Rank and Set Restrictions for Homogeneity Analysis in R: The “homals” Package

Patrick Mair[1], Jan de Leeuw[2]
[1]Wirtschaftsuniversität Wien, Augasse 2-6, 1090 Vienna, Austria
[2]University of California, Los Angeles, 8142 Math Sciences Bldg., Los Angeles, CA 90095-1554

Abstract
The model family proposed by Gifi (1990) is a flexible framework for the analysis of multivariate data. The common properties shared by all Gifi-models are the specification of a loss function solved by alternating least squares and transformations of the variables which lead to quantifications of the categories. The latter issue implies the concept of "optimal scaling" and allows to account for the scaling level of the variables. Starting from the basic model of homogeneity analysis, we present various extensions in terms of rank restrictions (nonlinear principal component analysis) and restrictions on sets of variables (nonlinear canonical correlation analysis). We focus on recent methodological developments and on the R package "homals" that allows for the computation of these models and provides new visualization techniques of the results.

Key Words: homals, Gifi-models, optimal scaling, homogeneity analysis, nonlinear PCA, nonlinear CCA

1. Introduction

Scaling categorical variables is a problem that has been studied extensively within the area of psychometrics. As a common problem in social sciences, researchers have often the situation of 5-point Likert items with response categories varying from “I completely disagree” (1) to “I completely agree” (5). Researchers have to keep in mind that the values of such a “1-to-5”scale are completely ad-hoc and in some sense arbitrary. This is ok as long as this type of data is regarded as ordinal. But in many applications researchers consider these data as numerical (or interval scaled) and they compute sum scores, means, etc.

In some cases, even an ad-hoc ordinal assumption can be misleading. As an example, let us consider the Galo dataset (Peschar, 1975) that, within the context of optimal scaling, is analyzed in Gifi (1990) and de Leeuw and Mair (2007). This dataset includes the variable “father’s profession” with categories *lower white collar*, *middle white collar*, *professional/managers*, *shopkeepers*, *schooled labor*, and *unskilled labor*. In this case, an order is not ad-hoc determinable. Therefore we need statistical methodology to scale the categories.

This brings us to the idea of *optimal scaling* (see Takane, 2005, for an overview). Basically, optimal scaling is a procedure which transforms the observed response categories according to some specified criterion. In other words, the distances between the categories are stretched and squeezed until we reach a particular optimum.

As a first step each variable involved in the analysis is considered as categorical (i.e. nominal). Now, if we have a priori knowledge about the order of the categories, we can take into account this property by posing order restrictions on the original scale: For instance having a Likert scale we could say that the scaled value of category (1) should be lower than the one for category (2). In turn, the scaled value for category (2) should be lower than the one for category (3) and so on. As a consequence, the original order of the variables is preserved (monotone transformations). Having a numerical variable involves the additional restriction that the distances between the original categories should be preserved. For instance, in the case of a Likert scale this would imply that the distances between the transformed categories are constant (affine transformations). This view on scaling properties of variables allows us to model variables with different scale levels.

In order to achieve optimal scaling we have to define a target criterion: The categories are scaled such that they are optimal with respect to this particular criterion. Gifi-models are not only models of optimal scaling but they also reduce the dimensionality of the problem in the sense of a principal component analysis (PCA), or, more precise, of multiple correspondence analysis (CA). The problem is formulated by means of a loss function and it is solved by the *alternating least squares* algorithm (ALS). Eventually, we get (optimally scaled) category and object scores on each dimension. As we will see, the resulting scores can be represented graphically in various ways.

In this paper we extend the basic HOMALS formulation (Section 2.1) in terms of rank restrictions (Section 2.2) and set restrictions (Section 2.3) on the category scores (or quantifications). In each section we focus on an additional "special task" (e.g. missing data, active/inactive variables, scale levels etc.) offered by the homals package (de Leeuw & Mair, 2007) in R (R development core team, 2008). All technical details can be found in Gifi (1990), Michailidis and de Leeuw (1998), as well as in de Leeuw and Mair (2007). An example using the R package homals is computed for each method.

2. Gifi Models for Nonlinear Multivariate Analysis

2.1 Homogeneity Analysis

From a conceptual point of view homogeneity analysis (HOMALS) is a synonym for multiple correspondence analysis (CA; see e.g. Greenacre & Blasius, 2006). CA is typically solved by singular value decomposition (SVD) whereas in HOMALS we use ALS on a least squares loss function which represents a criterion of departure from homogeneity to be minimized. We start with the following basic definitions. For $i = 1, \ldots, n$ objects, data on m (categorical) variables are collected. Each of the corresponding $j = 1, \ldots, m$ variables takes on k_j different values (their levels or categories). They are coded using $n \times k_j$ binary indicator or dummy matrices G_j. The whole set of indicator matrices can be collected in a block matrix $G = [G_1, \cdots, G_m]$.

The first special task we consider is the incorporation of missing values. Missing observations are coded as complete zero row sums: if object *i* is missing on variable *j*, then the row sum of G_j is 0. Otherwise the row sum becomes 1 because the category entries are disjoint. At this point all row sums of G_j are collected in the diagonal matrix M_j. Let us denote the average of the M_j matrices by $M_\bullet$. Note that G_j and M_j are matrices based on the observed data.

As mentioned above, HOMALS computes object and category scores on p dimensions. In other words, we have to solve a projection problem $\mathbb{R}^m \to \mathbb{R}^p$ with $p \ll m$. X is the $n \times p$ matrix for the object scores and Y_j the $k_j \times p$ matrix containing the category quantifications. Both matrices have to be determined during optimization. Based on these definitions, the following loss function can be established:

$$\sigma(X; Y_1, \ldots, Y_m) = \frac{1}{m} \sum_{j=1}^{m} tr\left(X - G_j Y_j\right)' M_j \left(X - G_j Y_j\right) \qquad (1)$$

which is optimized under the normalization conditions $\mathbf{u}' M_\bullet X = 0$ and $X'MX = I$ such that the trivial solution of complete 0-scores is avoided.

The basic idea of ALS minimization for solving this problem is the following: At iteration $t = 0$ we begin with a starting solution $X^{(0)}$ for the object scores. Consequently, we can update the category scores $Y_j^{(1)}$. In a next step we update the object scores $X^{(1)}$ and normalize them. Based on these normalized object scores, we update the category in the next iteration and so forth. The algorithm stops when the loss function does not decrease significantly anymore (i.e. the loss difference between two iterations is below a specified threshold ε).

We demonstrate this methodology using the senator dataset (ADA, 2008). The votes of all 100 senators on 20 issues were selected by Americans for Democratic Action. The votes selected cover a full spectrum of domestic, foreign, economic, military, environmental and social issues. In many instances we have chosen procedural votes: amendments, motions to table, or votes on rules for debate. The senators' responses are binary. In general, Democrat candidates have many more "yes" responses than Republican candidates. A full description of the items can be found in the help file of the homals package. Note that this dataset contains several non-responses.

The first column of the data set represents the party affiliation (50 Republicans, 49 Democrats and 1 Independent). This variable we will consider as "inactive". This means that this variable is not part of the optimization process. The category and object scores are based on the 20 (active) items. After optimization, the affiliation is scored by means of a (cone restricted) SVD without influencing the remaining scores computed by ALS.

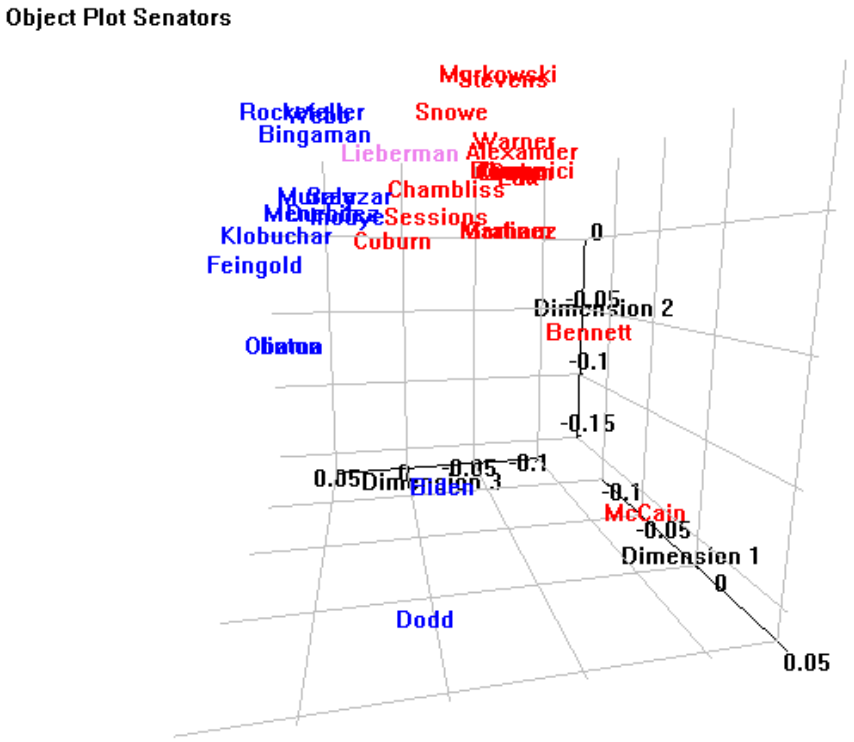

Figure 1: Object scores for homals solution on senate dataset.

We compute a 3-dimensional solution and the plot of the object scores for a subset of senators is given in Figure 1. The party affiliation is colored accordingly (blue = Democrats, red = Republicans, violet = Independent). Obviously, dimension 3 separates between the parties. This can be seen in the span plots in Figure 2 as well, where we slice through the 3-D cube and represent the scores on a 2-dimensional plane. Lieberman, which is Independent, is right in the middle between the two parties. The second dimension involves liberalism. For instance, on the Republican side, Murkowski, Stevens, and Snowe are considerably liberal. The same applies to Rockefeller, Webb, and Bingaman whereas Biden is rather conservative. It has to be pointed out that Obama, Clinton (both have the same response pattern

and therefore the same object scores), McCain, and Dodd have a noticeable amount of non-responses. Therefore their position is slightly apart from their colleagues.

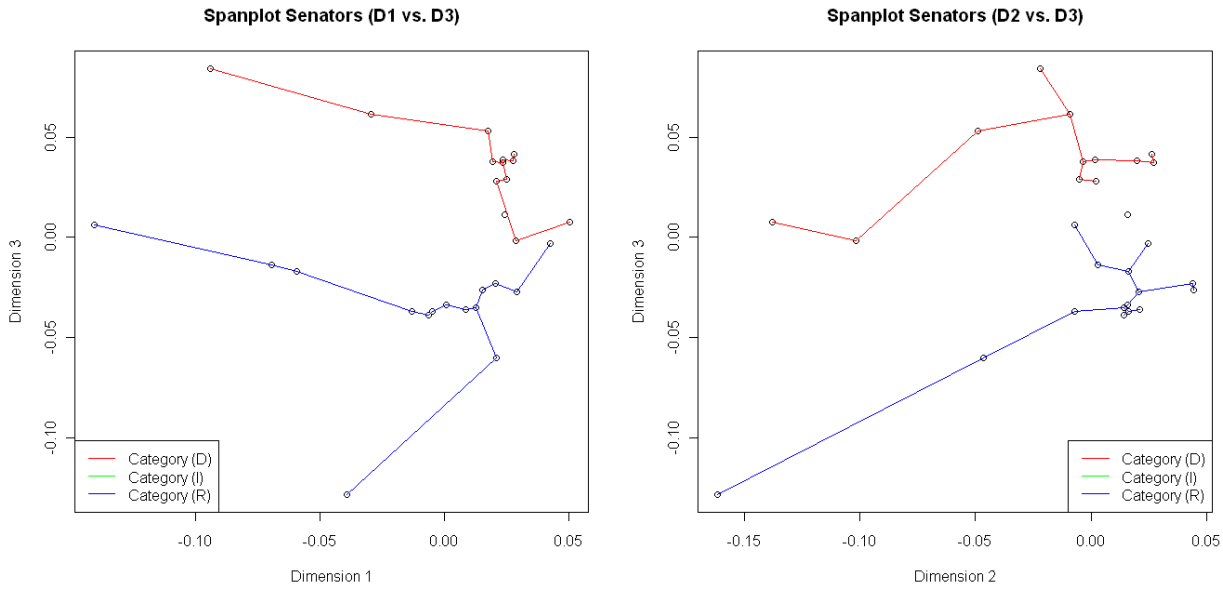

Figure 2: Span plots for object scores based on Figure 1.

The homals package offers many more options for plotting the scores such as Voronoi plots, hull plots, projection plots, start plots, loadings plots etc. Some plots will be shown in the next sections of this paper. Further descriptions and illustrations can be found in de Leeuw & Mair (2007).

2.2 Rank Restrictions: Nonlinear PCA

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

Having an $n \times m$ data matrix with metric variables, PCA is a common technique to reduce the dimensionality of the data set, that is to project the variables into a subspace $\mathbb{R}^p$. 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 an $n \times p$ matrix of component scores and an $m \times p$ matrix of component loadings. In the case of nonmetric variables, nonlinear PCA (NLPCA) is appropriate (de Leeuw, 2006; Michailidis, 2005). As always in Gifi terminology, the term "nonlinear" pertains to nonlinear transformations of the observed categories.

The crucial difference to homogeneity analysis concerns the category score matrix Y_j. In classical HOMALS, as described in the section above, Y_j is unrestricted. In NLPCA, Y_j is decomposed by a linear combination

$$Y_j = Z_j A_j', \qquad (2)$$

where Z_j is the restricted quantification matrix of dimension $k_j \times r_j$ (r_j represents the lower rank). A_j is the weight matrix and of dimension $p \times r_j$.

From a practical point of view the most important special case (see Michailidis & de Leeuw, 1998) is the rank-1 restricted formulation

$$Y_j = \mathbf{z}_j \mathbf{a}_j' .$$

This restricted formulation has two major implications: First, we are able to fit more parsimonious HOMALS models (without reducing the number of dimensions) that are straightforward to interpret. We have object scores on the one hand and single (rank-restricted) category quantifications on the other hand. Second, it is straightforward to incorporate different scale levels on the variables as quoted in the Introduction. The corresponding mathematical elaborations can be found in de Leeuw and Mair (2007, p. 6).

Using the general linear decomposition in (2), the loss function given in equation (1) changes to

$$\sigma(X;Z;A) = \frac{1}{m}\sum_{j=1}^{m} tr\left(X - G_j Z_j A_j'\right)' M_j \left(X - G_j Z_j A_j'\right), \qquad (3)$$

and is again minimized by ALS.

To illustrate the difference between unrestricted HOMALS and (rank-restricted) NLPCA we use the sleeping bag data from Prediger (1997; see also Michailidis, 2005). The variables temperature, weight, price, material, and quality rating were collected on 21 sleeping bags. The first three variables are numerical, material is nominal, and the quality rating (scale from 1 to 3; the higher the better) is ordinal.

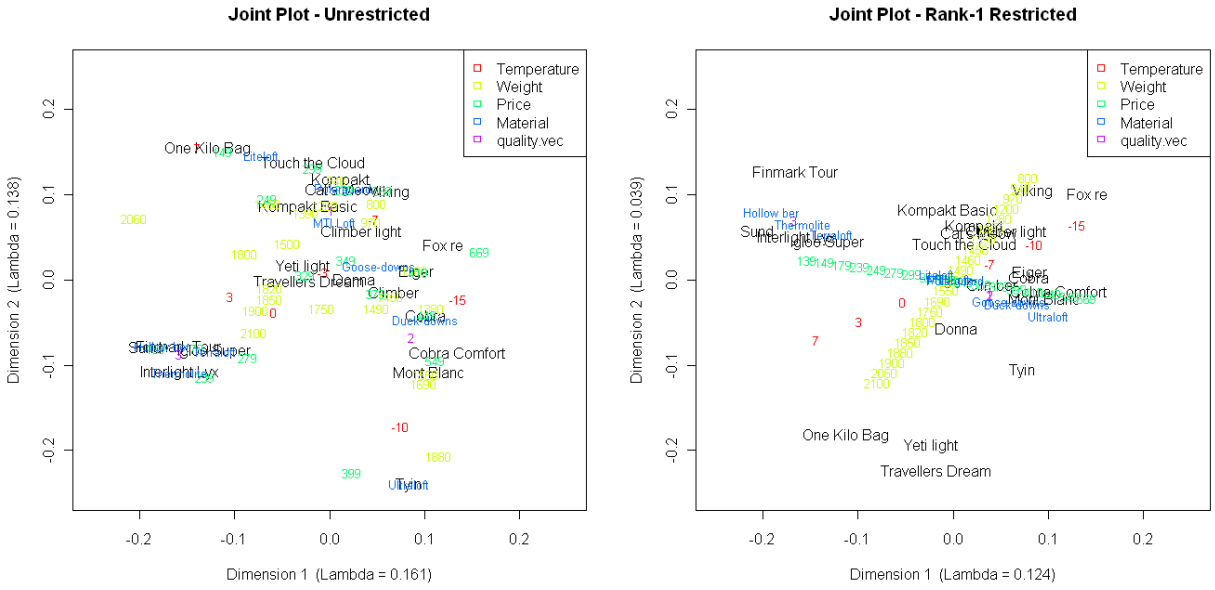

Figure 3: Unrestricted and rank-1 restricted homals solution for sleeping bags

On the left hand side of Figure 3 we see the joint plot (object and category scores) of a HOMALS solution as described in the former section. Within a CA context this representation is sometimes referred to as CA map. On the axes labels the eigenvalues are given. They can be used as a goodness-of-fit measure. In this 2-dimensional HOMALS solution their sum is .299.

On the right hand side of Figure 3 a rank-1 restricted NLPCA is computed which incorporates the scale level of the variables. Only one set of category scores is estimated. Obviously the category scores of dimension 2 are just linear combinations of the scores of dimension 2. As a practical implication, from a Marketing perspective, this more restricted model (sum of eigenvalues is .163) eases the positioning of products. For instance, suppliers of sleeping bags can immediately see the effects and the market position in the case of a price change. They might fall into a different segment with different competitors. Or, they could provide a material of higher quality and again examine the new market position. Of course, this is also possible with the basic HOMALS but in such restricted NLPCA representations

the implication of such a change is easier to explain since the bags are basically shifted along the linear alignments determined by the rank-restricted category scores.

2.3 Set Restrictions: Nonlinear CCA

As a further extension (see van der Burg, de Leeuw, & Verdegaal, R., 1988; van der Burg, de Leeuw, & Dijksterhuis, 1994) we can define sets of variables. This follows the tradition of CCA. That said, we point out that nonlinear CCA (NLCCA) is not restricted to two sets only but rather to $v = 1, \ldots, K$ sets. Note that in basic HOMALS each variable forms a single set (i.e., $K = m$).

The aim of homogeneity analysis was to find *p* orthogonal vectors in *m* indicator matrices G_j . This approach can be extended in terms of computing *p* orthogonal vectors in *K* general matrices G_v , each of dimension $n \times m_v$ where $j = 1, \ldots, m_v$ is the number of variables in set *v*. Each G_v can be represented as a block matrix consisting of the indicator matrices G_{v_j} of the variables that belong to the particular set *v*. Hence, the loss function becomes

$$\sigma\left(X; Y_{v_1}, \ldots, Y_{K_{m_v}}\right) = \frac{1}{K}\sum_{v=1}^{K} tr\left(X - \sum_{j=1}^{m_v} G_{v_j} Y_{v_j}\right)' M_v \left(X - \sum_{j=1}^{m_v} G_{v_j} Y_{v_j}\right) \qquad (4)$$

where *X* is the $n \times p$ matrix of object scores, G_{v_j} is $n \times k_j$, and Y_{v_j} is the $k_j \times p$ matrix of category scores. As before, missing values are taken into account in M_v . Again, this loss function is minimized by ALS.

At this point we can think of a combination of NLPCA and NLCCA: Rank restrictions within sets of variables. The loss function given in equation (4) changes straightforwardly according to equation (3). In the homals package all these models can be computed by means of the same function; called `homals()`, naturally. Rank restrictions can be imposed by the `rank` argument and set definitions by the `sets` argument. The `level` argument allows the incorporation of different scale levels.

Consequently, we present an example that includes all these options and extensions using the Neumann dataset (Wilson, 1926). Willard Gibbs discovered a theoretical formula connecting 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 discussed the systematic and accidental divergences (residuals).

Thus we have three numerical variables where we impose rank-1 restrictions and incorporate the scale level accordingly. In addition we define two sets: The first one consists of temperature and pressure, the second one of density only. This set definition allows us to emulate linear regression (without being able to specify density as response). Note that the numerical scale level involves linear transformations of the scores. To visualize the implication of different scale definitions, we compute the same model by treating the variables as ordinal, and, subsequently, as nominal.

The results are visualized in the transformation plots in Figure 4. By regarding the variables as numerical we see clearly the linear (affine) transformation in terms of $ax + b$. Assuming an ordinal scale we have a monotone transformation of the scores (red line in Figure 4). Finally, in the nominal case, the scores are transformed in a clearly nonlinear (non-monotone) manner.

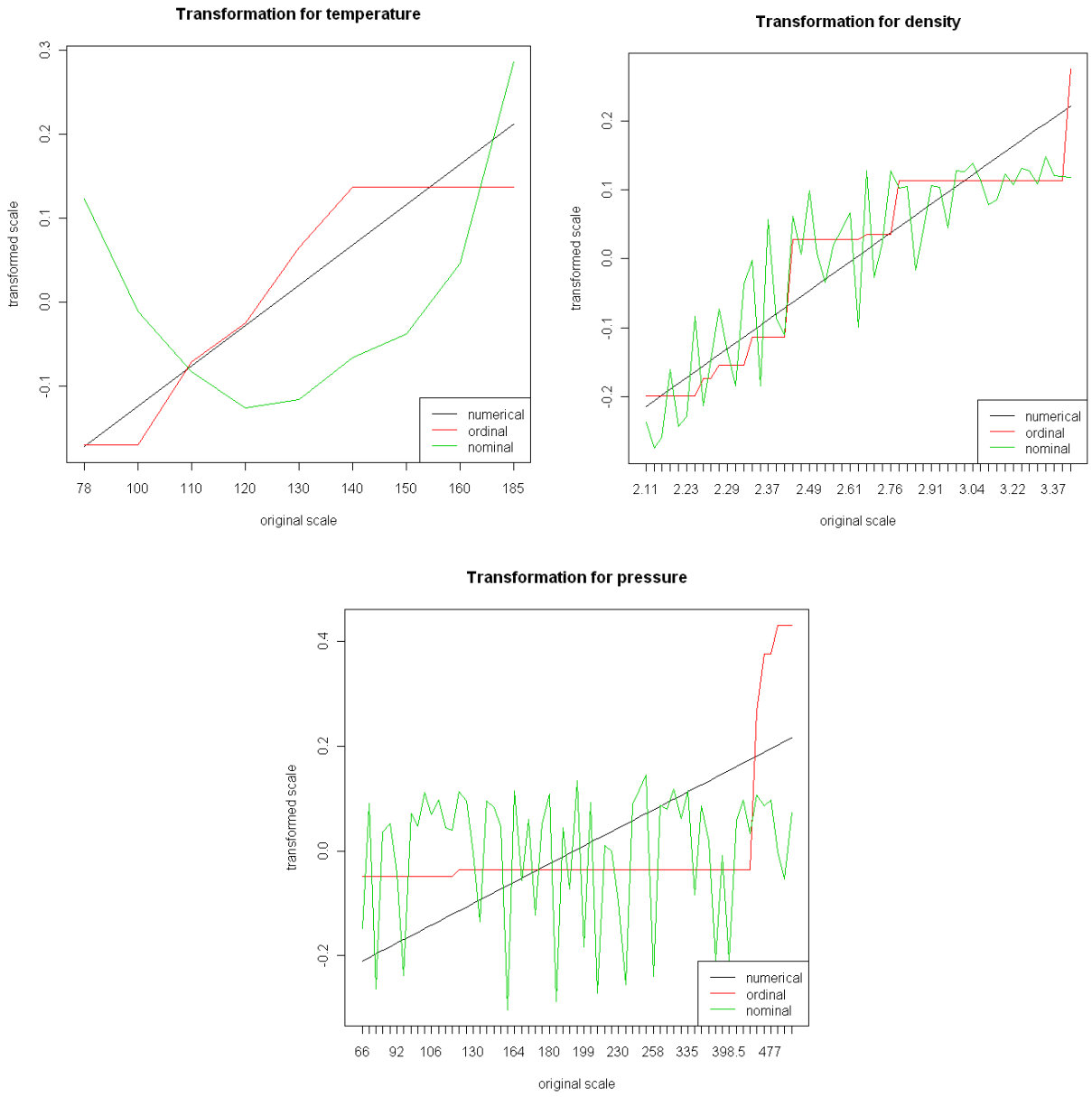

Figure 4: Transformation plots on Neumann data for different scale levels.

3. Conclusion and Further Developments

Gifi models and the corresponding R implementation by means of the package “homals” offer a powerful framework for scaling and visualizing multivariate categorical data. Embedding these models into a flexible environment like R allows the user to further process the results such as, for instance, using the HOMALS scores for subsequent statistical models. We have shown extensions in terms of rank restrictions on the category quantifications and set definitions based on CCA. In addition, based on the idea that each variable can basically be considered as categorical, different scale levels can be incorporated by means of level restrictions (monotone, linear) on the scores.

Having numerical variables with a considerable amount of different values, this approach is not efficient and can be quite time-consuming. This problem can be solved be the incorporation of B-splines (see various chapters in van Rijckevorsel & de Leeuw, 1988). This idea adds a vast amount of flexibility in terms of score transformations and we are not limited to the transformations shown in Figure 4 anymore. This will be one future development of the homals package. Furthermore, we are working on the "aspect" framework which adds a different perspective to the Gifi methodology (see Mair & de Leeuw, 2008) and offers many additional possibilities of analyzing multivariate data.

References

ADA (2008). Americans for Democratic Action: A newsletter for liberal activists, 63(1). URL: www.adaction.org.

de Leeuw, J. (2006). Nonlinear principal component analysis and related techniques. In M. Greenacre and J. Blasius (Eds.), *Multiple Correspondence Analysis and Related Methods* (107-134). Boca Raton, FL: Chapman & Hall/CRC.

de Leeuw, J., & Mair, P. (2007). Homogeneity analysis in R: The package homals. *UCLA Statistics Preprint Series, 525*. URL: http://preprints.stat.ucla.edu/

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

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

Mair, P., & de Leeuw, J. (2008). A general framework for multivariate analysis with optimal scaling: The R package aspect. *UCLA Statistics Preprint Series*. URL: http://preprints.stat.ucla.edu/

Michailidis, G. (2005). Principal components and extensions. In B. Everitt and D. C. Howell (Eds.), *Encyclopedia of Statistics for Behavioral Sciences*. Chichester: Wiley.

Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. Statistical Science, 13, 307–336.

Peschar J. L. (1975). School, Milieu, Beroep. Tjeek Willink, Groningen, The Netherlands.

Prediger, S. (1997): Symbolic objects in formal concept analysis. In G. Mineau and A. Fall (Eds.), *Proccedings of the Second International Symposium on Knowledge, Retrieval, Use and Storage for Efficiency*, Vancouver.

R Development Core Team (2008). 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, and D. Howell (Eds.), *Encyclopedia of Statistics for Behavioral Sciences* (pp. 1479-1482). Chichester: Wiley.

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.

van Rijckevorsel, J. L. A., & de Leeuw, J. (1988). Component and correspondence analysis: Dimension reduction by functional approximation. New York: Wiley.

Wilson, E. B. (1926). Empiricism and rationalism. *Science*, 64, 47–57.