

# Homogeneity Analysis in R: The Package **homals**

Jan de Leeuw

University of California, Los Angeles

Patrick Mair

Wirtschaftsuniversität Wien

---

## Abstract

Homogeneity analysis combines 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:* homogeneity analysis, correspondence analysis, nonlinear principal component analysis, nonlinear canonical correlation analysis, **homals**, R.

---

## 1. Introduction

During the last 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 (respectively their categories) onto a common metric plane.

Currently, R ([R Development Core Team 2007](#)) offers a variety of routines to compute CA and related models. An overview of corresponding 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. Starting from basic homogeneity analysis we elaborate various extensions such as nonlinear canonical analysis and nonlinear principal component analysis.

### 2.1. Establishing the loss function

Homogeneity analysis is based on the criterion of minimizing the departure from homogeneity. Homogeneity is measured by a loss function. To write the corresponding basic equations the following definitions are needed. For  $i = 1, \dots, n$  objects, data on  $m$  (categorical) variables are collected where each of the  $j = 1, \dots, m$  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 dummy matrix for each variable. The whole set of indicator matrices can be collected in a block matrix

$$G \triangleq \begin{bmatrix} G_1 & : & G_2 & : & \dots & : & G_m \end{bmatrix}. \quad (1)$$

Missing observations are coded as complete zero row sums; if object  $i$  is missing on variable  $j$ , then row sum  $i$  of  $G_j$  is 0, otherwise row sum becomes 1 since the category entries are disjunct. This corresponds to the first missing option presented in Gifi (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, all row sums of  $G_j$  are collected in the diagonal matrix  $M_j$ . Suppose  $M_\star$  is the sum of the  $M_j$  and  $M_\bullet$  is their average. Furthermore, we define

$$D_j \triangleq G_j' M_j G_j = G_j' G_j, \quad (2)$$

where  $D_j$  is the diagonal matrix ( $k_j \times k_j$ ) with the relative 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:

$$\sigma(X; Y_1, \dots, Y_m) \triangleq \frac{1}{m} \sum_{j=1}^m \text{tr}(X - G_j Y_j)' M_j (X - G_j Y_j) \quad (3)$$

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.

## 2.2. Geometry of the loss function

In the **homals** package we motivate homogeneity analysis as graphical method to explore multivariate data sets. The *joint plot* where the object scores and the category quantifications are mapped in a joint space, can be considered as the classical or standard homals plot. The category points are the center 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, it indicates how much a variable contributes to relative loss. The distance between two object scores is related to the “similarity” between 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 categories  $k_j$  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:  $Y_j^{(t)} = D_j^{-1} G_j' X^{(t)}$
2. Update object scores:  $\tilde{X}^{(t)} = M_{\star}^{-1} \sum_{j=1}^m G_j Y_j^{(t)}$
3. Normalization:  $X^{(t+1)} = \text{orth}(\tilde{X}^{(t)})$

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 **homals** the left singular vectors of  $\tilde{X}^{(k)}$ , here denoted as **lsvect**, 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

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

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

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

In each iteration 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, Michailides, and Wang 1999). Using the block matrix notation, we have to solve the generalized singular value problem of the form

$$GY = M_\star X \Lambda, \quad (6)$$

$$G'X = DY \Lambda, \quad (7)$$

or equivalently one of the two generalized eigenvalue problems

$$GD^{-1}G'X = M_\star X \Lambda^2, \quad (8)$$

$$G'M_\star^{-1}GY = DY \Lambda^2. \quad (9)$$

Here the eigenvalues  $\Lambda^2$  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 Michailides 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 the sparseness of  $G$ , **homals** is able to handle large data sets.

### 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 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 a  $m \times p$  matrix of *component loadings*.

However, having nonmetric variables, nonlinear PCA (NLPCA) can be used. The term “non-linear” 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

$$Y_j = \mathbf{z}_j \mathbf{a}_j', \quad (10)$$

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.

Straightforwardly, Equation 10 can be extended to the general case

$$Y_j = Z_j A_j' \quad (11)$$

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 dummies now becomes

$$Q \triangleq \begin{bmatrix} G_1 Z_1 & : & G_2 Z_2 & : & \cdots & : & G_m Z_m \end{bmatrix}. \quad (12)$$

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

$$A \triangleq \begin{bmatrix} A_1 & : & A_2 & : & \cdots & : & A_m \end{bmatrix} \quad (13)$$

results. Then, Equation 3 becomes

$$\begin{aligned} \sigma(X; Z; A) &= \sum_{j=1}^m \mathbf{tr}(X - G_j Z_j A_j')' M_j (X - G_j Z_j A_j') = \\ &= \mathbf{tr}(Q - X A')'(Q - X A') + m(p - r_\bullet). \end{aligned} \quad (14)$$

This shows that  $\sigma(X; Y_1, \dots, Y_m) \geq m(p - r_\bullet)$  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

$$\sigma(Z) \triangleq \min_{X,A} \sigma(X; Z; A) = \sum_{s=p+1}^{r_*} \lambda_s^2(Z) + m(p - r_*), \quad (15)$$

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

Now 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 (Gifi 1990; Michailidis and de Leeuw 1998). Using  $\hat{Y}_j = D_j^{-1}G'_jX$  this leads to

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

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.,

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

Now, level constraints for nominal, ordinal, 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  $\mathbf{u}'D_jZ_j = 0$ ,  $Z'_jD_jZ_j = I$ , and  $\mathbf{u}'D_jY_j = 0$ . The stationary equations are

$$A_j = Y'_jD_jZ_j, \quad (18a)$$

$$Y_jA_j = Z_jW + \mathbf{u}\mathbf{h}', \quad (18b)$$

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

$$\mathbf{h} = \frac{1}{\mathbf{u}'D_j\mathbf{u}} A'_jY'_jD_j\mathbf{u} = \mathbf{0}, \quad (19)$$

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

$$\bar{Y}_j\bar{Y}'_j\bar{Z}_j = \bar{Z}_jW. \quad (20)$$

If  $\bar{Y}_j = K\Lambda L'$  is the SVD of  $\bar{Y}_j$ , then we see that  $\bar{Z}_j = K_rO$  with  $O$  an arbitrary rotation matrix. Thus,  $Z_j = D_j^{-1/2}K_rO$ , and  $A_j = \bar{Y}'_j\bar{Z}_j = L_r\Lambda_rO$ . 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  $Z'_jD_jZ_j = I$  (and optionally additional conditions on  $Z_j$ ). If we minimize over  $A_j$ , we can also solve the problem  $\mathbf{tr}(Z'_jD_jY_jY'_jD_jZ_j)$  with  $Z'_jD_jZ_j = I$ .

In the case of numerical variables, the first column in  $Z_j$  denoted by  $\mathbf{z}_{j0}$  is fixed, the rest is free. Hence, the loss function in (17) changes to

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

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  $p \times (r_j - 1)$ . As minimization condition  $\mathbf{z}_{j0}' D_j Z_j = 0$  is required.

Note that level constraints can be imposed additionally to rank constraints. If the data set has variables with different scale levels, **homals** allows for setting level constraints for each variable  $j$  separately.

### 3.2. 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 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  $g = 1, \dots, K$  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 ( $j = 1, \dots, m_v$ ) in set  $v$ . Thus,

$$G_v \triangleq \begin{bmatrix} G_{v_1} & : & G_{v_2} & : & \dots & : & G_{v_{m_v}} \end{bmatrix}. \quad (22)$$

The loss function can be stated as

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

$X$  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  $X M_{\bullet} X = I$  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 onto 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 4.2.

### 3.3. Cone restricted SVD

In this final methodological section we show how the loss functions of these models can be solved in term 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.,

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

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,

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

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

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

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  $\bar{A} = \hat{A}(S^{-1})'$ . Then  $\tilde{Z}\hat{A}' = Z^+\bar{A}'$ , and thus

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

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

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

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



## 4. The R package **homals**

At this point we show how the models described in the sections above can be computed using the package **homals** in R ([R Development Core Team 2007](#)) available on [CRAN](#).

The core routine of the package is **homals**. The extended models can be fitted through appropriate settings on the parameters **sets**, **rank**, and **level**. An object of class "homals" is created and the following methods are provided: **print**, **summary**, **plot**, **plot3d**, **scatterplot3d** and **predict**.

The package offers a wide variety of plots; some of them are discussed in [Michailidis and de Leeuw \(1998\)](#) and [Michailidis and de Leeuw \(2001\)](#). In the **plot** method the user can specify the type of plot through the argument **plot.type**. For some plot types three-dimensional versions are provided in **plot3d** (dynamic) and **plot3dstatic**:

- Object plot ("objplot"): Plots the scores of the objects (rows in data set) on two or three dimensions.
- Category plot ("catplot"): Plots the rank-restricted category quantifications for each variable separately. Three-dimensional plot is available.
- Voronoi plot ("vorplot"): Produces a category plot with Voronoi regions.
- Joint plot ("jointplot"): The object scores and category quantifications are mapped in the same (two- or three-dimensional) device.
- Graph plot ("graphplot"): Like joint plot with additional connections of the scores/quantifications.
- Hull plot ("hullplot"): For each single variable the object scores are mapped onto two dimensions and the convex hull for each response category is drawn.
- Label plot ("labplot"): Similar to object plot, the object scores are plotted but for each variable separately with the corresponding category labels. A three-dimensional version is provided.
- Span plot ("spanplot"): Like label plot, it maps the object scores for each variable and it connects them by the shortest path within each response category.
- Star plot ("starplot"): Again, the object scores are mapped on 2 or 3 dimensions. In addition, these points are connected with the category centroid.
- Loss plot ("lossplot"): Plots the rank-restricted category quantifications against the unrestricted for each variable separately.
- Projection plot ("prjplot"): For variables of rank 1 the category scores (two-dimensional) are projected onto a straight line determined by the rank restricted category quantifications.
- Vector plot ("vecplot"): For variables of rank 1 the object scores (two-dimensional) are projected onto a straight line determined by the rank restricted category quantifications.
- Transformation plot ("trfplot"): Plots variable-wise the original (categorical) scale against the transformed (metric) scale  $Z_j$  for each solution.

- Loadings plot ("loadplot"): Plots the loadings  $\mathbf{a}_j$  and connects them with the origin. Note that if  $r_j > 1$  only the first solution is taken.

#### 4.1. Simple Homogeneity Analysis

The first example is a simple (i.e., no level or rank restrictions, no sets defined) three-dimensional homogeneity analysis for the `senate` data set (ADA 2002). The data consists of 2001 senate votes on 20 issues selected by Americans for Democratic Action. The votes selected cover a full spectrum of domestic, foreign, economic, military, environmental and social issues. We tried to select votes which display sharp liberal/conservative contrasts. As a consequence, Democrat candidates have much more “yes” responses than Republican candidates. A full description of the items can be found in the corresponding package help file. The first column of the data set (i.e., 50 Republicans vs. 49 Democrats and 1 Independent) is inactive and will be used for validation.

```
> data(senate)
> res <- homals(senate, active = c(FALSE, rep(TRUE, 20)), ndim = 3)
> plot3d(res, plot.type = "objplot", sphere = FALSE, bgpng = NULL)
> plot(res, plot.type = "spanplot", plot.dim = c(1, 2), var.subset = 1)
> plot(res, plot.type = "spanplot", plot.dim = c(1, 3), var.subset = 1)
> plot(res, plot.type = "spanplot", plot.dim = c(2, 3), var.subset = 1)
> plot3dstatic(res, plot.type = "loadplot")
```

Figure 1 shows four “wings” of senators which we will denote by north, south, west and east. The west and the north wing are composed by Republicans, the east and south wing by Democrats. Note that the 3D-plot is rotated in a way that Dimension 3 is horizontally aligned, Dimension 2 is vertically aligned, and Dimension 1 is the one aback. The two-dimensional slices show that Dimension 1 vs. 2 does not distinguish between Democrats and Republicans. If Dimension 3 is involved, as in the two bottom plots in Figure 1, the separation between Democrats and Republicans is obvious. To distinguish within north-west and south-east, respectively, Item 19 has to be taken into account:

*V19: S 1438. Military Base Closures. Warner (R-VA) motion to authorize an additional round of U.S. military base realignment and closures in 2003. Motion agreed to 53-47. September 25, 2001. A “yes” vote is a +.*

Republicans belonging to the north wing as well as Democrats belonging to the east wing gave a “yes” vote. South-Wing Democrats and West-Wing Republicans voted with “no”. It is well known that the response on this item mainly depends on whether there is a military base in the senator’s district or not: Those senators who have a military base in their district do not want to close it since such a base provides working places and is an important income source for the district. Hence, these are the determining factors and not the party affiliation. This result is underpinned by Figure 2 where Item 19 is clearly separated from the remaining items.

Given a (multiple) homals solution, we can reconstruct the indicator matrix by assigning each object to the closest points of the variable.

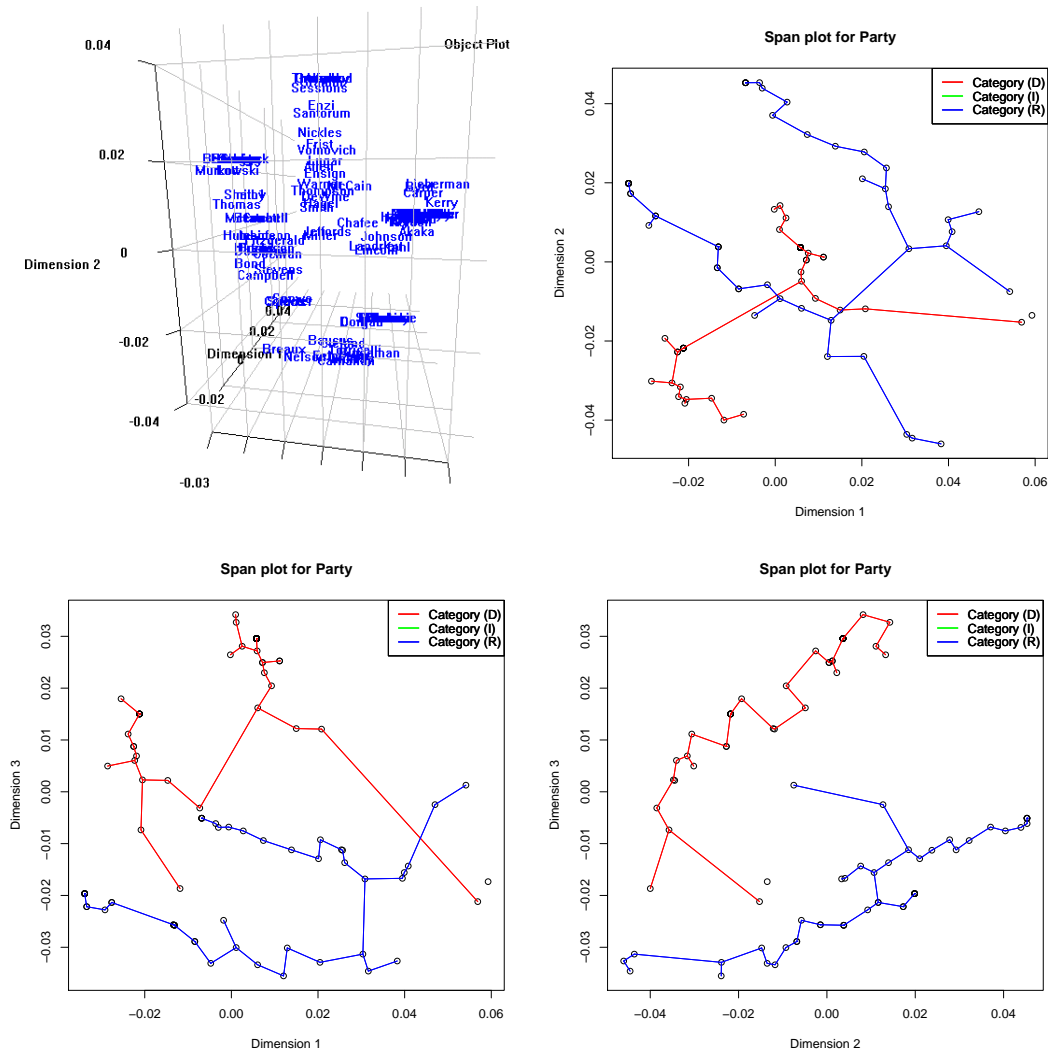


Figure 1: 3D Object Plot and Span Plots for Senate Dataset

```
> p.res <- predict(res)
> p.res$cl.table$Party
```

```
pre
obs  (D) (I) (R)
(D)  49  1  0
(I)   0  1  0
(R)   0  8 41
```

From the classification table we see that 91% of the party affiliations are correctly classified. Note that in the case of such a simple homals solution it can happen that a lower dimensional solution results in a better classification rate than a higher dimensional. The reason is that in simple homals the classification rate is not the criterion to be optimized.

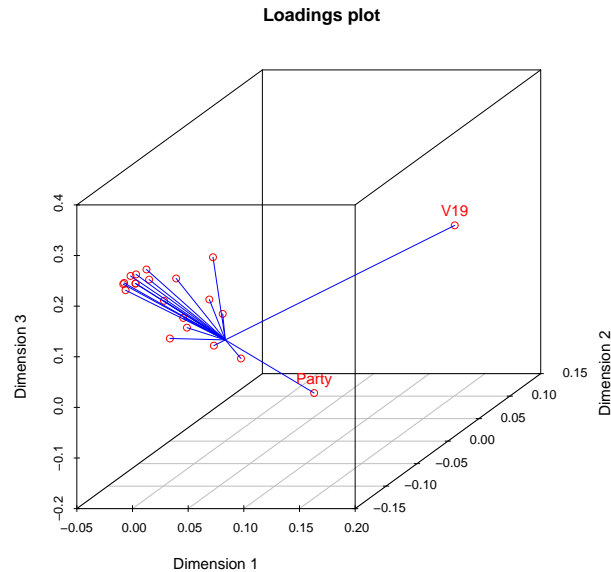


Figure 2: Loadings Plot for Senate Dataset

To show additional plotting features of the **homals** package we run a three-dimensional homogeneity analysis on the mammals dentition dataset ([Hartigan 1975](#)). In this dataset dental characteristics are used in the classification of mammals. The teeth are divided into four groups: incisors, canines, premolars, and molars. Within each group top and bottom teeth are classified.

```
> data(mammals)
> res <- homals(mammals, ndim = 3)
> plot(res, plot.type = "graphplot")
> plot3dstatic(res, plot.type = "starplot", var.subset = 3, box = FALSE)
```

On the left hand side of Figure 3 we have a graph plot where the object scores are drawn as green stars and the category quantifications as red circles. The objects are connected with the respective category responses in the dataset.

For the variable-wise star plot we pick out the top canines with zero canines coded as 1 and one canine coded as 2 (the full coding description can be found in package help file). In the star plot the object scores are connected with the corresponding category centroid. For animals with more than one canine the elk and the reindeer, which has the same object scores, are quite distant from the centroid. All other animals lie close around the category centroid.

## 4.2. Predictive Models and Canonical Correlation

The **sets** argument allows for partitioning the variables into sets in order to emulate canonical correlation analysis and predictive models. As outlined above, if the variables are partitioned into asymmetric sets of one variable vs. the others, we can put this type of homals model into

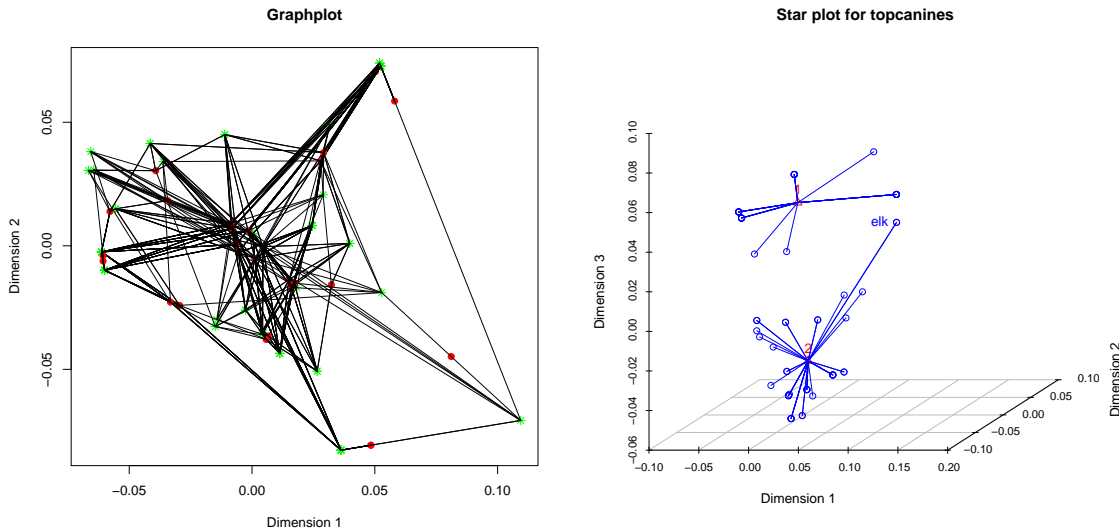


Figure 3: Graph plot and 3D star plot for Mammals data

a predictive modeling context. If not, the interpretation in terms of canonical correlation is more appropriate.

To demonstrate this, we use the `galo` dataset (Peschar 1975) where data of 1290 school children in the sixth grade of elementary school in the city of Groningen (Netherlands) were collected. The variables are Gender, IQ (categorized into 9 ordered categories), Advice (teacher categorized the children into 7 possible forms of secondary education, i.e., Agr = agricultural; Ext = extended primary education; Gen = general; Grls = secondary school for girls; Man = manual, including housekeeping; None = no further education; Uni = pre-University) and SES (parent's profession in 6 categories). In this example it could be of interest to predict Advice from Gender, IQ, and SES.

```
> data(galo)
> res <- homals(galo, active = c(rep(TRUE, 4), FALSE), sets = list(c(1,
+   2, 4), 3, 5))
> plot(res, plot.type = "vorplot", var.subset = 3)
> plot(res, plot.type = "labplot", var.subset = 2)
> predict(res)
```

Classification rate:

	Variable	Cl. Rate	%Cl. Rate
1	gender	0.5690	56.90
2	IQ	0.6333	63.33
3	advice	0.6318	63.18
4	SES	0.2907	29.07
5	School	0.0302	3.02

A rate of .6318 correctly classified teacher advices results. The Voronoi plot in Figure 4

shows the Voronoi regions for the same variable. A labeled plot is given for the IQs which shows that on the upper half of the horseshoe there are mainly children with IQ-categories 7-9. Distinctions between these levels of intelligence are mainly reflected by Dimension 1. For the lower horseshoe half it can be stated that both dimensions reflect differences in lower IQ-categories.

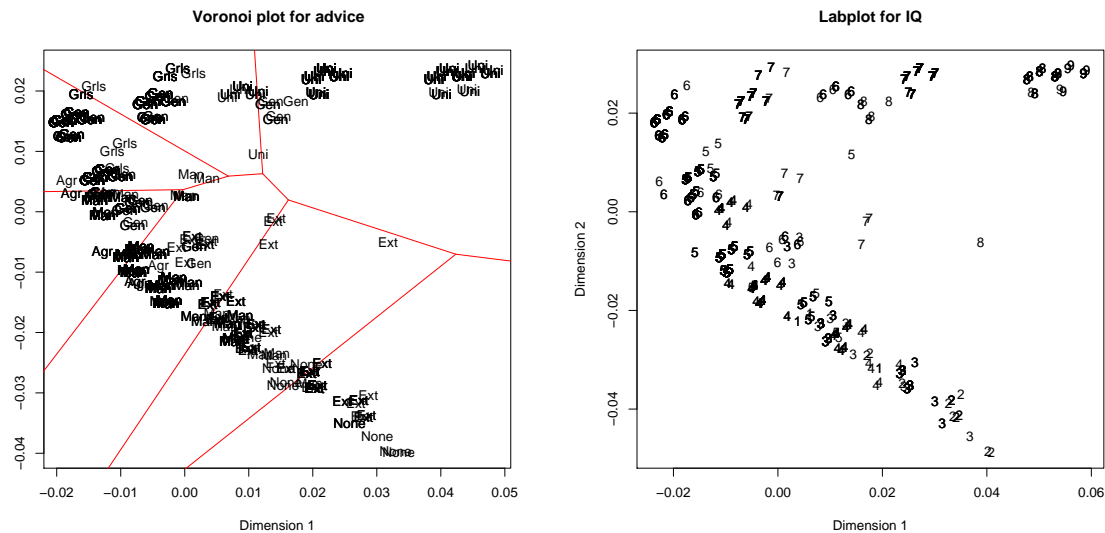


Figure 4: Voronoi Plot and Label Plot for Galo Data

Using the classical iris dataset, the aim is to predict Species from Petal/Sepal Length/Width. The polynomial level constraint is posed on the predictors and the response is treated as nominal. A hull plot for the response, a label plot Petal Length and loss plots for all predictors are produced.

```
> data(iris)
> res <- homals(iris, sets = list(1:4, 5), level = c(rep("polynomial",
+ 4), "nominal"), rank = 2, itermax = 2000)
> plot(res, plot.type = "hullplot", var.subset = 5, cex = 0.7)
> plot(res, plot.type = "labplot", var.subset = 3, cex = 0.7)
> plot(res, plot.type = "lossplot", var.subset = 1:4, cex = 0.7)
```

For this two-dimensional homals solution, 100% of the iris species are correctly classified. The hullplot in Figure 5 shows that the species are clearly separated on the two-dimensional plane. In the label plot the object scores are labeled with the response on Petal Length and it becomes obvious that small lengths form the setosa “cluster”, whereas iris virginica are composed by observations with large petal lengths. Iris versicolor have medium lengths.

The loss plots in Figure 6 show the fitted rank-2 solution (red lines) against the unrestricted solution. The implication of the polynomial level restriction for the fitted model is obvious.

To show another homals application of predictive (in this case regression) modeling we use the Neumann dataset (Wilson 1926): Willard Gibbs discovered a theoretical formula connecting

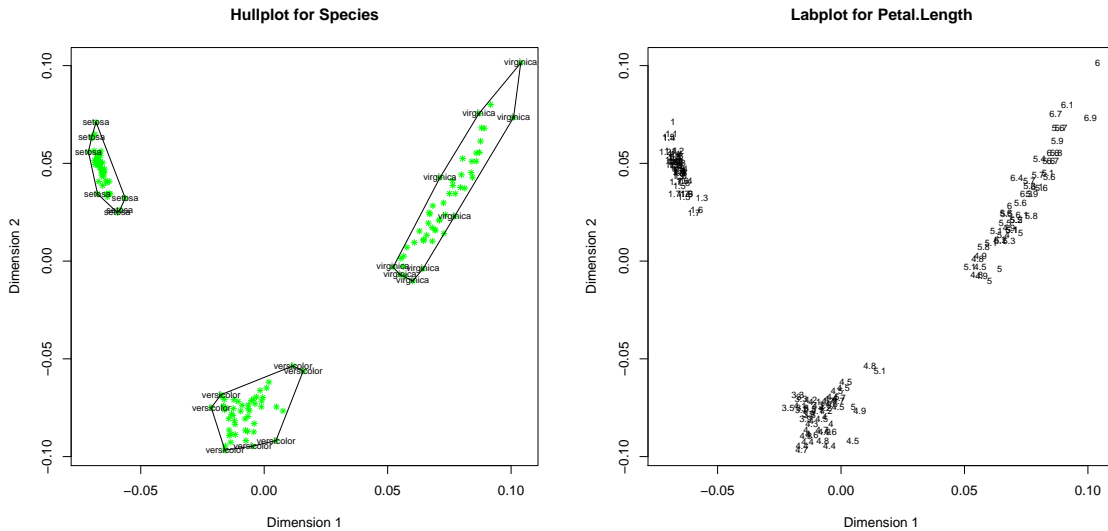


Figure 5: Hullplot and Label Plot for Iris Data

the density, the pressure, and the absolute temperature of a mixture of gases with convertible components. He applied this formula and the estimated constants to 65 experiments carried out by Neumann, and he discusses the systematic and accidental divergences (residuals). In **homals** such a linear regression problem can be emulated by setting numerical levels. Constraining the levels to be ordinal, we get a monotone regression (Gifi 1990).

```
> data(neumann)
> res.lin <- homals(neumann, sets = list(3, 1:2), level = "numerical",
+   rank = 1)
> res.mon <- homals(neumann, sets = list(3, 1:2), level = "ordinal",
+   rank = 1)
> plot(res.lin, plot.type = "loadplot", main = "Loadings Plot Linear Regression")
> plot(res.mon, plot.type = "loadplot", main = "Loadings Plot Monotone Regression")
```

The points in the loadings plot in Figure 7 correspond to regression coefficients.

#### 4.3. NLPCA on Roskam data

Roskam (1968) collected preference data where 39 psychologists ranked all nine areas (see Table 1) of the Psychology Department at the University of Nijmegen.

Using this data set we will perform two-dimensional NLPCA by restricting the rank onto 1. Note that the objects are the areas and the variables are the psychologists. Thus, the input data structure is a  $9 \times 39$  data frame. Note that the scale level is set to “ordinal”.

```
> data(roskam)
> res <- homals(roskam, rank = 1, level = "ordinal")
> plot(res, plot.type = "objplot")
> plot(res, plot.type = "labplot", var.subset = 2, main = "Labelplot Rater 2")
```

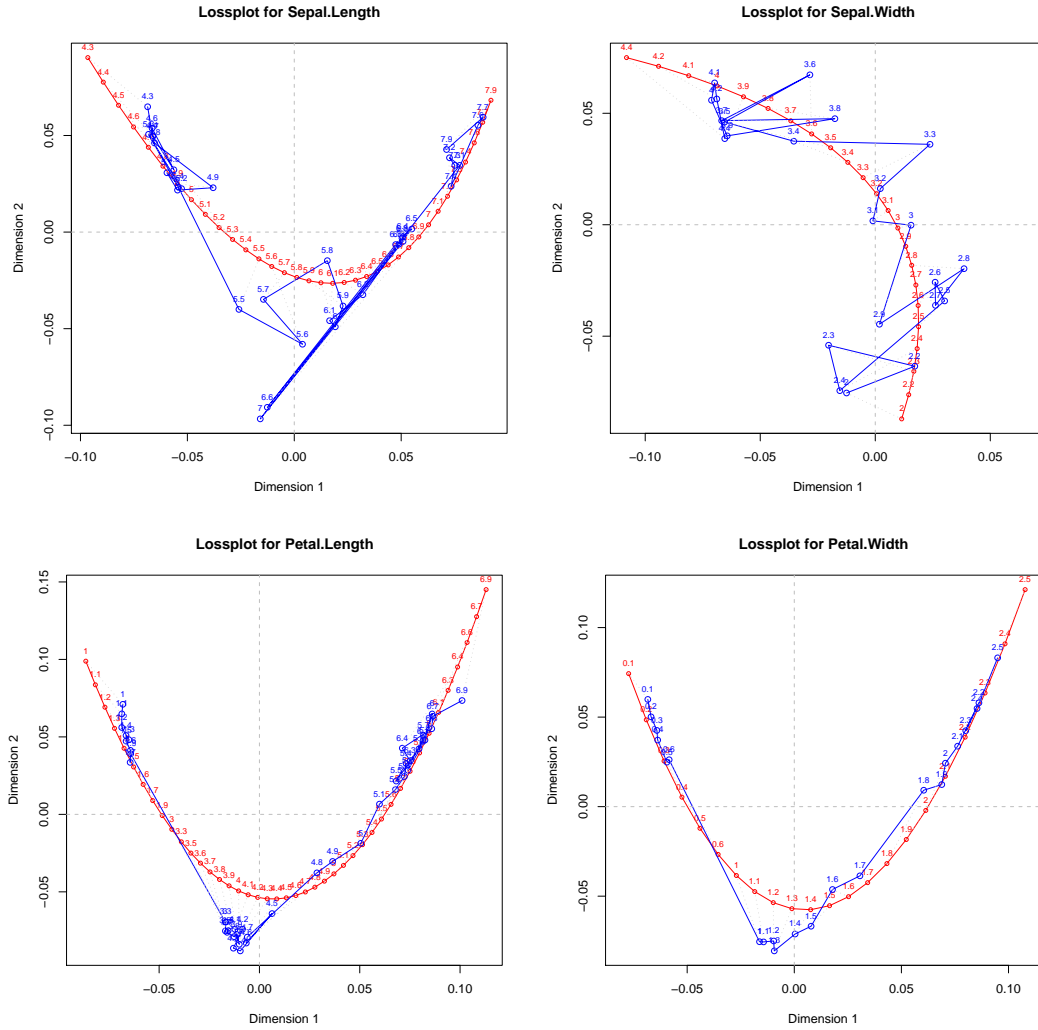


Figure 6: Loss plots for Iris Predictors

The object plot in Figure 8 shows interesting rating “twins” of departmental areas: mathematical and experimental psychology, industrial psychology and test construction (which are close to the former two areas), educational and social psychology, clinical and cultural psychology. Physiological and animal psychology are somewhat separated from the other areas. The label plot allows to look closer at a particular rater; we pick out rather #2. Obviously this rater is attracted to areas like social, cultural and clinical psychology rather than to methodological fields. Further analyses of this dataset within a PCA context can be found in [de Leeuw \(2006\)](#).



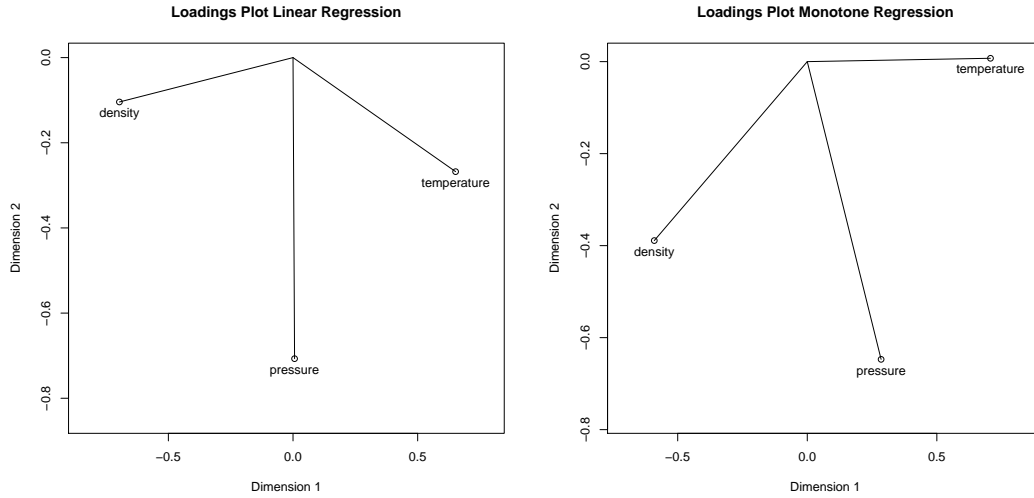


Figure 7: Loading Plots for Neumann Regression

SOC	Social Psychology
EDU	Educational and Developmental Psychology
CLI	Clinical Psychology
MAT	Mathematical Psychology and Psychological Statistics
EXP	Experimental Psychology
CUL	Cultural Psychology and Psychology of Religion
IND	Industrial Psychology
TST	Test Construction and Validation
PHY	Physiological and Animal Psychology

Table 1: Psychology Areas in Roskam Data.

## 5. Discussion

In this paper theoretical foundations of the methodology used in **homals** are elaborated as well as package application and visualization issues are presented. Basically, **homals** covers the models 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. To conclude, **homals** provides flexible easy-to-use routines which allow researchers from different areas to compute, interpret, and visualize models belonging to the Gifi-family.

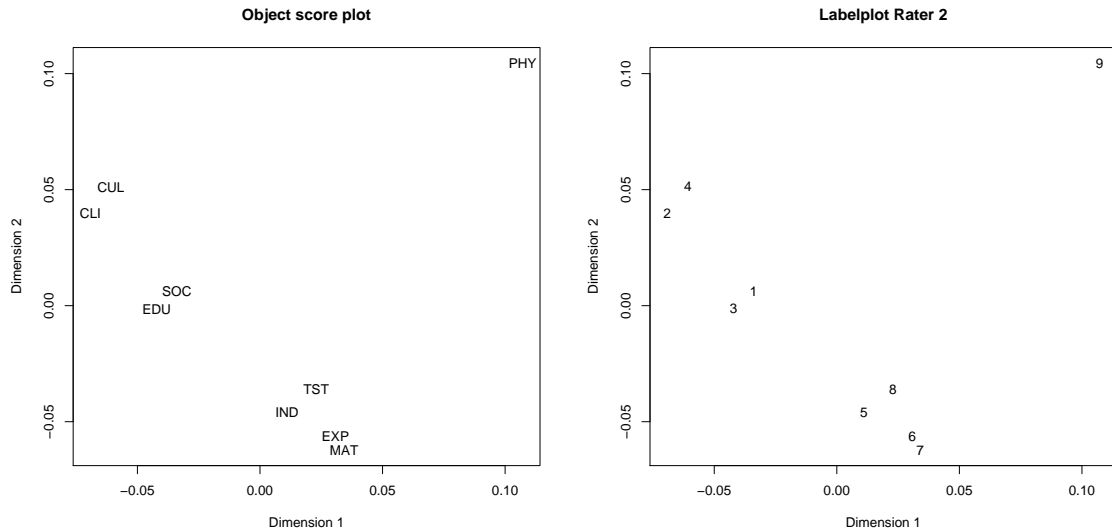


Figure 8: Plots for Roskam data

## References

- ADA (2002). "Voting Record: Shattered Promise of Liberal Progress." *ADA Today*, **57**(1), 1–17.
- 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, Michailides 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.
- 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.
- Hartigan JA (1975). *Clustering Algorithms*. Wiley, New York.
- Mair P, Hatzinger R (2007). "Psychometrics Task View." *R-News*, **7/2**, in press.
- Michailidis G, de Leeuw J (1998). "The Gifi System of Descriptive Multivariate Analysis." *Statistical Science*, **13**, 307–336.

- Michailidis G, de Leeuw J (2001). “Data Visualization through Graph Drawing.” *Computational Statistics*, **16**, 435–450.
- 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.
- Peschar JL (1975). *School, Milieu, Beroep*. Tjeek Willink, Groningen, The Netherlands.
- 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>.
- Roskam E (1968). *Metric Analysis of Ordinal Data in Psychology (PhD Thesis)*. University of Leiden, Leiden, The Netherlands.
- 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 Factors.” *Psychometrika*, **53**, 177–197.
- Wilson EB (1926). “Empiricism and Rationalism.” *Science*, **64**, 47–57.