization condition.

Note that level constraints can be imposed additionally to rank constraints. If the data set has variables with different scale levels, the **homals** package allows for setting level constraints for each variable j separately. Unlike in Gifi (1990) and Michailidis and de Leeuw (1998) it is not necessary to have rank 1 restrictions in order to allow for different scaling levels. Our implementation allows for multiple ordinal, multiple numerical etc. level constraints.

3.4. Nonlinear canonical correlation analysis

In Gifi terminology, nonlinear canonical correlation analysis (NLCCA) is called "OVERALS" (van der Burg, de Leeuw, and Verdegaal 1988; van der Burg, de Leeuw, and Dijksterhuis 1994). This is due to the fact that it has most of the other Gifi-models as special cases. In this section the relation to homogeneity analysis is shown. The **homals** package allows for the definition of *sets* of variables and thus, for the computation NLCCA between $g = 1, \ldots, K$ sets of variables.

Recall that the aim of homogeneity analysis is to find p orthogonal vectors in m indicator matrices G_j. This approach can be extended in order to compute p orthogonal vectors in K general matrices G_v, each of dimension $n \times m_v$ where m_v is the number of variables $(j = 1, \ldots, m_v)$ in set v. Thus,

$$G_v \triangleq \begin{bmatrix} G_{v_1} & \vdots & G_{v_2} & \vdots & \cdots & \vdots & G_{v_{m_v}} \end{bmatrix}. \tag{22}$$

The loss function can be stated as

$$\sigma(X;Y_1,\ldots,Y_K)\triangleq\frac{1}{K}\sum_{v=1}^{K}\mathbf{tr}\left(X-\sum_{j=1}^{m_v}G_{v_j}Y_{v_j}\right)'M_v\left(X-\sum_{j=1}^{m_v}G_{v_j}Y_{v_j}\right). \tag{23}$$

X is the $n\times p$ matrix with object scores, G_{v_j} is $n\times k_j$, and Y_{v_j} is the $k_j\times p$ matrix containing the coordinates. Missing values are taken into account in M_v which is the element-wise minimum of the M_j in set v. The normalization conditions are $XM_\bullet X=I$ and $\mathbf{u}'M_\bullet X=0$ where $M_\bullet$ is the average of M_v.

Since NLPCA can be considered as special case of NLCCA, i.e., for $K=m$, all the additional restrictions for different scaling levels can straightforwardly be applied for NLCCA. Unlike classical canonical correlation analysis, NLCCA is not restricted to two sets of variables but allows for the definition of an arbitrary number of sets. Furthermore, if the sets are treated in an asymmetric manner predictive models such as regression analysis and discriminant analysis can be emulated. For $v=1,2$ sets this implies that G_1 is $n\times 1$ and G_2 is $n\times m-1$. Corresponding examples will be given in Section **??**.

3.5. Cone restricted SVD

In this final methodological section we show how the loss functions of these models can be solved in terms of cone restricted SVD. All the transformations discussed above are projections on some convex cone $\mathcal{K}_j$. For the sake of simplicity we drop the j and v indexes and we look only at the second term of the partitioned loss function (see Equation 17), i.e.,

$$\sigma(Z,A)=\mathbf{tr}(ZA'-\hat{Y})'D(ZA'-\hat{Y}), \tag{24}$$

over Z and A, where $\hat{Y}$ is $k\times p$, Z is $k\times r$, and A is $p\times r$. Moreover the first column z_0 of Z is restricted by $z_0\in\mathcal{K}$, with $\mathcal{K}$ as a convex cone. Z should also satisfy the common normalization conditions $u'DZ=0$ and $Z'DZ=I$.

The basic idea of the algorithm is to apply alternating least squares with rescaling. Thus we alternate minimizing over Z for fixed A and over A for fixed Z. The "non-standard" part of the algorithm is that we do not impose the normalization conditions if we minimize over Z. We show below that we can still produce a sequence of normalized solutions with a non-increasing sequence of loss function values.

Suppose $(\hat{Z},\hat{A})$ is our current best solution. To improve it we first minimize over the non-normalized Z, satisfying the cone constraint, and keeping A fixed at $\hat{A}$. This gives $\tilde{Z}$ and a corresponding loss function value $\sigma(\tilde{Z},\hat{A})$. Clearly,

$$\sigma(\tilde{Z},\hat{A})\leq\sigma(\hat{Z},\hat{A}), \tag{25}$$

but $\tilde{Z}$ is not normalized. Now update Z to Z^+ using the weighted Gram-Schmidt solution $\tilde{Z}=Z^+S$ for Z where S is the Gram-Schmidt triangular matrix. The first column $\tilde{z}_0$ of $\tilde{Z}$ satisfies the cone constraint, and because of the nature of Gram-Schmidt, so does the first column of Z^+. Observe that it is quite possible that

$$\sigma(Z^+,\hat{A})>\sigma(\hat{Z},\hat{A}). \tag{26}$$

This seems to invalidate the usual convergence proof, which is based on a non-increasing sequence of loss function values. But now also adjust $\hat{A}$ to $\overline{A} = \hat{A}(S^{-1})'$. Then $\tilde{Z}\hat{A}' = Z^+\overline{A}'$, and thus

$$\sigma(\tilde{Z}, \hat{A}) = \sigma(Z^+, \overline{A}). \tag{27}$$

Finally compute A^+ by minimizing $\sigma(Z^+, A)$ over A. Since $\sigma(Z^+, A^+) \leq \sigma(Z^+, \overline{A})$ we have the chain

$$\sigma(Z^+, A^+) \leq \sigma(Z^+, \overline{A}) = \sigma(\tilde{Z}, \hat{A}) \leq \sigma(\hat{Z}, \hat{A}). \tag{28}$$

In any iteration the loss function does not increase. In actual computation, it is not necessary to compute $\overline{A}$, and thus it also is not necessary to compute the Gram-Schmidt triangular matrix S.

4. The R package homals

4.1. Categorical Principal Component Analysis

Standard PCA assumes that the data are metric (i.e. equidistant categories within and across variables) and assumes a linear relationship among the observed variables. Having ordinal variables such as Likert items, these assumptions are often not fulfilled in practice. In this case we have two options: run an ordinal factor analysis (FA) based on polychoric correlations as implemented in the `fa.poly()` function in the **psych** package, or run a nonlinear (ordinal) PCA as presented here. Apart from the conceptual differences between FA and PCA in general, the advantage of NLPCA over polychoric FA is that we do not have to pose any underlying distribution assumption on our data, whereas a polychoric (or tetrachoric in the binary case) correlation assumes that the categories are realizations of an underlying latent normal distribution.

In **homals** there is the `princals()` function which performs NLPCA. Since NLPCA is just a rank-1 restricted version of general homogeneity analysis, internally `princals()` uses `homals()` as engine. An effort was made to make the PRINCALS output as PCA-like (i.e. `princomp()`-like) as possible in terms of comparable eigenvalues, explaned amount of variance on each dimension, and loadings.

Standard PCA is solved by an eigenvalue decomposition of the input correlation matrix R based on the original data which gives us eigenvalue vector λ of length m. Subsequently, the amount of explained variance for each dimension can be computed by dividing each eigenvalue by the sum of the m eigenvalues. One of the cores outputs of any Gifi model is that it provides a "new" data matrix where the original categories are optimally scaled for each dimension. Now can now compute the correlation matrix R^* on the new data matrix (it does not matter which one we use since the p matrices are linearly dependent) and perform an eigenvalue decomposition. This gives us the eigenvalue vector λ^* of length m. As above, we can compute the amount of explained variance for each of the m dimensions. In addition, in order to evaluate the amount of "improvement" of NLPCA over PCA we can compute the eigenvalue ratio, e.g. for the first dimension we have λ_1^*/λ_1. This gives us a measure for the violations of equidistance and linearity in our original data.

Regarding the loadings, in standard PCA they are normed to $\|w\|^2 = 1$. In order to make the loadings w^* from NLPCA comparable to standard PCA, they need to be normalized the same way. The `princals()` function performs all these computations internally and returns the p eigenvalues based on the R^*, the amount of variance explained for each of the p dimensions, and the standardized loadings.

Through the `level` argument the user can specify the scale levels of the variables (`"ordinal"` as default. If all variables are set to `"numerical"`, PRINCALS mimics standard PCA. In terms of plotting possibilities, a generic plot function allows for a loadings plot (default), a scree plot, transformation plots, and a biplot by specifying the `plot.type` argument accordingly.

Now we show an ordinal PCA example on the ABC dataset which reproduces the analysis in Ferrari and A. (2012). ABC is a ficticious company which launched a customer satisfaction survey. In this analysis we use six items, each of them on a 5-point Likert scale, covering certain aspects of customer satisfaction: equipment, sales support, technical support, training, purchase, and pricing.

First, we start with a full-dimensional PRINCALS solution and examine the scree plot.

```
ABC6 <- ABC[, 6:11]
fitfull <- princals(ABC6, ndim = 6)
fitfull

## Call: princals(data = ABC6, ndim = 6)
##
## Loss: 0.000667735
## Number of iterations: 19
##
## Eigenvalues:
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6
## 2.31083 1.00248 0.76242 0.73749 0.67682 0.50996

summary(fitfull)

##
## Importance of Components:
##                          Comp.1  Comp.2  Comp.3  Comp.4  Comp.5   Comp.6
## Explained Variance (%)  38.5138 16.7081 12.7070 12.2915 11.2804   8.4993
## Cumulative Variance (%) 38.5138 55.2218 67.9288 80.2203 91.5007 100.0000
```

The scree plot is given in in the left panel of Figure **??**.

Second, we fit a two-dimensional ordinal solution. The loadings plot is given in Figure 1 (right panel).

```
fit2d <- princals(ABC6, ndim = 2)
fit2d

## Call: princals(data = ABC6, ndim = 2)
```

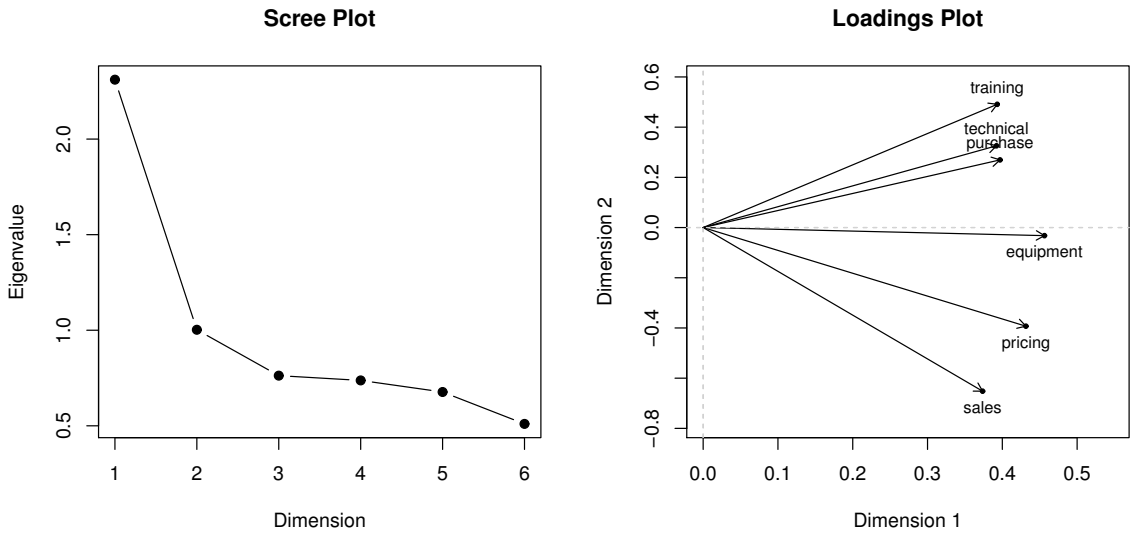

Figure 1: Left panel: Scree plot for full-dimensional PRINCALS. Right panel: Loadings plot for two-dimensional solution.

```
##
## Loss: 0.0005475663
## Number of iterations: 38
##
## Eigenvalues:
##  Comp.1  Comp.2
## 2.94546 0.85430

summary(fit2d)

##
## Importance of Components:
##                            Comp.1  Comp.2
## Explained Variance (%)  49.0909 14.2383
## Cumulative Variance (%) 49.0909 63.3293
```

We see that we explain around 63% of the variance.

```
op <- par(mfrow = c(1,2))
plot(fitfull, plot.type = "screeplot")
plot(fit2d)
par(op)
```

Now we fit a one-dimensional ordinal PCA.

```
fit1d <- princals(ABC6, ndim = 1)
fit1d

## Call: princals(data = ABC6, ndim = 1)
##
## Loss: 0.0004027617
## Number of iterations: 9
##
## Eigenvalues:
##  Comp.1
## 2.98415

summary(fit1d)

##
## Importance of Components:
##                            Comp.1
## Explained Variance (%)  49.7359
## Cumulative Variance (%) 49.7359
```

Let us compare it with the outcome of a standard PCA solution using `princomp()` and compute the ratio of the first eigenvalues.

```
ABC6m <- sapply(ABC6, function(x) as.numeric(levels(x))[x])
fitpc <- princomp(ABC6m)
fitpc

## Call:
## princomp(x = ABC6m)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.6732946 0.9832548 0.8268262 0.7708195 0.6209486 0.5968660
##
##  6  variables and  208 observations.

fit1d$eigenvalues/(fitpc$sdev^2)[1]       ## eigenvalue ratio NLPCA/PCA

##   Comp.1
## 1.065801
```

The eigenvalue ratio of suggests 1.07 a slight improvement of NLPCA over standard PCA. This implies that the response categories are approximately equidistant and the relationship between the variables is not far from linear. If we want to mimic standard PCA with PRINCALS, we declare all the variables as numeric.

```
fit1dm <- princals(ABC6, ndim = 1, level = "numerical")
fit1dm

## Call: princals(data = ABC6, ndim = 1, level = "numerical")
##
## Loss: 0.0004178631
## Number of iterations: 6
##
## Eigenvalues:
##  Comp.1
## 2.87104

fit1dm$eigenvalues/(fitpc$sdev^2)[1]       ## eigenvalue ratio NLPCA/PCA

##   Comp.1
## 1.025404
```

The size of the eigenvalue ratio decreased since we are essentially doing the same thing. Finally, we can also abandon the order assumption in the response categories and treat all variables as nominal. This is the least restrictive of our one-dimensional PRINCALS models.

```
fit1dc <- princals(ABC6, ndim = 1, level = "nominal")
fit1dc

## Call: princals(data = ABC6, ndim = 1, level = "nominal")
##
## Loss: 0.0004026197
## Number of iterations: 11
##
## Eigenvalues:
##  Comp.1
## 2.98521
```

We see that the nominal PCA leads pretty much to the same fit as the ordinal version which suggests that the ordinal scale level for the variables holds.

5. Discussion

In this paper theoretical foundations of the methodology used in the **homals** package are elaborated and package application and visualization issues are presented. Basically, **homals** covers the techniques described in Gifi (1990): Homogeneity analysis, NLCCA, predictive models, and NLPCA. It can handle missing data and the scale level of the variables can be taken into account. The package offers a broad variety of real-life datasets and furthermore provides numerous methods of visualization, either in a two-dimensional or in a three-dimensional way. Future enhancements will be to replace indicator matrices by more general B-spline bases

and to incorporate weights for observations. To conclude, **homals** provides flexible, easy-to-use routines which allow researchers from different areas to compute, interpret, and visualize methods belonging to the Gifi family.

References

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

de Leeuw J (2006). "Nonlinear Principal Component Analysis and Related Techniques." In M Greenacre, J Blasius (eds.), *Multiple Correspondence Analysis and Related Methods*, pp. 107–134. Chapman & Hall/CRC, Boca Raton, FL.

de Leeuw J, Michailidis G, Wang D (1999). "Correspondence Analysis Techniques." In S Ghosh (ed.), *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pp. 523–546. Dekker, New York.

Ferrari PA, A B (2012). "Nonlinear principal component analysis." In RS Kenett, S Salini (eds.), *Modern Analysis of Customer Surveys with Applications in R*, pp. 333–356. Wiley, New York.

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

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

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

Mair P, de Leeuw J (2009). "Rank and Set Restrictions for Homogeneity Analysis in R." In *JSM 2008 Proceedings, Statistical Computing Section*. American Statistical Association., Alexandria, VA.

Mair P, Hatzinger R (2007). "Psychometrics Task View." *R-News*, **7/3**, 38–40.

Meulman JJ (1996). "Fitting a Distance Model to Homogeneous Subsets of Variables: Points of View Analysis of Categorical Data." *Journal of Classification*, **13**, 249–266.

Michailidis G, de Leeuw J (1998). "The Gifi System of Descriptive Multivariate Analysis." *Statistical Science*, **13**, 307–336.

Nenadic 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. URL http://www.jstatsoft.org/v20/i03/paper.

R Development Core Team (2015). *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.

Takane Y (2005). "Optimal Scaling." In B Everitt, D Howell (eds.), *Encyclopedia of Statistics for Behavioral Sciences*, pp. 1479–1482. Wiley, Chichester.

van der Burg E, de Leeuw J, Dijksterhuis G (1994). "OVERALS: Nonlinear Canonical Correlation with k Sets of Variables." *Computational Statistics & Data Analysis*, **18**, 141–163.

van der Burg E, de Leeuw J, Verdegaal R (1988). "Homogeneity Analysis with k Sets of Variables: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika*, **53**, 177–197.