Open Research Online

The Open University's repository of research publications
and other research outputs

Generalised procrustes analysis with optimal scaling: exploring data from a power supplier

Journal Item

For guidance on citations see FAQs.

Version: Accepted Manuscript

oro.open.ac.uk

Generalised Procrustes Analysis with Optimal Scaling: exploring data from a power supplier.

Jaap Wieringa[1]

Garmt Dijksterhuis[1, 3]*

John Gower[2]

Frederieke van Perlo[1]

[1] Department of Marketing, Faculty of Economics and Business, University of Groningen, NL.
[2] Department of Mathematics and Statistics, The Open University, Milton Keynes, UK.
[3] Unilever Food & Health Research Institute, Vlaardingen, the Netherlands

* Corresponding author.
Unilever Food & Health Research Institute
G.B. Dijksterhuis
PO Box 114
3130 AC Vlaardingen
The Netherlands
t: +31(0)10-460 6275
f: +31(0)10-460 5236
e: garmt.dijksterhuis@unilever.com

Abstract

Generalised Procrustes Analysis (GPA) is a method for matching several, possibly large, data sets by fitting to each other using transformations, typically rotations. The linear version of GPA has been applied in a wide range of contexts. A non-linear extension of GPA is developed which uses Optimal Scaling (OS). The approach is suited to match data sets that contain nominal variables. A data base of a Dutch power supplier that contains many categorical variables unfit for the usual linear GPA methodology is used to illustrate the approach.

Key words: Generalised Procrustes Analysis; Optimal Scaling; Multivariate Data Analysis; switching behaviour.

1   Introduction

Generalised Procrustes Analysis (GPA, Gower 1975) is a versatile set of methods for matching several data sets. Originally it was designed to match configurations. However, it can also match more general data sets, including the case that we consider in this paper where different variables are observed in each data set (see Gower and Dijksterhuis 2004). Recently GPA has been generalised to include PLS regression (Dijksterhuis, Martens and Martens 2005) and combined with canonical variate analysis by Gardner, Gower and LeRoux (2006).

The applicability of GPA is limited to those cases where all variables in all data sets are measured on a ratio scale or on an interval scale. In the present paper we develop an extension that can be used to match data sets containing nominal variables. We propose to code the category levels in indicators and to match the data sets using an optimal scaling version of Generalised Procrustes Analysis that we develop in this paper. Our method conceptually fits into the Gifi (1990) system of nonlinear multivariate data analysis, as we employ an Alternating Least Squares (ALS) algorithm to compute both optimal scores for the categories and an optimal Procrustes fit of the data sets.

The GPA loss function is $\sum_{h<k}^{K}\left\|\mathbf{X}_h\mathbf{Q}_h-\mathbf{X}_k\mathbf{Q}_k\right\|$, where $\mathbf{X}_1,\cdots,\mathbf{X}_K$ are data sets consisting of quantitative variables, and by $\left\|\mathbf{A}\right\|$ we mean $\operatorname{tr}(\mathbf{A}'\mathbf{A})$. The columns of each $\mathbf{X}_k$ refer to different variables and the rows correspond to (the same) observational units (e.g. respondents). Columns of zeroes are padded such that all data sets have the same number of columns (this is further explained in Section 2). As in all applications of GPA, we seek the rotation matrices $\mathbf{Q}_1,\cdots,\mathbf{Q}_K$ that minimize this loss function. To accommodate categorical variables, we rewrite the loss function as $\sum_{h<k}^{K}\left\|\mathbf{G}_h\mathbf{Z}_h\mathbf{Q}_h-\mathbf{G}_k\mathbf{Z}_k\mathbf{Q}_k\right\|$, where $\mathbf{G}_k$ is an indicator matrix for the kth set of categorical variables and $\mathbf{Z}_k$ is a matrix containing the category

quantifications (cf. Gifi 1990) that need to be determined. Our aim is to assign optimal scores to the category levels, in the sense that the scores minimise the GPA least-squares criterion. The underlying idea of our approach is that the Procrustean fit may be improved by assigning appropriate quantifications to the levels of the nominal variables. We name this optimal scaling version of GPA, not surprisingly, OS-GPA.

OS-GPA is related to OVERALS (van der Burg 1988, Gifi 1990), but differs on one important issue. In GPA the transformation matrices, $\mathbf{Q}_k$, are orthogonal rotation matrices, whereas the transformation matrices are unconstrained in OVERALS. The difference in treatment of the row-objects in the data sets is that the distance between the row-points is invariant in GPA, but is allowed to vary under the transformations in OVERALS. This restriction is particularly useful in applications where the rows correspond to observational units, which is often the case in applications of GPA where the row-points are observational units whose inter-relations, here distances, should remain fixed (cf. Dijksterhuis and Gower 1991, Steenkamp, Van Trijp and Ten Berge 1994). In our application, the row-objects are customers and the distances between them indicate similarities based on the estimated scores on the variables in the data sets. In the marketing setting of the application, differences between customers (the row-points) enable the interpretation of effects of the different variables. The differences between customers also enable the identification of meaningful segments of customers. For such reasons the distances between row-points are to remain fixed in the analysis.

As a simple alternative to our approach, one might consider a GPA of the indicator matrices $\mathbf{G}_k$ themselves. We note that a Procrustean rotation of an indicator matrix does not necessary result in an indicator matrix and one might have reservations about centring and/or scaling each $\mathbf{G}_k$. So at first sight, applying GPA directly to $\mathbf{G}_k$ is not attractive. However, the squared distance between a pair of rows of an indicator matrix is twice the number of

mismatches among the categorical variables. This represents the complement of the extended matching coefficient (EMF) which counts the number of matches, a direct extension of the simple matching coefficient for binary variables (see Gower and Hand, 1996). The EMF is a simple but effective measure of similarity, so the configurations defined by the $\mathbf{G}_k$ are not unacceptable. Indeed, we may do multidimensional scaling analyses (MDS, cf. Borg and Groenen, 1997) of the EMF matrices, replacing them by coordinates that generate the EMFs. A GPA of these matrices of derived coordinates would be equivalent to a GPA of the indicator matrices themselves. That the MDS coordinates are not defined up to an arbitrary rotation is immaterial in the context of GPA. The derived coordinates may be regarded as a form of quantification, though different occurrences of the same category level will be given multidimensional "scores". We compare the results obtained by this simple approach with those obtained by assigning optimal scores.

We apply OS-GPA to a data base of a Dutch power supplier. The data base contains many categorical variables originating from different business units within the company. This induces natural splits in the data base. The underlying research problem for the power supplier is to characterize their customers, e.g. for identifying potential drivers of switching behaviour. The latter is a managerially relevant issue as the power market has recently been liberalised in the Netherlands.

2 Theory

Let $\mathbf{X}_k$, $k = 1, \cdots, K$, denote K data matrices of size ($n \times p_k$), in which the corresponding rows refer to identical objects. For $\mathbf{X}_k$ with $p_k < \max_k p_k$ columns with zeroes are padded so that $p_1 = p_2 = \cdots = p_K = \max_k p_k \equiv p$. Padding with zeroes neither increases the dimensionality of the data, nor does it affect inter-row distances (see Gower and Dijksterhuis 2004, p.34). We assume that all data are nominal, *i.e.* consist of scores on a finite number of categories. The *j*th

categorical variable in set k can be coded into an $n \times c_{kj}$ indicator matrix $\mathbf{G}_{kj}$, where c_{kj} equals the number of categories of variable j in set k. Let $\mathbf{z}_{kj}$ denote the $c_{kj} \times 1$ vector of quantifications (cf. Gifi 1990, Section 2.2) for the jth variable in set k. We assume that the k sets contain different variables. In this case mean values cannot be compared because the variables differ, so orientational information is all that is available.

Using this notation the data matrix $\mathbf{X}_k$ can be written as

$$\mathbf{X}_k = [\mathbf{G}_{k1}\mathbf{z}_{k1}, \mathbf{G}_{k2}\mathbf{z}_{k2}, \cdots, \mathbf{G}_{kp}\mathbf{z}_{kp}] \tag{1}$$

where, as explained above, the indicator matrices are zero for variables that do not occur. As an illustration, consider the following data matrix for a data set k with two variables, where the first has three categories and the second has two:

$$\mathbf{X}_k = \begin{bmatrix} 3 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}.$$

The corresponding indicator matrices $\mathbf{G}_{k1}$ and $\mathbf{G}_{k2}$ and quantification vectors $\mathbf{z}_{k1}$ and $\mathbf{z}_{k2}$ are:

$$\mathbf{G}_{k1} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \ \mathbf{G}_{k2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \mathbf{z}_{k1} = \begin{bmatrix} z_{11} \\ z_{12} \\ z_{13} \end{bmatrix}, \text{ and } \mathbf{z}_{k2} = \begin{bmatrix} z_{21} \\ z_{22} \end{bmatrix}.$$

The Procrustes problem is to minimise

$$\sum_{h<k}^{K} \left\| \mathbf{X}_h\mathbf{Q}_h - \mathbf{X}_k\mathbf{Q}_k \right\|, \tag{2}$$

where $\mathbf{Q}_1, \cdots, \mathbf{Q}_K$ are $p \times p$ orthogonal matrices. The basic idea of GPA with optimal scaling is to alternate between steps (i) and (ii) in the following:

(i) For given $\mathbf{z}_{kj}$, obtain $\mathbf{Q}_1, \cdots, \mathbf{Q}_K$ by conventional GPA;

(ii) For given $\mathbf{Q}_k$ determine optimal scores $\mathbf{z}_{kj}$ ($k = 1, \cdots, K$; $j = 1, \cdots, p_k$).

Step (i) is extensively described by Gower and Dijksterhuis (2004). The interesting step is step (ii), where optimal quantifications $\mathbf{z}_{kj}$ ($k=1,\cdots,K$; $j=1,\cdots,p_k$) are determined for given rotation matrices $\mathbf{Q}_k$ ($k=1,\cdots,K$).

Note that in step (ii) the maximum value of j is p_k, implying that padding is irrelevant when estimating the optimal scores. In step (i) any necessary columns of each current $\mathbf{X}_k$ are padded with zeros when updating the estimated orthogonal matrices. Thus, when estimating $\mathbf{z}_{kj}$ summations over j from $1,\cdots,p$ are understood to terminate at p_k.

To proceed with the analysis, we take into account the usual requirement of GPA that all variables are centred to zero mean (thus dealing with translation) *i.e.*

$$\mathbf{1}'\mathbf{G}_{kj}\mathbf{z}_{kj} = 0 \text{ (for all } k, j), \text{ which implies } \mathbf{1}'\mathbf{L}_{kj}\mathbf{z}_{kj} = 0 \text{ (for all } k, j), \tag{3}$$

where $\mathbf{1}$ is a vector of ones and $\mathbf{L}_{kj} = \mathbf{G}'_{kj}\mathbf{G}_{kj}$. In GPA we often have a size constraint, such as

$$\operatorname{tr}\sum_{k=1}\mathbf{X}'_k\mathbf{X}_k = K.$$ Such a constraint is essential when $\mathbf{X}_k$ is defined in terms of quantifications, as in (1), to rule out the trivial solution $\mathbf{z}_{kj} = 0$ (for all k, j). It turns out that pathological solutions are avoided only if the $\mathbf{z}_{kj}$ satisfy the constraint $\mathbf{z}'_{kj}\mathbf{L}_{kj}\mathbf{z}_{kj} = 1$ for all k and $j \leq p_k$. Thus, $\operatorname{tr}\left(\mathbf{X}'_k\mathbf{X}_k\right) = p_k$ (see the end of this section for remarks on p_k scaling).

In Van Buuren and Dijksterhuis (1988) the same constraint is used but is not included in the optimisation. They calculate unconstrained solutions and impose the constraint in the algorithm by standardizing the quantifications $\mathbf{z}_{kj}$ obtained in each iteration. The optimization in our approach explicitly takes centring and standardising of the matrices $\mathbf{G}_{kj}\mathbf{z}_{kj}$ into account which van Buuren and Dijksterhuis (1988) also perform but in a separate step, *i.e.* not included in the objective function.

We now develop our OS-GPA approach. The Procrustes problem is

$$\text{minimise} \sum_{h<k}^{K} \left\| \mathbf{X}_h \mathbf{Q}_h - \mathbf{X}_k \mathbf{Q}_k \right\|,$$

subject to

$$\mathbf{1}' \mathbf{L}_{kj} \mathbf{z}_{kj} = 0, \text{ for all } k, j \text{ and } \mathbf{z}'_{kj} \mathbf{L}_{kj} \mathbf{z}_{kj} = 1, \text{ for all } k \text{ and } j \text{ with } j \leq p_k.$$

A Lagrangian for this problem is

$$L = \sum_{h<k}^{K} \left\| \mathbf{X}_h \mathbf{Q}_h - \mathbf{X}_k \mathbf{Q}_k \right\| - 2 \sum_{k=1}^{K} \sum_{j=1}^{p} \mu_{kj} \mathbf{1}' \mathbf{L}_{kj} \mathbf{z}_{kj} - \sum_{k=1}^{K} \sum_{j=1}^{p} \lambda_{kj} \left(\mathbf{z}'_{kj} \mathbf{L}_{kj} \mathbf{z}_{kj} - 1 \right). \tag{4}$$

From (4) we isolate the terms involving $\mathbf{z}_{kj}$. The terms involving $\mathbf{z}_{kj}$ in $\sum_{h<k}^{K} \left\| \mathbf{X}_h \mathbf{Q}_h - \mathbf{X}_k \mathbf{Q}_k \right\|$

occur in $\sum_{h \neq k}^{K} \left\| \mathbf{X}_h \mathbf{Q}_h - \mathbf{X}_k \mathbf{Q}_k \right\|$ and are:

$$(K-1)\operatorname{tr}\left(\mathbf{X}'_k\mathbf{X}_k\right)-2\operatorname{tr}\left(\mathbf{Q}'_k\mathbf{X}'_k\sum_{h\neq k}^{K}\mathbf{X}_h\mathbf{Q}_h\right)=(K-1)\left[\operatorname{tr}\left(\mathbf{X}'_k\mathbf{X}_k\right)-2\operatorname{tr}\left(\mathbf{Q}'_k\mathbf{X}'_k\overline{\mathbf{Y}}_{-k}\right)\right],\quad(5)$$

where $\overline{\mathbf{Y}}_{-k}$ is the k-excluded group average $\frac{1}{K-1}\sum_{h\neq k}^{K}\mathbf{X}_h\mathbf{Q}_h$. Observe that

$$\operatorname{tr}\left(\mathbf{X}'_k\mathbf{X}_k\right)=\sum_{j=1}^{p_k}\mathbf{z}'_{kj}\mathbf{L}_{kj}\mathbf{z}_{kj}\qquad(6)$$

and

$$\operatorname{tr}(\mathbf{Q}'_k\mathbf{X}'_k\overline{\mathbf{Y}}_{-k})=\operatorname{tr}\left(\mathbf{G}_{k1}\mathbf{z}_{k1},\mathbf{G}_{k2}\mathbf{z}_{k2},\cdots,\mathbf{G}_{kp}\mathbf{z}_{kp}\right)'\overline{\mathbf{Y}}_{-k}\mathbf{Q}'_k$$

$$=\operatorname{tr}\begin{pmatrix}\mathbf{z}'_{k1}\mathbf{G}'_{k1}\\ \vdots\\ \mathbf{z}'_{kp_k}\mathbf{G}'_{kp}\end{pmatrix}\left(\mathbf{y}_{k1},\mathbf{y}_{k2},\cdots,\mathbf{y}_{kp}\right),\qquad(7)$$

where $\mathbf{y}_{kj}$ is defined as the jth column of $\overline{\mathbf{Y}}_{-k}\mathbf{Q}'_k$. For a specific j the only term of (7)

involving $\mathbf{z}_{kj}$ is $\mathbf{z}'_{kj}\mathbf{G}'_{kj}\mathbf{y}_{kj}$. Thus, the only terms in (5) that involve $\mathbf{z}_{kj}$ follow from (6) and (7)

to give

$$\mathbf{z}'_{kj}\mathbf{L}_{kj}\mathbf{z}_{kj}-2\mathbf{z}'_{kj}\mathbf{G}'_{kj}\mathbf{y}_{kj}.\qquad(8)$$

Note that the objective of the optimisation problem reduces to a linear term subject to the

quadratic constraint $\mathbf{z}'_{kj}\mathbf{L}_{kj}\mathbf{z}_{kj}=1$. Thus, finally, the only terms in the Lagrangian (4) involving

$\mathbf{z}_{kj}$ are

$$\mathbf{z}'_{kj}\mathbf{L}_{kj}\mathbf{z}_{kj}-2\mathbf{z}'_{kj}\mathbf{G}'_{kj}\mathbf{y}_{kj}-2\mu_{kj}\left(\mathbf{1}'\mathbf{L}_{kj}\mathbf{z}_{kj}\right)-\lambda_{kj}\left(\mathbf{z}'_{kj}\mathbf{L}_{kj}\mathbf{z}_{kj}\right).\qquad(9)$$

Differentiating (4) with respect to $\mathbf{z}_{kj}$ therefore results in the first-order condition

$$\left(1-\lambda_{kj}\right)\mathbf{L}_{kj}\mathbf{z}_{kj}=\mathbf{G}'_{kj}\mathbf{y}_{kj}+\mu_{kj}\left(\mathbf{L}_{kj}\mathbf{1}\right).\qquad(10)$$

Pre-multiplying (10) by $\mathbf{1}'$ and imposing the centring constraint gives

$$\mathbf{1}'\mathbf{G}'_{kj}\mathbf{y}_{kj}+\mu_{kj}\left(\mathbf{1}'\mathbf{L}_{kj}\mathbf{1}\right)=0,\ i.e.\ \mu_{kj}=-\frac{1}{n}\mathbf{1}'\mathbf{y}_{kj}\qquad(11)$$

and if we pre-multiply (10) by $\mathbf{z}'_{kj}$, we obtain

$$\lambda_{kj} = 1 - \mathbf{z}'_{kj}\mathbf{G}'_{kj}\mathbf{y}_{kj}. \tag{12}$$

Substituting (11) and (12) into (10) gives

$$(\mathbf{z}'_{kj}\mathbf{G}'_{kj}\mathbf{y}_{kj})\mathbf{L}_{kj}\mathbf{z}_{kj} = \mathbf{G}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{y}_{kj},$$

where $\mathbf{N} = \dfrac{1}{n}\mathbf{11}'$, so that we can now solve for $\mathbf{z}_{kj}$:

$$\mathbf{z}_{kj} = [\mathbf{y}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{G}_{kj}\mathbf{L}_{kj}^{-1}\mathbf{G}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{y}_{kj}]^{-1/2}\mathbf{L}_{kj}^{-1}\mathbf{G}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{y}_{kj}. \tag{13}$$

We show in Appendix B that (13) does indeed give a minimum of the objective function, except in pathological situations where there is an exact fit, but possibly non-unique.

When the different sets have differing numbers of variables this may introduce spurious size effects. These can be circumvented by adapting the normalisation, using p_k-scaling (cf. Gower and Dijksterhuis 2004), where each variable is divided by $\sqrt{p_k}$, p_k being the number of variables in set k. This scaling results in a slightly different expression for the $\mathbf{z}_{kj}$:

$$\mathbf{z}_{kj} = \frac{1}{\sqrt{p_k}}[\mathbf{y}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{G}_{kj}\mathbf{L}_{kj}^{-1}\mathbf{G}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{y}_{kj}]^{-1/2}\mathbf{L}_{kj}^{-1}\mathbf{G}'_{kj}(\mathbf{I}-\mathbf{N})\mathbf{y}_{kj}.$$

One outcome of the OS-GPA analysis gives a group average $\mathbf{Y}$, *i.e.* a position of each customer in a multidimensional space; there is also a position of each customer, per set of variables, *i.e.* one such point in the space of each of the K sets: $\mathbf{G}_k\mathbf{Z}_k\mathbf{Q}_k$, $k=1,\cdots,K$. The $\mathbf{Z}_k$ contain the quantifications for the categories of the variables. These can all be plotted in the space of the first few, typically two, dimensions of the principal components analysis (PCA, cf. Jolliffe 2002) on $\mathbf{Y}$. A PCA on $\mathbf{Y}$ gives a low dimensional representation of the high dimensional OS-GPA result in $\mathbf{Y}$. An algorithm for our approach is listed in Appendix A.

3   Application

The data refer to 500 customers of a Dutch power supplier. The data base contains data from four different business units within the company (hence, $K = 4$). The first data set contains

three variables that quantify consumption of the power supplier's services, *viz.* variables that measure usage of electricity, gas and cable TV. The second set consists of five different cost variables, such as the number of inbound phone contacts and the paying behaviour of the customers. The third set contains eight different demographic variables of the customer. The fourth set contains 17 satisfaction variables as measured in a survey. Table 1 gives an overview of the variables in the four data sets and their measurement scales.

Table 1   The four data sets, a description of the variables, and the number of categories.

Table 1

The four data sets were matched using the OS-GPA developed in the previous section. We have analysed the variables assuming that all are nominal. To this end, we discretised variables that are measured on an ordinal, ratio or interval scale, without restricting their category quantifications. Such an analysis on a nominal level enables non-linear relationships between variables to be taken into account.

We decomposed the overall loss value (2) into 500 respondent-specific losses, which are presented in Figure 1. The losses can be seen to run smoothly from low values to higher values. A number of customers can be identified to have a relatively high loss value. The high large loss levels are caused by the fact that these customers have atypical values for some of the variables. It can be useful to determine the reasons for this atypicality as these customers may require special attention. Such customers may *e.g.* shortly decide to switch to another power supplier, which could perhaps be prevented by giving them some attention by offering them a special deal.

Figure 1

A decomposition of the total loss value in losses per data set (not shown) does not identify important differences in fit between the four sets.

The first two principal components of the GPA group average contain 30% of the variance in the GPA group average and will be used to illustrate the results of the OS-GPA method. The results can be given in a biplot (Blasius, Eilers and Gower 2009), as they are the result of a proper PCA. As there are 500 row points, and a total of 33 variables, which together have 152 categories in total, a plot with all this information easily becomes cluttered. We therefore present for each data set a biplot in Figure 2. This plot contains the position of the 500 customers, the rows from $\mathbf{G}_k\mathbf{Z}_k\mathbf{Q}_k\mathbf{V}$, labelled by the set number (1 through 4) and the positions of the quantified variable categories, the rows from $\mathbf{Z}_k\mathbf{Q}_k\mathbf{V}$, where $\mathbf{V}$ is the matrix from the SVD of $\mathbf{Y} = \mathbf{USV}'$ (see Appendix A, algorithm step 4). The latter are indicated by open dots labelled by three numbers (variable no., category no, set no.) which correspond to the numbers in Table 1. The category scores belonging to the same variable are connected by a straight line through the origin. This is not an artefact but is a natural consequence of the scores being on a linear coordinate axis, whose projection is shown. Note that not every plot contains the full number of 500 customers because many points may coincide. The number of non-coinciding points is limited by the number of profiles, that is, by the number of variables and the number of categories per variable. The first set contains three variables which have 4, 4 and 2 categories, which totals to 32 different possible profiles. The other sets, no. 2, 3 and 4, contain 144, 1944 and over 30×10^{12} possible profiles, respectively. Hence, only the biplots for sets 3 and 4 can show the 500 customer points although, even there, some points may coincide.

Figure 2

Visual inspection of the cloud of customer points shows two outliers in the points in set four. Two clearly quite atypical customers, in terms of the satisfaction variables, can be discerned. Compared to the other customers, they must have extreme scores on category 6 of variable 2 and category 2 of variable 11. The lower number of possible profiles makes the biplots for set 1 and set 2 easier to inspect.

Two groups of customers can be identified when inspecting the group average configuration (see Figure 3, indicated by the ellipses), the right group is characterised by the variables and categories indicated in the figure.

Figure 3

We display the category quantifications for all variables in each of the four sets in the separate panels of Figure 4. This figure shows that indeed, most variables should be quantified as nominal variables. For instance variable 1 of set 3, indicating 'type of house/area' (a nominally scaled variable), clearly shows non-monotonic category quantifications. The quantifications for the ordinal categories of variable 1 in set 2, indicating the number of automatically paid contracts are also non-monotonic. This indicates a non linear relationship between this variable and other variables.

An analysis that assumes a ratio, an interval, or an ordinal scaling would harm the optimality of the GPA solution. Note that if the quantifications of two categories of a variable are the same, one could combine these categories.

Figure 4

We analysed the same data sets by performing a GPA on the indicator matrices $\mathbf{G}_k$, $k = 1, \cdots, 4$. The loss value (2) for this analysis is 2.74, while the value of (2) for the OS-GPA method was 2.5. This confirms that the OS-GPA method attains a greater agreement between the data sets, because the category quantifications are optimal with respect to the minimisation of (2), while the 'quantifications' obtained by the GPA on the $\mathbf{G}_k$ do not necessarily minimize (2).

4   Discussion and conclusion

The inclusion of an optimal scaling step to ordinary GPA is a useful extension of the method. It allows for categorical variables in different sets to be compared under the orthogonal affine assumptions of GPA. A visual inspection of the results enables a scan of potential groupings of row-objects (here customers), of outliers and of ways of treating the categories of variables. The quantifications of most variables show that an ordinal or numerical treatment would have been sub-optimal compared to our OS-GPA approach.

We have considered the case where each of the K sets use different variables. When the same variables are used in each set, then only one set of quantifications is required. This is not the ideal structure for using Procrustean methods because now it would usually be better to analyse how means vary from set to set. One situation where the approach may be useful is when it is suspected that although variables may have the same names across sets, they may not be similarly interpreted. Then it is interesting to compare two analyses, one assuming different quantifications and the other the same quantifications. The details of the modifications required for evaluating common optimal scores are available on request.

Appendix A

Algorithm in pseudo code:

1. Initialisation

a. indicator matrices $\mathbf{G}_k$ ($n \times \sum_{j=1}^{p} c_{kj}$), where $\mathbf{G}_k = \left[\mathbf{G}_{k1}, \mathbf{G}_{k2}, \cdots, \mathbf{G}_{kp} \right]$;

b. rotation matrices $\mathbf{Q}_k$ ($p \times p$), $k = 1, \cdots, K$ are identity matrices of appropriate order;

c. category quantification matrices $\mathbf{Z}_k$ ($\sum_{j=1}^{p} c_{kj} \times p_k$), $\mathbf{Z}_k = [\mathbf{z}_{k1}, \mathbf{z}_{k2}, \cdots, \mathbf{z}_{kp_k}]$.

2. Main ALS iteration until $\sum_{h<k}^{K} \left\| \mathbf{G}_h \mathbf{Z}_h \mathbf{Q}_h - \mathbf{G}_k \mathbf{Z}_k \mathbf{Q}_k \right\| < \varepsilon$, a small positive number,

a. optimal scaling: $\bar{\mathbf{Y}}_{-k} = \frac{1}{K-1} \sum_{h \neq k}^{K} \mathbf{X}_h \mathbf{Q}_h$, with $\bar{\mathbf{Y}}_{-k} \mathbf{Q}'_k$ centred as in equation (13);

$\mathbf{L}_{kj} = \mathbf{G}'_{kj} \mathbf{G}_{kj}$; $\mathbf{z}_{kj} = \mathbf{L}_{kj}^{-1} \mathbf{G}'_j \bar{\mathbf{Y}}_{-k} \mathbf{Q}'_k / \sqrt{\mathbf{Q}_k \bar{\mathbf{Y}}'_{-k} \mathbf{G}_k \mathbf{L}_{kj}^{-1} \mathbf{G}'_k \bar{\mathbf{Y}}_{-k} \mathbf{Q}'_k}$ providing $\mathbf{z}_{kj}$ for given $\mathbf{Y}$ and $\mathbf{Q}$, for each $k = 1, \cdots, K$ (possibly taking into account p_k-scaling);

b. GPA, providing rotation matrices $\mathbf{Q}_k$ minimising $\sum_{h<k}^{K} \left\| \mathbf{G}_h \mathbf{Z}_h \mathbf{Q}_h - \mathbf{G}_k \mathbf{Z}_k \mathbf{Q}_k \right\|$ for given $\mathbf{Z}_k$, for each $k = 1, \cdots, K$;

3. convergence test, going back to step 2 when $\sum_{h<k}^{K} \left\| \mathbf{G}_h \mathbf{Z}_h \mathbf{Q}_h - \mathbf{G}_k \mathbf{Z}_k \mathbf{Q}_k \right\| < \varepsilon$;

4. perform PCA on $\mathbf{Y} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{G}_k \mathbf{Z}_k \mathbf{Q}_k$, give the $\mathbf{G}_k \mathbf{Z}_k \mathbf{Q}_k$ the same rotation, say $\mathbf{V}$, with

$\mathbf{Y} = \mathbf{USV}'$ the SVD of $\mathbf{Y}$; as well as the $\mathbf{Z}_k \mathbf{Q}_k$; plot $\mathbf{YV}$, $\mathbf{G}_k \mathbf{Z}_k \mathbf{Q}_k \mathbf{V}$ and $\mathbf{Z}_k \mathbf{Q}_k \mathbf{V}$.

Appendix B

Here we show that the value of $\mathbf{z}_{kj}$ given by (13) minimises the objective function (9). For simplicity, in this section we drop the suffices k and j. To establish a minimum requires the second derivative of (9), given by differentiating (10). If we do this and use (12), we get the second differential as:

$$(1-\lambda)\mathbf{L1} = \left(\mathbf{z}'\mathbf{G}'\mathbf{y}\right)\mathbf{L1}. \tag{B1}$$

On substituting for $\mathbf{z}$ from (13), we have:

$$
\begin{aligned}
\left(\mathbf{z}'\mathbf{G}'\mathbf{y}'\right)\mathbf{L1} &= [\mathbf{y}'(\mathbf{I}-\mathbf{N})\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'(\mathbf{I}-\mathbf{N})\mathbf{y}]^{-1/2}\left(\mathbf{y}'(\mathbf{I}-\mathbf{N})\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'\mathbf{y}\right)\mathbf{L1} \\
&= [\mathbf{y}'(\mathbf{I}-\mathbf{N})\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'(\mathbf{I}-\mathbf{N})\mathbf{y}]^{-1/2}\,\mathbf{y}'\left(\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'-\mathbf{11}'/n\right)\mathbf{y}\mathbf{L1},
\end{aligned} \tag{B2}
$$

where the term in square brackets is assumed positive (confirmed below) and $\mathbf{L1}$ necessarily has positive frequency elements.

Now $\left(\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'\right)\mathbf{1}=\mathbf{G}'\mathbf{1}=\mathbf{1}$, so $\mathbf{1}$ is an eigenvector of the positive semi definite matrix $\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'$. It follows that

$$\mathbf{y}'\left(\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'-\mathbf{11}'/n\right)\mathbf{y} = \mathbf{y}'\left(\sum_{i=1}^{p_k-1}\gamma_i\mathbf{v}_i\mathbf{v}_i'\right)\mathbf{y} = \sum_{i=1}^{p_k-1}\gamma_i(\mathbf{v}_i'\mathbf{y})^2, \tag{B3}$$

where γ_i, $\mathbf{v}_i$ are an eigenvalue, necessarily positive, and associated eigenvector of the matrix on the left-hand-side. It follows that, apart from a pathological case about to be discussed, all the components of the second differential (B2) are positive, which is the condition for a minimum. Additionally this confirms that the positive square root must be taken in (13).

The pathological case is in the unlikely event where $\mathbf{y}$ happens to be $\mathbf{1}$ or some other null-vector of the matrix on the left-hand-side of (B3). Then (B2) is zero and we have not demonstrated a minimum. When $\mathbf{y}=\mathbf{1}$, then the second term of (8) becomes zero as $\mathbf{z}'\mathbf{G}'\mathbf{1}=\mathbf{z}'\mathbf{L1},$ which is constrained to be zero. Thus (8) has an infinite number of exact solutions satisfying $\mathbf{z}'\mathbf{L}\mathbf{z}=\mathbf{1}$ and $\mathbf{1}'\mathbf{L}\mathbf{z}=0$. Similarly, when $\mathbf{y}$ is any other null vector, then $\mathbf{1}'\mathbf{y}=0$, so that $\left(\mathbf{G}\mathbf{L}^{-1}\mathbf{G}'\right)\mathbf{y}=0$, which in turn implies that $\mathbf{G}'\mathbf{y}=0$. Again the second term of (8) vanishes and we have the same exact solutions as when $\mathbf{y}=\mathbf{1}$. Thus the pathological solutions, should they occur, may not correspond formally to analytical minima but they give exact fits that cannot be bettered.

Acknowledgements

References

Blasius, J., Eilers, P.H.C., Gower, J.C., 2009, Better biplots, Computational Statistics & Data Analysis, in press.

Borg, I., Groenen, P., 1997. Modern multidimensional scaling. Theory and applications. Springer, New York.

Dijksterhuis, G.B., Gower, J.C., 1991. The Interpretation of Generalized Procrustes analysis and Allied Methods. Food Quality and Preference 3, 67-87.

Dijksterhuis, G.B., Martens, H., Martens, M., 2005. Combined Procrustes analysis and PLSR for internal and external mapping of data from multiple sources. Computational Statistics & Data Analysis 48(1), 47-62.

Gardner, S., Gower, J.C., le Roux, N.J., 2006. A synthesis of canonical variate analysis, generalised canonical correlation and Procrustes analysis. Computational Statistics & Data Analysis 50(1), 107-134.

Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, Chichester.

Gower, J.C., 1975. Generalised Procrustes Analysis. Psychometrika 40, 33-51.

Gower, J.C., Dijksterhuis, G.B., 2004. Procrustes Problems. Oxford University Press, Oxford.

Gower, J.C. and Hand, D.J., 1996. Biplots. Chapman and Hall, London.

Jolliffe, I., 2002. Principal component analysis. 2nd edition. Springer, New York.

Steenkamp, J.B.E.M., Van Trijp, H.C.M., Ten Berge, J.M.F., 1994. Perceptual Mapping Based on Idiosyncratic Sets of Attributes. Journal of Marketing Research 31(1), 15-27.

Van Buuren, S., Dijksterhuis, G.B., 1988. Procrustes Analysis of discrete data. In Jansen, M.G.H., van Schuur, W.H., (Eds.), The many faces of multivariate analysis. Proceedings of the SMABS-88 conference. Volume I. RION, Groningen.

Van der Burg, E., 1988. Nonlinear Canonical Correlation and Some Related Techniques. DSWO, Press Leiden.
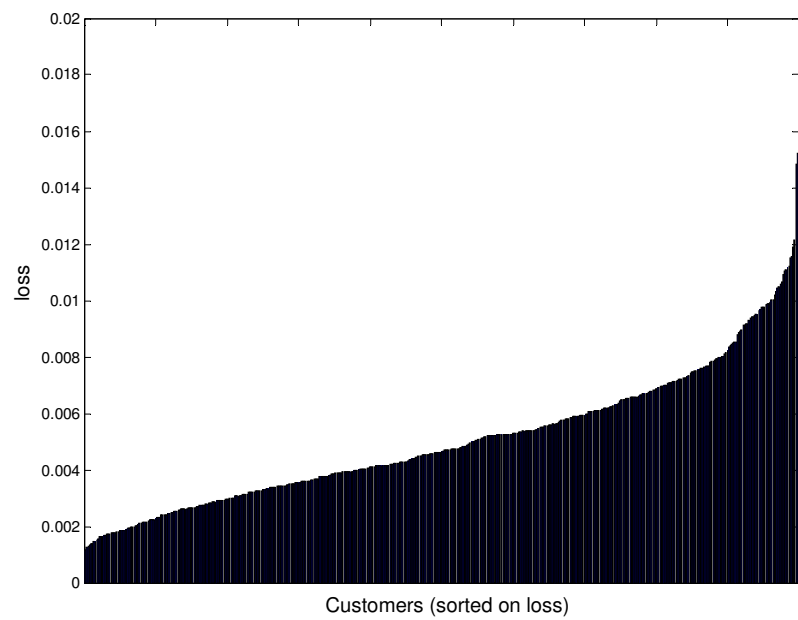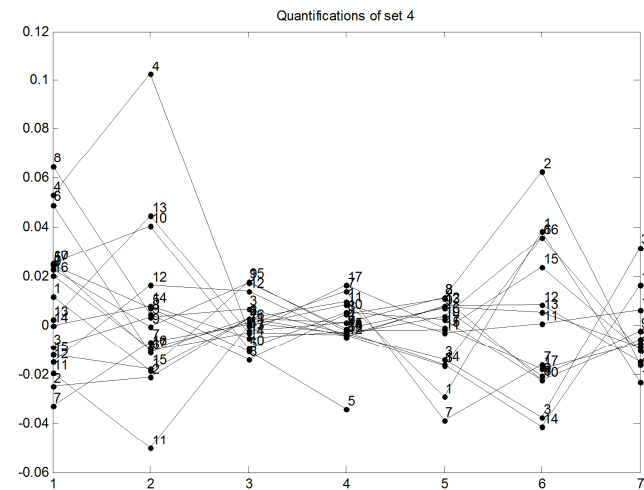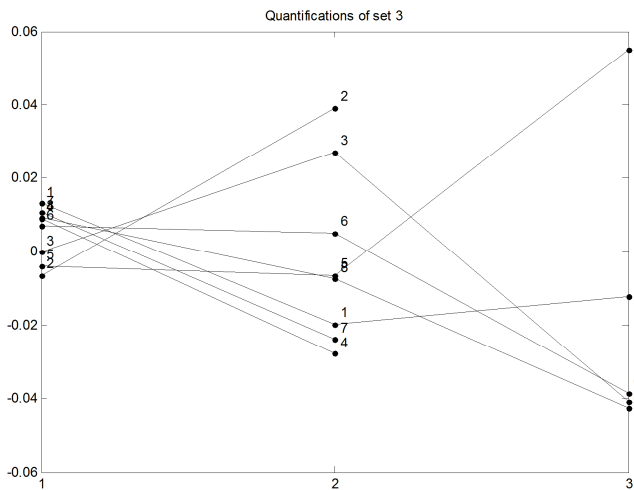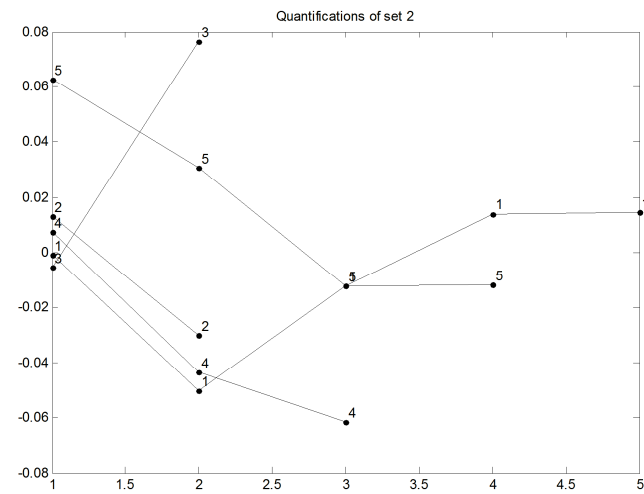
FIGURES

Figure 1

Loss distributed over 500 customers, sorted from low to high loss.

Set 1: Power Consumption

Set 2: Costs

Set 3: Demographics

Set 4: Satisfaction

Figure 2

Biplots of the first two dimensions of customer-points ($\mathbf{G}_k\mathbf{Z}_k\mathbf{Q}_k\mathbf{V}$) and variable category points ($\mathbf{Z}_k\mathbf{Q}_k\mathbf{V}$) per set labelled by three numbers

(variable, category, set).

Figure 3

The first two dimensions of a PCA of $\mathbf{Y}$, the group averages, of the 500 customers labelled by their ID number, and variable categories, see equation (13), labelled by three numbers (variable, category, set). The two ellipses indicate a potentially interesting clustering of customers.
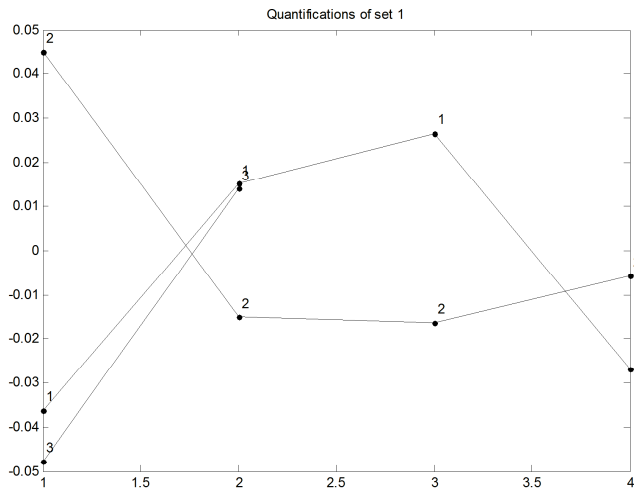
2   Figure 4

3   Category quantifications $\mathbf{z}_{kj}$, (one panel per set $k = 1, \cdots, 4$). Numbers refer to variables within each set.

4