

Econometric Institute Report Nr. EI 2014-23

Least-squares Bilinear Clustering of Three-way Data

**Pieter C. Schoonees Patrick J.F. Groenen
Michel van de Velden**

Econometric Institute, Erasmus University Rotterdam
The Netherlands

February 16, 2015

Corresponding author: P. C. Schoonees, Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: schoonees@gmail.com.

We gratefully acknowledge the input of Karl Jöreskog, whose question at a seminar in Uppsala sowed the seeds for this research.

Least-squares Bilinear Clustering of Three-way Data

Pieter C. Schoonees Patrick J.F. Groenen Michel van de Velden
Erasmus University Rotterdam

Abstract

A least-squares bilinear clustering framework for modelling three-way data, where each observation consists of an ordinary two-way matrix, is introduced. The method combines bilinear decompositions of the two-way matrices into overall means, row margins, column margins and row-column interactions with clustering along the third way. Different clusterings are defined for each part of the decomposition, so that up to four different classifications are defined jointly. The computational burden is greatly reduced by the orthogonality of the bilinear model, such that the joint clustering problem reduces to separate ones which can be handled independently. Three of these sub-problems are specific cases of k -means clustering; a special algorithm is formulated for the row-column interactions, which are displayed in clusterwise biplots. The method is illustrated via two empirical examples and interpreting the interaction biplots are discussed.

Keywords: three-way data, bilinear decomposition, k -means cluster analysis, least-squares estimation, biplots.

1. Introduction

Three-way data appear regularly in research, such as when a number of respondents are asked to rate several objects based on a set of characteristics in a marketing survey. Such data can be collected in a three-way array, with slices along one dimension containing the data matrices for the different individuals. Several models have been formulated for least-squares approximation of such three-way arrays, such as CANDECOMP/PARAFAC (Carroll and Chang 1970; Harshman 1970) and TUCKALS3 (Kroonenberg and De Leeuw 1980). Much research is available on theoretical aspects of these models, nonuniqueness properties, and estimation algorithms (e.g. Kruskal 1977; Kiers and Krijnen 1991; ten Berge and Sidiropoulos 2002; Faber, Bro, and Hopke 2003). However, these models suffer from disadvantages including that they are relatively complicated, hard to fit and that the interpretation of their graphical representations require sound knowledge of the models (e.g. Kiers 2000; Krijnen, Dijkstra, and Stegeman 2008).

Here our aim is to simplify the analysis of such data by using a simple model for the two-way matrix slices of the three-way data array, combined with clustering over the third way. The model we elaborate on is the well-known bilinear decomposition of a matrix into an overall mean, row means, column means, and a low-rank decomposition of the remaining row-column interactions (see [Gower and Hand 1996](#), for example). This can be viewed as a two-way analysis of variance-type decomposition into an overall mean effect, marginal effects and row-column interaction effects, and is also known as a biadditive model.

The novelty of the least-squares bilinear clustering model proposed in this paper, henceforth referred to as LSBCLUST, is that the clusters over the third way of the array are introduced jointly for each of the terms in the bilinear decomposition. Importantly, however, we show that this clustering can be done separately for each of the terms because of the orthogonality of the terms in the bilinear model (with one exception – see [Section 2.1](#)). This property greatly reduces the computational cost of the method, and aids interpretation of the results. Additionally, we show how to construct biplots for the interaction effects so that these can be easily interpreted, and how different choices for the identifiability constraints in the bilinear model lead to different submodels.

The main ideas of LSBCLUST are summarized in [Figure 1](#). The data array represented by \mathbf{X} is decomposed into overall means, row means, column means and row-column interaction. Each of these components are modelled clusterwise, with different sets of clusters introduced for each of the components of the decomposition. The clusters are represented by different colours and/or labelled effects in the figure. Within each cluster the effects are determined by modelling the cluster means, and in the case of the interactions low-rank decompositions of the cluster means are used. This enables us to elicit only the most prominent structure in the interactions and improves interpretability by allowing biplots to be constructed. We note that adjusting for different row and column margins pre-analysis is routinely done in correspondence analysis (e.g. [Greenacre 2007](#)).

The problem is formulated as a least-squares loss function in [Section 2](#). We also discuss the separability of the different clustering problems in this section. In [Section 3](#) we develop an alternating least-squares algorithm for minimizing the loss function and discuss the construction of biplots for the interactions, as well as model selection. [Section 4](#) contains two illustrative empirical applications, and [Section 5](#) concludes.

2. Problem Formulation

Consider as starting point data consisting of J objects that have been rated by N individuals on K attributes. Define the indices i, j and k such that these identify the respondents, objects and attributes respectively. It follows that $i = 1, \dots, N$, $j = 1, \dots, J$ and $k = 1, \dots, K$. For each individual, form the $J \times K$ matrix \mathbf{X}_i where each row j gives the scores of person i

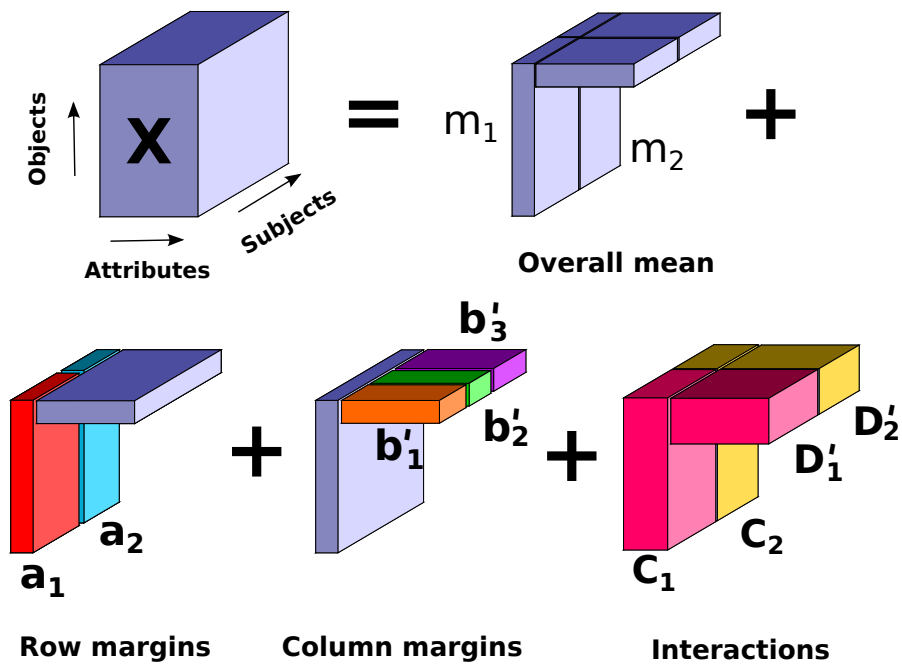


Figure 1: A diagram illustrating our LSBCLUST approach to analysing three-way data. Colours and/or labels indicate different clusters, while pale blue indicates constant vectors or matrices. Different clusters are introduced for all four parts of the bilinear decomposition of the two-way matrix slices. The cluster means are modelled within each such cluster, and the clustering and modelling steps are undertaken jointly.

for object j on all K attributes. For now, we assume that there are no missing values and that the values in the \mathbf{X}_i 's are commensurable, that is, measured on the same scale. The collection of these \mathbf{X}_i 's can also be viewed as a three-way array, with the objects, attributes and individuals constituting the three dimensions.

The \mathbf{X}_i 's are modelled by a variation of the bilinear (or biadditive) decomposition in the least-squares loss function

$$L(m, \mathbf{a}, \mathbf{b}, \mathbf{C}, \mathbf{D}) = \sum_{i=1}^N \left\| \mathbf{X}_i - \left(m \mathbf{1}_J \mathbf{1}_K' + \mathbf{a} \mathbf{1}_K' + \mathbf{1}_J \mathbf{b}' + \mathbf{C} \mathbf{D}' \right) \right\|^2. \quad (1)$$

Here m denotes an overall constant or average effect, \mathbf{a} the row effects, \mathbf{b} the column effects, $\mathbf{C} \mathbf{D}'$ a low-rank decomposition of the interaction effects between rows and columns, and $\|\cdot\|$ the Frobenius norm. Also, $\mathbf{1}_K$ denotes the length- K vector of ones. Representing the interaction effects as inner products permit these to be displayed in biplots (Gower and Hand 1996; Gower, Lubbe, and Le Roux 2011). To ensure uniqueness of the model, the usual sum-to-zero constraints $\mathbf{a}' \mathbf{1}_J = \mathbf{b}' \mathbf{1}_K = 0$ and $\mathbf{1}_J' \mathbf{C} = \mathbf{1}_K' \mathbf{D} = \mathbf{0}$ must be imposed. Additionally, the columns of \mathbf{C} and \mathbf{D} are required to be orthogonal and of equal length (for more information, see Denis and Gower 1994). Model (1) has an analytic solution.

Our main contribution is to embed (1) in a general modelling framework by adding several different types of clusters (latent classes or segments) while allowing for a variety of parameter constraints. Different choices for these constraints lead to different submodels. Clusters are introduced to separate the respondents on four different characteristics: the first with respect to the overall average, the second for the row effects, the third for the column effects, and the fourth for the interaction effects. For modelling the interactions, we allow for three options: (a) a common \mathbf{C}_1 for representing the rows and a differential \mathbf{D}_u for each interaction cluster indexed by u ; (b) a differential \mathbf{C}_u but common \mathbf{D}_1 for each interaction cluster; or (c) both \mathbf{C}_u and \mathbf{D}_u are specific to the interaction cluster. Options (a) and (b) are more parsimonious than (c) and are particularly useful for linking the graphical representations (biplots) of the clusterwise interaction effects through common row or column representations. In Section 3.3 we consider generalized Procrustes analysis (Gower and Dijksterhuis 2004) as an interpretative aid for option (c).

Let $\mathbf{G}^{(o)}$ be the $N \times R$ matrix of cluster memberships for the overall constant, which has $g_{ir}^{(o)} = 1$ if person i belongs to cluster r and $g_{ir}^{(o)} = 0$ otherwise ($r = 1, 2, \dots, R$). Similarly, $\mathbf{G}^{(r)}$ is the $N \times S$ matrix of cluster memberships for the row effects, $\mathbf{G}^{(c)}$ the $N \times T$ matrix of cluster memberships for the column effects, and $\mathbf{G}^{(i)}$ the $N \times U$ matrix of cluster memberships for the interaction effects. Now, by incorporating the clustering, the least-squares loss function

becomes

$$\begin{aligned}
L(\mathbf{G}^{(o)}, \mathbf{G}^{(r)}, \mathbf{G}^{(c)}, \mathbf{G}^{(i)}, \mathbf{m}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
&= \sum_{i=1}^N \sum_{r=1}^R \sum_{s=1}^S \sum_{t=1}^T \sum_{u=1}^U g_{ir}^{(o)} g_{is}^{(r)} g_{it}^{(c)} g_{iu}^{(i)} \left\| \mathbf{X}_i - \left(m_r \mathbf{1}_J \mathbf{1}'_K + \mathbf{a}_s \mathbf{1}'_K + \mathbf{1}_J \mathbf{b}'_t + \mathbf{C}_u \mathbf{D}'_u \right) \right\|^2 \\
&= \sum_{i,r,s,t,u} g_{ir}^{(o)} g_{is}^{(r)} g_{it}^{(c)} g_{iu}^{(i)} L(i|r, s, t, u).
\end{aligned} \tag{2}$$

Here $\mathbf{m} = (m_1, \dots, m_R)'$, $\mathbf{A} = [\mathbf{a}_1 \ \dots \ \mathbf{a}_S]$, $\mathbf{B} = [\mathbf{b}_1 \ \dots \ \mathbf{b}_T]$, and $\mathbf{C}' = [\mathbf{C}'_1 \ \dots \ \mathbf{C}'_U]$ and $\mathbf{D}' = [\mathbf{D}'_1 \ \dots \ \mathbf{D}'_U]$. In case options (a) or (b) are used, $\mathbf{C}_u \mathbf{D}'_u$ in (2) should be replaced by $\mathbf{C}_1 \mathbf{D}'_u$ or $\mathbf{C}_u \mathbf{D}'_1$ respectively.

2.1. Separability of Different Cluster Types

Here we show that the joint clustering can be simplified significantly into four separate clustering problems. Define the matrix $\mathbf{J}_J^{(\delta)}$ to be of the form

$$\mathbf{J}_J^{(\delta)} = \mathbf{I}_J - \frac{\delta}{J} \mathbf{1}_J \mathbf{1}'_J, \tag{3}$$

where $\delta \in \{0, 1\}$. Hence when $\delta = 1$, $\mathbf{J}_J^{(\delta)}$ is the J -dimensional centring matrix; otherwise it reduces to the identity matrix. The shorthand notation $\mathbf{J}_J = \mathbf{J}_J^{(1)}$ will also be used. Applying a generalization of the well-known double-centring operation to \mathbf{X}_i yields

$$\mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)} = \left(\mathbf{I}_J - \frac{\delta_1}{J} \mathbf{1}_J \mathbf{1}'_J \right) \mathbf{X}_i \left(\mathbf{I}_K - \frac{\delta_2}{K} \mathbf{1}_K \mathbf{1}'_K \right)'. \tag{4}$$

Expanding and rearranging the above implies that

$$\mathbf{X}_i = -\delta_1 \delta_2 \frac{\mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K}{JK} \mathbf{1}_J \mathbf{1}'_K + \frac{\delta_1}{J} \mathbf{1}_J \mathbf{1}'_J \mathbf{X}_i + \frac{\delta_2}{K} \mathbf{X}_i \mathbf{1}_K \mathbf{1}'_K + \mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)}. \tag{5}$$

Furthermore, (5) can be rewritten as

$$\begin{aligned}
\mathbf{X}_i &= (\delta_1 \delta_3 + \delta_2 \delta_4 - \delta_1 \delta_2) \frac{\mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K}{JK} \mathbf{1}_J \mathbf{1}'_K + \frac{\delta_1}{J} \mathbf{1}_J \mathbf{1}'_J \mathbf{X}_i \mathbf{J}_K^{(\delta_3)} \\
&\quad + \frac{\delta_2}{K} \mathbf{J}_J^{(\delta_4)} \mathbf{X}_i \mathbf{1}_K \mathbf{1}'_K + \mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)}
\end{aligned} \tag{6}$$

by inserting additional centring matrices operating on the row and column means. The binary variables in $\boldsymbol{\delta}' = (\delta_1, \delta_2, \delta_3, \delta_4)$ act as switches which determine which centring constraints are applied and are chosen by the user according to the problem at hand. The choice of these switches determines the biadditive model to be fitted. Although there are 16 different choices

for the δ 's, there are fewer unique specifications. For example, notice that if $\delta_1 = 0$, the second term on the right-hand side of (6) vanishes so that δ_3 does not have any effect. The choices $\delta_3 = 0$ and $\delta_3 = 1$ therefore leads to the same decomposition when $\delta_1 = 0$. A similar relationship exists between δ_2 and δ_4 .

By using the same centring approach in the model specification, we can drop the sum-to-zero constraints from the formulation. This is done by redefining the terms in the summation in (2) as

$$L(i|r, s, t, u) = \left\| \mathbf{X}_i - \left((\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2)m_r\mathbf{1}_J\mathbf{1}'_K + \delta_2\mathbf{J}_J^{(\delta_4)}\mathbf{a}_s\mathbf{1}'_K + \delta_1\mathbf{1}_J\mathbf{b}'_t\mathbf{J}_K^{(\delta_3)} + \mathbf{J}_J^{(\delta_1)}\mathbf{C}_u\mathbf{D}'_u\mathbf{J}_K^{(\delta_2)} \right) \right\|^2. \quad (7)$$

Note that using $\delta_1 = 1$, $\delta_2 = 1$, $\delta_3 = 1$ or $\delta_4 = 1$ enforces the sum-to-zero constraints on the columns of \mathbf{C}_u or \mathbf{D}_u , or on \mathbf{b}_t or \mathbf{a}_s respectively. For example, estimating the parameters in $\mathbf{J}_J\mathbf{a}_s$ is equivalent to estimating \mathbf{a}_s subject to $\mathbf{1}'_J\mathbf{a}_s = 0$.

We can now associate each of the terms in (6) with the corresponding terms in the model (7), by substituting (6) in (7) so that

$$L(i|r, s, t, u) = \left\| (\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - m_r \right) \mathbf{1}_J \mathbf{1}'_K + \delta_1 \mathbf{1}_J \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i - \mathbf{b}'_t \right) \mathbf{J}_K^{(\delta_3)} + \delta_2 \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \mathbf{1}'_K + \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}'_u \right) \mathbf{J}_K^{(\delta_2)} \right\|^2. \quad (8)$$

It can be shown (see [Appendix A](#) for an overview) that for all choices of $\boldsymbol{\delta}$, with the exception of the case $\boldsymbol{\delta}' = (1, 1, 0, 0)$, the decomposition is orthogonal such that

$$\begin{aligned} L(i|r, s, t, u) &= JK \left\| (\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - m_r \right) \right\|^2 + J \left\| \delta_1 \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i - \mathbf{b}'_t \right) \mathbf{J}_K^{(\delta_3)} \right\|^2 \\ &\quad + K \left\| \delta_2 \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \right\|^2 + \left\| \mathbf{J}_J^{(\delta_1)} (\mathbf{X}_i - \mathbf{C}_u \mathbf{D}'_u) \mathbf{J}_K^{(\delta_2)} \right\|^2 \\ &= L_{(o)}(i|r) + L_{(r)}(i|s) + L_{(c)}(i|t) + L_{(i)}(i|u). \end{aligned} \quad (9)$$

This equality follows from the fact that all the cross-products are zero. Furthermore, the orthogonality leads to a profound simplification in the clustering, since now the loss function

(2) equals

$$\begin{aligned}
& L(\mathbf{G}^{(o)}, \mathbf{G}^{(r)}, \mathbf{G}^{(c)}, \mathbf{G}^{(i)}, \mathbf{m}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
&= \sum_{i,r,s,t,u} \left\{ g_{ir}^{(o)} L_{(o)}(i|r) + g_{is}^{(r)} L_{(r)}(i|s) + g_{it}^{(c)} L_{(c)}(i|t) + g_{iu}^{(i)} L_{(i)}(i|u) \right\} \\
&= JK \sum_{i=1}^N \sum_{r=1}^R g_{ir}^{(o)} \left\| (\delta_1 \delta_3 + \delta_2 \delta_4 - \delta_1 \delta_2) \left(\frac{1}{JK} \mathbf{1}_J' \mathbf{X}_i \mathbf{1}_K - m_r \right) \right\|^2 \\
&\quad + K \sum_{i=1}^N \sum_{s=1}^S g_{is}^{(r)} \left\| \delta_2 \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \right\|^2 \\
&\quad + J \sum_{i=1}^N \sum_{t=1}^T g_{it}^{(c)} \left\| \delta_1 \left(\frac{1}{J} \mathbf{1}_J' \mathbf{X}_i - \mathbf{b}_t' \right) \mathbf{J}_K^{(\delta_3)} \right\|^2 \\
&\quad + \sum_{i=1}^N \sum_{u=1}^U g_{iu}^{(i)} \left\| \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right\|^2 \\
&= L_{(o)}(\mathbf{G}^{(o)}, \mathbf{m}) + L_{(r)}(\mathbf{G}^{(r)}, \mathbf{A}) + L_{(c)}(\mathbf{G}^{(c)}, \mathbf{B}) + L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D}). \tag{10}
\end{aligned}$$

Consequently, the joint clustering reduces to separate clusterings on the overall mean, row margins, column margins and interactions respectively. To see why simplification in (10) is possible after utilizing (9), consider the summation

$$\sum_{r=1}^R \sum_{s=1}^S \sum_{t=1}^T \sum_{u=1}^U g_{ir}^{(o)} g_{is}^{(r)} g_{it}^{(c)} g_{iu}^{(i)} \left\| \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right\|^2. \tag{11}$$

As the term $\left\| \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right\|^2$ only depends on the subscript u , the summations of the other cluster types are equal to one because for each i we have $\sum_{r=1}^R g_{ir}^{(r)} = 1$, $\sum_{s=1}^S g_{is}^{(c)} = 1$, and $\sum_{u=1}^U g_{iu}^{(i)} = 1$. The same holds for the other terms.

Different choices for $\boldsymbol{\delta}$ lead to different model specifications, the choice of which should be guided by the application. The nine different cases are summarized in Table 2.1. Note that Model 6 does not lead to an orthogonal decomposition and is therefore not discussed further. The most general model is Model 1, which amounts to a clusterwise low-rank decomposition of the mean. All the other models are essentially special cases of this model, stemming from specific forms for \mathbf{C}_u or \mathbf{D}_u or both. For example, Model 4 can be written as $\mathbf{C}_u^* \mathbf{D}_u^{*'} where $\mathbf{C}_u^* = \begin{bmatrix} \mathbf{1}_J & \mathbf{J}_J \mathbf{C}_u \end{bmatrix}$ and $\mathbf{D}_u^* = \begin{bmatrix} \mathbf{b}_t & \mathbf{D}_u \end{bmatrix}$. We note that the added clustering adds additional restrictions on the parameters which is not the case in ordinary bilinear models. Besides the choice of $\boldsymbol{\delta}$, the choice of either (a) $\mathbf{C}_1 \mathbf{D}_u'$, (b) $\mathbf{C}_u \mathbf{D}_1'$ or (c) $\mathbf{C}_u \mathbf{D}_u'$ must be made.$

3. Algorithm

Model	δ_1	δ_2	δ_3	δ_4	Model for \mathbf{X}_i	
1	0	0	0	0		$\mathbf{C}_u \mathbf{D}'_u$
2	0	1	0	0	$\mathbf{a}_s \mathbf{1}'_K +$	$\mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K$
3	0	1	0	1	$m_r \mathbf{1}_J \mathbf{1}'_K + \mathbf{J}_J \mathbf{a}_s \mathbf{1}'_K +$	$\mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K$
4	1	0	0	0		$\mathbf{1}_J \mathbf{b}'_t + \mathbf{J}_J \mathbf{C}_u \mathbf{D}'_u$
5	1	0	1	0	$m_r \mathbf{1}_J \mathbf{1}'_K +$	$\mathbf{1}_J \mathbf{b}'_t \mathbf{J}_K + \mathbf{J}_J \mathbf{C}_u \mathbf{D}'_u$
6	1	1	0	0	$-m_r \mathbf{1}_J \mathbf{1}'_K +$	$\mathbf{a}_s \mathbf{1}'_K + \mathbf{1}_J \mathbf{b}'_t + \mathbf{J}_J \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K$
7	1	1	0	1		$\mathbf{J}_J \mathbf{a}_s \mathbf{1}'_K + \mathbf{1}_J \mathbf{b}'_t + \mathbf{J}_J \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K$
8	1	1	1	0		$\mathbf{a}_s \mathbf{1}'_K + \mathbf{1}_J \mathbf{b}'_t \mathbf{J}_K + \mathbf{J}_J \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K$
9	1	1	1	1	$m_r \mathbf{1}_J \mathbf{1}'_K + \mathbf{J}_J \mathbf{a}_s \mathbf{1}'_K + \mathbf{1}_J \mathbf{b}'_t \mathbf{J}_K + \mathbf{J}_J \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K$	

Table 1: A summary of the models implied by different choices of $\boldsymbol{\delta}$. Note that Model 6 is not orthogonal and is only included for completeness.

Due to the form of the loss function (10), we can treat each of the components separately. Conveniently, the loss functions $L_{(o)}(\mathbf{G}^{(o)}, \mathbf{m})$, $L_{(r)}(\mathbf{G}^{(r)}, \mathbf{A})$ and $L_{(c)}(\mathbf{G}^{(c)}, \mathbf{B})$ are specific k -means problems (e.g. [Everitt, Landau, Leese, and Stahl 2011](#)), on the rows of (say) the data matrix $\mathbf{Y} : N \times d$. In each of these cases, \mathbf{Y} can be defined as follows:

- For estimating $\mathbf{G}^{(o)}$ and \mathbf{m} , \mathbf{Y} has a single column ($d = 1$) containing the overall means $\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K$ of the $\{\mathbf{X}_i\}$;
- For estimating $\mathbf{G}^{(r)}$ and \mathbf{A} , the rows of \mathbf{Y} ($d = J$) consist of the row mean vectors $\frac{1}{K} \mathbf{1}'_K \mathbf{X}_i$; and
- For estimating $\mathbf{G}^{(c)}$ and \mathbf{B} , the rows of \mathbf{Y} ($d = K$) are the column mean vectors $\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i$.

Hence optimizing $L_{(o)}(\mathbf{G}^{(o)}, \mathbf{m})$, $L_{(r)}(\mathbf{G}^{(r)}, \mathbf{A})$ and $L_{(c)}(\mathbf{G}^{(c)}, \mathbf{B})$ can resort to standard methods for k -means on the overall mean, row margins and column margins respectively. Also, there are a variety of tools available for selecting R, S and T . We stress that caution is required with respect to local minima in k -means clustering, which can be alleviated by using multiple random starts.

The degrees-of-freedom for estimating the clusterwise overall means are $(\delta_1 \delta_3 + \delta_2 \delta_4 - \delta_1 \delta_2)R$, while the degrees-of-freedom associated with estimating the clusterwise row and column means are $\delta_2 S(J - \delta_4)$ and $\delta_1 T(K - \delta_3)$ respectively. The degrees-of-freedom for the interactions depends on the choice of (a), (b) or (c), and are deferred to the next section.

3.1. An Algorithm for the Interaction Clustering

We proceed to formulate a special algorithm for minimizing $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$ based on block-relaxation methods (see for example [De Leeuw 1994](#)). The proposed algorithm iterates over

optimizing one set of parameters while keeping the others fixed. Specifically, in step (1) we consider $\mathbf{G}^{(i)}$ fixed and simultaneously update \mathbf{C} and \mathbf{D} , while step (2) consists of updating $\mathbf{G}^{(i)}$ while keeping \mathbf{C} and \mathbf{D} fixed at the values obtained in step (1). Finally, steps (1) and (2) are repeated until numerical convergence of the loss function is observed. This algorithm is guaranteed to converge monotonically, but only to local minima. It must be initialized by a starting configuration for $\mathbf{G}^{(i)}$. To increase the likelihood of locating the global minimum, it is advisable to use multiple (random) starting values for $\mathbf{G}^{(i)}$.

We now describe the steps of our algorithm in more detail. To minimize $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$ over \mathbf{C} and \mathbf{D} for fixed $\mathbf{G}^{(i)}$, it is useful to rewrite it as follows:

$$\begin{aligned} L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D}) &= \sum_{i=1}^N \sum_{u=1}^U g_{iu}^{(i)} \left\| \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \bar{\mathbf{X}}_u + \bar{\mathbf{X}}_u - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right\|^2 \\ &= \sum_{i=1}^N \sum_{u=1}^U g_{iu}^{(i)} \left\| \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \bar{\mathbf{X}}_u \right) \mathbf{J}_K^{(\delta_2)} \right\|^2 + \sum_{u=1}^U N_u \left\| \mathbf{J}_J^{(\delta_1)} \left(\bar{\mathbf{X}}_u - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right\|^2. \end{aligned} \quad (12)$$

Here $N_u = \sum_{i=1}^N g_{iu}^{(i)}$ is the cardinality of cluster u , and $\bar{\mathbf{X}}_u = \frac{1}{N_u} \sum_{i=1}^N g_{iu}^{(i)} \mathbf{X}_i$ is the cluster mean. Equality holds as the cross-product equals zero. Note also that (12) shows $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$ to be decomposable into two parts: the first term gives the deviations of the observations from their respective cluster means, while the last term is the modelling part which models the cluster means by a low-rank decomposition $\mathbf{C}_u \mathbf{D}_u'$.

It is evident from (12) that estimating the \mathbf{C}_u 's and \mathbf{D}_u 's only requires minimization of the modelling term. In order to do this, we first make the following definitions:

$$\begin{aligned} \mathbf{D}_* &= \begin{bmatrix} \sqrt{N_1} \mathbf{D}_1' \mathbf{J}_K^{(\delta_2)} & \sqrt{N_2} \mathbf{D}_2' \mathbf{J}_K^{(\delta_2)} & \cdots & \sqrt{N_U} \mathbf{D}_U' \mathbf{J}_K^{(\delta_2)} \end{bmatrix}'; \\ \mathbf{C}_* &= \begin{bmatrix} \sqrt{N_1} \mathbf{C}_1' \mathbf{J}_J^{(\delta_1)} & \sqrt{N_2} \mathbf{C}_2' \mathbf{J}_J^{(\delta_1)} & \cdots & \sqrt{N_U} \mathbf{C}_U' \mathbf{J}_J^{(\delta_1)} \end{bmatrix}'; \\ \bar{\mathbf{X}}_{(C)} &= \begin{bmatrix} \sqrt{N_1} \bar{\mathbf{X}}_1 \mathbf{J}_K^{(\delta_2)} & \sqrt{N_2} \bar{\mathbf{X}}_2 \mathbf{J}_K^{(\delta_2)} & \cdots & \sqrt{N_U} \bar{\mathbf{X}}_U \mathbf{J}_K^{(\delta_2)} \end{bmatrix}; \\ \bar{\mathbf{X}}_{(R)} &= \begin{bmatrix} \sqrt{N_1} \bar{\mathbf{X}}_1' \mathbf{J}_J^{(\delta_1)} & \sqrt{N_2} \bar{\mathbf{X}}_2' \mathbf{J}_J^{(\delta_1)} & \cdots & \sqrt{N_U} \bar{\mathbf{X}}_U' \mathbf{J}_J^{(\delta_1)} \end{bmatrix}'. \end{aligned} \quad (13)$$

Distinction must be made between the three cases where (a) $\mathbf{C}_1 \mathbf{D}_u'$, (b) $\mathbf{C}_u' \mathbf{D}_1$ or (c) $\mathbf{C}_u' \mathbf{D}_u$ applies. For a specified rank P , we now consider finding updates for \mathbf{C} and \mathbf{D} in each of these cases.

(a) Suppose that $\mathbf{C}_1 \mathbf{D}_u'$ applies. The last term in (12) can be rewritten as

$$\left\| \mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} - \mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}_*' \right\|^2. \quad (14)$$

Hence by the Eckart-Young theorem (Eckart and Young 1936), the best rank- P least-squares approximation of $\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)}$ is given by the truncated singular value decomposition (SVD) of $\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)}$. Consequently we can update \mathbf{C} and \mathbf{D} as follows:

$$\begin{aligned}\mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 &= \mathbf{U} \mathbf{\Gamma}^\alpha \mathbf{L} \\ \mathbf{D}_* &= \mathbf{V} \mathbf{\Gamma}^{1-\alpha} \mathbf{L}\end{aligned}\tag{15}$$

where $\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} = \mathbf{U} \mathbf{\Gamma} \mathbf{V}'$ is the appropriate SVD and the square matrix \mathbf{L} of dimension $\min\{J, UK\}$ is given by

$$\mathbf{L} = \begin{bmatrix} \mathbf{I}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.\tag{16}$$

Hence multiplication by \mathbf{L} sets all singular values except the first P equal to zero. The parameter $0 \leq \alpha \leq 1$ is typically taken to be 0.5, but can be set by the user to improve the interpretability of the graphical output (see Section 4). In this notation $\mathbf{\Gamma}^\alpha$ denotes the diagonal matrix where the diagonal contains the singular values to the power α . The degrees-of-freedom for the interactions are $P(J + UK - P - \delta_1 - \delta_2 U)$ in this case. See Appendix C for a short overview of how to derive this and the other degrees-of-freedom stated in the current section.

(b) In case $\mathbf{C}_u \mathbf{D}_1'$ applies, we can rewrite the last term in (12) as

$$\left\| \bar{\mathbf{X}}_{(R)} \mathbf{J}_K^{(\delta_2)} - \mathbf{C}_* \mathbf{D}_1' \mathbf{J}_K^{(\delta_2)} \right\|^2.\tag{17}$$

Analogously to (a), the update is based on the SVD $\bar{\mathbf{X}}_{(R)} \mathbf{J}_K^{(\delta_2)} = \mathbf{U} \mathbf{\Gamma} \mathbf{V}'$ and is

$$\begin{aligned}\mathbf{C}_* &= \mathbf{U} \mathbf{\Gamma}^\alpha \mathbf{L} \\ \mathbf{J}_K^{(\delta_2)} \mathbf{D}_1 &= \mathbf{V} \mathbf{\Gamma}^{1-\alpha} \mathbf{L}.\end{aligned}\tag{18}$$

Here \mathbf{L} is square with dimensions $\min\{UJ, K\}$. In this case, the degrees-of-freedom associated with the interactions are $P(UJ + K - P - \delta_1 U - \delta_2)$.

(c) Finally, when both \mathbf{C}_u and \mathbf{D}_u are cluster-specific, the update is based on the cluster-wise SVD's $\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)} = \mathbf{U}_u \mathbf{\Gamma}_u \mathbf{V}_u'$. We then have

$$\begin{aligned}\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u &= \mathbf{U}_u \mathbf{\Gamma}_u^\alpha \mathbf{L} \\ \mathbf{J}_K^{(\delta_2)} \mathbf{D}_u &= \mathbf{V}_u \mathbf{\Gamma}_u^{1-\alpha} \mathbf{L},\end{aligned}\tag{19}$$

with \mathbf{L} having dimensions $\min\{J, K\}$. The appropriate degrees-of-freedom are $UP(J + K - P - \delta_1 - \delta_2)$.

Hence the first step of our algorithm conveniently relies only on SVD's. Now in step (2) $\mathbf{G}^{(i)}$ is updated while regarding \mathbf{C} and \mathbf{D} as fixed. The updated $\mathbf{G}^{(i)}$ is constructed by simply assigning each i to the cluster with the closest mean, hence minimizing $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$ for each individual in a greedy manner. This entails setting $\mathbf{g}_{ik_i} = 1$ and zero elsewhere, where \mathbf{g}_i is the i th row of $\mathbf{G}^{(i)}$ and k_i is determined as

$$k_i = \arg \min_{u=1}^U \left\| \mathbf{J}_J^{(\delta_1)} (\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u') \mathbf{J}_K^{(\delta_2)} \right\|^2. \quad (20)$$

We prefer reporting the minimized value of $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$ after division by $\sum_{i=1}^N \left\| \mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)} \right\|^2$, since this standardized value lies within $[0, 1]$.

3.2. Fit Diagnostics

If $P \in \{1, 2, 3\}$ dimensions are used, biplots can be used to visualize the relationships between the J objects and K attributes for each of the clusters. Biplots generalize scatterplots of two variables to multiple variables (Gower and Hand 1996; Gower *et al.* 2011), and rely on low-rank inner product approximations.

Constructing biplots for the interactions here simply entails plotting the approximation of $\mathbf{J}_J^{(\delta_1)} \overline{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)}$ for each cluster. For example, for case (c) in (19) the object coordinates are given by $\mathbf{U}_u \mathbf{\Gamma}_u^\alpha \mathbf{L}$, while $\mathbf{V}_u \mathbf{\Gamma}_u^{1-\alpha} \mathbf{L}$ provides the coordinates for the attributes in P -dimensional space. The inner products between the pairs of rows in these matrices are rank- P approximations of the corresponding entries in $\mathbf{J}_J^{(\delta_1)} \overline{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)}$. Similar results are easily obtained for cases (a) and (b). We defer discussion of the interpretation of these biplots to Section 4, where empirical examples are examined.

It is possible to construct goodness-of-fit measures for the biplots used to visualize the interactions. These are based on the proportion of variation accounted for by the model. A fit value of one indicates perfect fit: the model captures all the variation in the data. In contrast, low fit values imply that a substantial amount of variation occurs in the subspace orthogonal to that identified by the model. An increase in fit can usually be achieved by increasing P , but even though overall fit is guaranteed to improve with increasing P , the fit of all individual rows and columns will not necessarily improve concurrently. In practice the choice $P = 2$ is the most convenient because the biplots can readily be displayed.

Measures can be defined for the overall fit, the J objects as well as for the K attributes. Again, we must distinguish between the three cases (a), (b) and (c). Case (c), where the applicable model is $\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}_u' \mathbf{J}_K^{(\delta_2)}$, is the simplest and therefore discussed first. The overall fit relies on the result

$$\left\| \mathbf{J}_J^{(\delta_1)} \overline{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)} \right\|^2 = \left\| \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}_u' \mathbf{J}_K^{(\delta_2)} \right\|^2 + \left\| \mathbf{J}_J^{(\delta_1)} (\overline{\mathbf{X}}_u - \mathbf{C}_u \mathbf{D}_u') \mathbf{J}_K^{(\delta_2)} \right\|^2, \quad (21)$$

which shows that the total sum-of-squares in cluster u can be decomposed into that explained by the model and the residual sum-of-squares. We therefore define the overall quality of fit within cluster u for P dimensions as

$$o_{\text{fit}} = \frac{\|\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K^{(\delta_2)}\|^2}{\|\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)}\|^2} = \frac{\text{tr } \mathbf{\Gamma}_u^2 \mathbf{L}}{\text{tr } \mathbf{\Gamma}_u^2}. \quad (22)$$

This is just the proportion of the variation in the cluster mean explained by the model. Here $\text{tr } \mathbf{A}$ denotes the trace of the square matrix \mathbf{A} , which is just the sum of its diagonal elements. Diagnostics for the J objects rely on the more general decomposition

$$\begin{aligned} & \left(\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)} \right) \left(\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)} \right)' \\ &= \left(\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K^{(\delta_2)} \right) \left(\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K^{(\delta_2)} \right)' \\ &+ \left(\mathbf{J}_J^{(\delta_1)} \left(\bar{\mathbf{X}}_u - \mathbf{C}_u \mathbf{D}'_u \right) \mathbf{J}_K^{(\delta_2)} \right) \left(\mathbf{J}_J^{(\delta_1)} \left(\bar{\mathbf{X}}_u - \mathbf{C}_u \mathbf{D}'_u \right) \mathbf{J}_K^{(\delta_2)} \right)'. \end{aligned} \quad (23)$$

Hence the total sum-of-squares for each of the objects can be decomposed orthogonally into the part explained by the model in the first term on the right-hand side of (23), and the residual sum-of-squares in the second term. The proportion of the variation explained by each of the rows, also known as sample predictivities (Gower *et al.* 2011), is therefore given by

$$\begin{aligned} \mathbf{r}_{\text{fit}} &= \left[\text{diag } \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K^{(\delta_2)} \mathbf{D}_u \mathbf{C}'_u \mathbf{J}_J^{(\delta_1)} \right] \left[\text{diag } \mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)} \bar{\mathbf{X}}_u' \mathbf{J}_J^{(\delta_1)} \right]^{-1} \mathbf{1}_J \\ &= \left[\text{diag } \mathbf{U}_u \mathbf{\Gamma}_u^{2\alpha} \mathbf{L} \mathbf{U}_u' \right] \left[\text{diag } \mathbf{U}_u \mathbf{\Gamma}_u^{2\alpha} \mathbf{U}_u' \right]^{-1} \mathbf{1}_J, \end{aligned} \quad (24)$$

with each element bounded on $[0, 1]$. In this context, $\text{diag } \mathbf{A}$ denotes the diagonal matrix constructed from the main diagonal of \mathbf{A} .

The column fit for case (c) can be defined analogously for each of the K attributes as

$$\begin{aligned} \mathbf{c}_{\text{fit}} &= \left[\text{diag } \mathbf{J}_K^{(\delta_2)} \mathbf{D}_u \mathbf{C}'_u \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}'_u \mathbf{J}_K^{(\delta_2)} \right] \left[\text{diag } \mathbf{J}_K^{(\delta_2)} \bar{\mathbf{X}}_u' \mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_u \mathbf{J}_K^{(\delta_2)} \right]^{-1} \mathbf{1}_K \\ &= \left[\text{diag } \mathbf{V}_u \mathbf{\Gamma}_u^{2\alpha} \mathbf{L} \mathbf{V}_u' \right] \left[\text{diag } \mathbf{V}_u \mathbf{\Gamma}_u^{2\alpha} \mathbf{V}_u' \right]^{-1} \mathbf{1}_K. \end{aligned} \quad (25)$$

These quantities are also known as axis predictivities (Gower *et al.* 2011). Diagnostics for cases (a) and (b) are deferred to [Appendix C](#).

The loss contribution for person i towards the interactions is defined as

$$L_{(i)}(i) = \sum_{u=1}^U g_{iu}^{(i)} L_{(i)}(i|u) = \sum_{u=1}^U g_{iu}^{(i)} \left\| \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}'_u \right) \mathbf{J}_K^{(\delta_2)} \right\|^2. \quad (26)$$

This gives an indication of badness-of-fit, and the sum over all persons gives the minimized value of $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$. These loss contributions account for possible differences in origin, scale and/or rotation between a person's interactions and the modelled cluster mean $\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}_u' \mathbf{J}_K^{(\delta_2)}$. A more informative manner of presenting these loss contributions may be as percentage contributions to $L_{(i)}(\mathbf{G}^{(i)}, \mathbf{C}, \mathbf{D})$.

An alternative measure of person fit which is bounded on $[-1, 1]$ is given by

$$p_{\text{fit}}(i) = \sum_{u=1}^U g_{iu}^{(i)} \frac{\text{tr} \mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)} \mathbf{D}_u \mathbf{C}_u'}{\left\| \mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)} \right\| \left\| \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}_u' \mathbf{J}_K^{(\delta_2)} \right\|}. \quad (27)$$

This only takes into account differences in rotation and origin, and high values indicate good fit whilst negative values indicate poor fit. When the origins coincide, the quantity (27) can be interpreted as a product-moment correlation coefficient between $\text{Vec} \mathbf{J}_J^{(\delta_1)} \mathbf{X}_i \mathbf{J}_K^{(\delta_2)}$ and $\text{Vec} \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{D}_u' \mathbf{J}_K^{(\delta_2)}$. The notation $\text{Vec} \mathbf{A}$ denotes the vector formed by concatenating the columns of a matrix \mathbf{A} into a single vector.

3.3. Biplot Interpretability

When neither the row nor column configurations are fixed across biplots, as is the case in (c), it can aid interpretation to rotate the configurations so that the axes lie more or less in the same direction. For any orthogonal matrix \mathbf{Q}_u , it holds for the inner product matrices that $\mathbf{C}_u \mathbf{D}_u' = (\mathbf{C}_u \mathbf{Q}_u) (\mathbf{D}_u \mathbf{Q}_u)'$, and hence these are invariant to orthogonal rotations. The problem of finding orthogonal matrices $\mathbf{Q}_u, u = 1, 2, \dots, U$, such that either the row or column configurations match each other as closely as possible is known as the generalized orthogonal Procrustes problem (Gower 1975; Gower and Dijksterhuis 2004).

Supposing without loss of generality that we use the $\mathbf{J}_J^{(\delta_1)} \mathbf{C}_u$ as axes in the biplots, a typical loss function for this problem is

$$\begin{aligned} L(\mathbf{Q}_1, \dots, \mathbf{Q}_U) &= \sum_{u < v}^U \left\| \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{Q}_u - \mathbf{J}_J^{(\delta_1)} \mathbf{C}_v \mathbf{Q}_v \right\|^2 \\ &= U \sum_{u=1}^U \left\| \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{Q}_u - \mathbf{H} \right\|^2 \end{aligned} \quad (28)$$

where $\mathbf{H} = U^{-1} \sum_{u=1}^U \mathbf{J}_J^{(\delta_1)} \mathbf{C}_u \mathbf{Q}_u$. A solution for (28) can be obtained through an alternating least-squares (ALS) algorithm; see Gower and Dijksterhuis (2004) for more details.

We note that two types of scalings can be used to make the biplot displays more attractive, namely α and λ scaling. First, since our choice of α does not change the inner product approximations (15), (18) and (19), we are free to choose it such that the resulting biplots are easy to interpret. In our software implementation we use as a heuristic method the value

of α which maximizes the minimum Euclidean distance over all row and column points to the origin. Alternatively, the user can choose any other quantile of these distances, such as the median, or specify the desired value of α explicitly.

Second, note that for matrices \mathbf{A} and \mathbf{B} it holds that $\mathbf{AB}' = (\lambda\mathbf{A})(\mathbf{B}'/\lambda)$, so that λ can also be freely chosen. Following Gower *et al.* (2011), we choose λ such that the average squared Euclidean distances from the two sets of points represented by the rows of the matrices in (15), (18) and (19) to the origin are equal. For case (c) in (19) for example, this amounts to choosing

$$\lambda = \sqrt[4]{\frac{J\|\mathbf{V}\mathbf{\Gamma}_u^{1-\alpha}\mathbf{L}\|^2}{K\|\mathbf{U}\mathbf{\Gamma}_u^\alpha\mathbf{L}\|^2}} = \sqrt[4]{\frac{J\text{tr}\mathbf{\Gamma}_u^{1-\alpha}\mathbf{L}}{K\text{tr}\mathbf{\Gamma}_u^\alpha\mathbf{L}}}. \quad (29)$$

3.4. Model Selection

Some questions remain, including how to select the number of segments for each of the four clustering problems, and what to do when missing values are present in the data. Here and in the next section we give a short description of viable options for dealing with these issues, which are by no means intended to be exhaustive. The applications of Section 4 will further illustrate some of the points raised here.

Selection of the number of clusters can be handled separately for each of the four clustering problems. Many criteria have been proposed in the literature, especially in the context of k -means and hierarchical clustering. Naturally, these can be utilized directly for the three k -means subproblems, and most can also be used for the interaction clustering. The simplest approach, and the one we use here for illustration, is probably the scree test (Cattell 1966). This method involves running the algorithm for several values of k and plotting the loss function against k . The user must then choose a value for k based on this so-called scree plot, such that the chosen k is close to an “elbow” in the plot. This indicates that adding additional groups to the analysis does not significantly increase how well the results describe the data.

A variation of the scree test is the convex hull (CHull) procedure of Ceulemans and Kiers (2006), which uses a measure of model complexity other than k , such as the degrees-of-freedom, for the scree plot. A convex hull is constructed from these points and a point close to the resulting elbow is selected, which represents a tradeoff between model complexity and goodness-of-fit. This approach has recently been used successfully in several component analysis contexts (e.g. Schepers, Ceulemans, and Van Mechelen 2008; Ceulemans and Kiers 2009; Ceulemans, Timmerman, and Kiers 2011; Lorenzo-Seva, Timmerman, and Kiers 2011).

Several other approaches have also been proposed in the literature, such as the Caliński-Harabasz criterion (Caliński and Harabasz 1974), the Krzanowski-Lai criterion (Krzanowski and Lai 1988), the silhouette plot (Kaufman and Rousseeuw 1990), the gap statistic (Tibshi-

rani, Walther, and Hastie 2001), the jump method (Sugar and James 2003) and bootstrapping (Dolnicar and Leisch 2010). Milligan and Cooper (1985), Hardy (1996) and Everitt *et al.* (2011) provide an assessment of some of these criteria and additional references. Experimenting with such alternative criteria are left for future research.

We note that it is also possible to select the number of clusters for all four subproblems at the same time. This may be easier to achieve with e.g. the CHull procedure than doing model selection separately. This approach can also be employed when the analyst wishes to treat the dimensionality of the interactions P as part of the model selection process.

4. Applications

We discuss two empirical examples: the first considers the evaluation of car manufacturers by a Dutch consumer panel, while the second considers the list of values data set (e.g. Van Rosmalen, van Herk, and Groenen 2010).

4.1. Car Manufacturers

This data set consists of 187 persons evaluating 10 car manufacturers on a set of 8 attributes, as collected via an online survey (Bijmolt and van de Velden 2012). The data is a subset of a larger set collected from panellists within the CentERpanel of Tilburg University in the Netherlands. The sample used was selected to be representative of the population of Dutch households. The car brands and items were presented in random order to each respondent, with the order of the items fixed per respondent. More information regarding the data collection can be found in Bijmolt and van de Velden (2012).

The manufacturers considered are ten international brands, namely Citroën, Fiat, Ford, Opel, Peugeot, Renault, Seat, Toyota, Volkswagen and Volvo. The task respondents were taxed with was to rate each of these brands on 8 different attributes using a 10-point rating scale. For 6 out of the 8 items, namely Affordability, Attractiveness, Safety, Sportiness, Reliability and Features, a score of 10 relates to the most desirable outcome. However, for the items Size and Operating Cost, a score of 10 reflects small cars and those with high operating costs respectively. Consequently, higher ratings on these items indicate more negative assessments.

We fit an LSBCLUST model with $\delta = (1, 1, 1, 1)$ so that the overall means, row means, columns means and interactions are estimated separately. Also, we use $P = 2$ dimensions and fix the coordinates of the 10 car brands across all interaction biplots (case (a)). The number of clusters to use for each of the four components must be determined. As mentioned in Section 3.4, this can be done separately for each subproblem. Here we fit LSBCLUST models for 1 to 15 clusters and inspect the resulting scree plots to select R, S, T and U . Based on these plots, we selected $R = 5$, $S = 8$, $T = 6$ and $U = 8$ clusters. We note that these choices

are subjective, should take into account the aims of the research and that alternative selection criteria can also be used. The number of random starts used for the interaction and k -means clustering were 100 and 1000 respectively.

The mean ratings for each of the five overall mean clusters are shown in Figure 2. Most interesting here are that 8 and 3 persons used very high and very low scores overall respectively (clusters O4 and O5). Inspecting the individuals belonging to Segment O5, we can identify individuals 50, 66 and 85, all of which respond with a rating of 1 to all items on all car brands, save for a single 2 assigned by person 66. These respondents obviously do not provide very interesting information in their answers, but since their corresponding row means, column means and interactions do not differ from the overall mean, we do not have to remove them from our analysis. These person are merely assigned to the row, column and interaction segments containing negligible effects (see below). In a similar vein, of the 8 persons in Segment O4, individuals 40, 47, 151 and 162 exclusively use rating category 10. The remaining persons in this cluster also almost exclusively use high ratings. LSBCLUST has therefore been able to identify the 11 persons in clusters O4 and O5 who provide very little sensible information.

Figure 3 displays the means of the eight car brand (row) clusters across all attributes. Effect sizes can be read off on the horizontal axis. Both Segments R7 and R8 consist of single individuals, namely persons 121 and 156 respectively. Person 121 used constant ratings for all car brands, except for Citroën (perhaps indicating that this person owns a Citroën). Specifically, 10's were assigned to Fiat and Toyota, 1's to Opel and Seat and 6's to the other brands. Segment R8 can be explained by only 10's being assigned to Fiat and Toyota and 1's to Peugeot and Volvo by person 156. Segment R6 contains 2.7% of respondents who scores Peugeot on average nearly 2.5 points below average, and rather prefers Fiat and Toyota, albeit with smaller effects (approximately 1.2 and 1.5 above average respectively). Segment R5 (8%) is not attracted to Opel with a negative mean effect of approximately 1.6 rating points, but does like Renault and Citroën somewhat. Persons in Segment R4 (9.6%) mainly dislike Volvo, scoring it 1.6 below average. The remaining segments can be interpreted similarly. Note that the largest cluster (R1, 48.7%) does not contain any large effects, indicating that these consumers do not have strong preferences for any of the brands across all attributes.

The attribute (column) mean effects for all six clusters are displayed in Figure 4. There are no singleton clusters here. Segment C6 (7%) consists of respondents who assign high scores to Safety and Reliability irrespective of the car brand, with effects of approximately 1.4 and 1.3 respectively. Operating Cost and Size are also assessed positively on average, taking into account that for these items lower scores are better. The effects are approximately -1.4 and -1.7 on these two items. Segment C5 (8%) is somewhat curious in that the large negative effect on Affordability (-1.6) dominates. These respondents give ratings below their average on Affordability, irrespective of the car brand. Affordability also has a large negative effect in

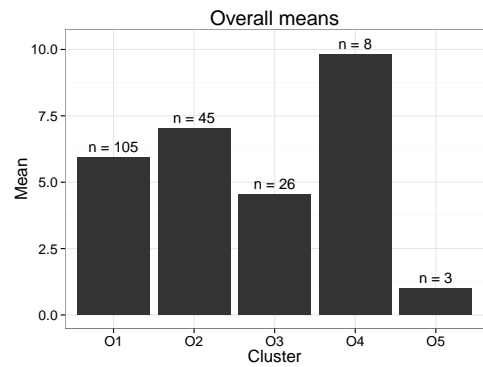


Figure 2: Clusterwise overall means detected for the cars data.

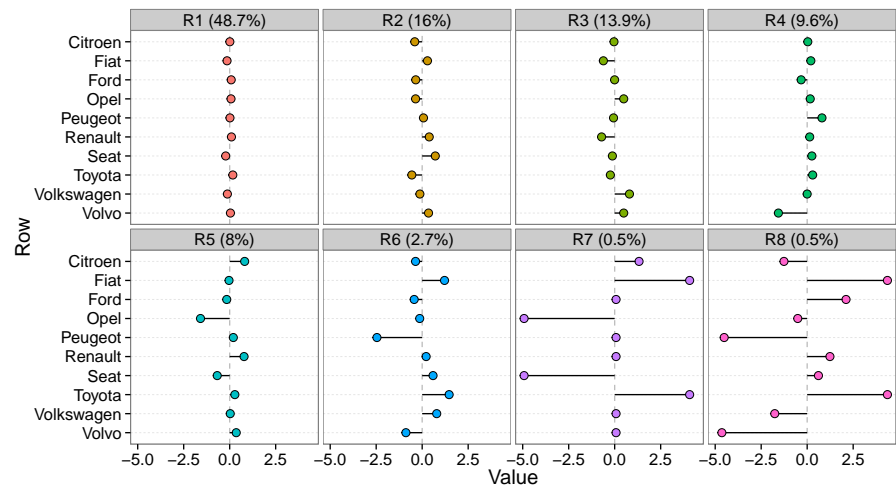


Figure 3: Car manufacturer (row) cluster means detected in the cars data. The size of the effects can be read off from the horizontal axis.

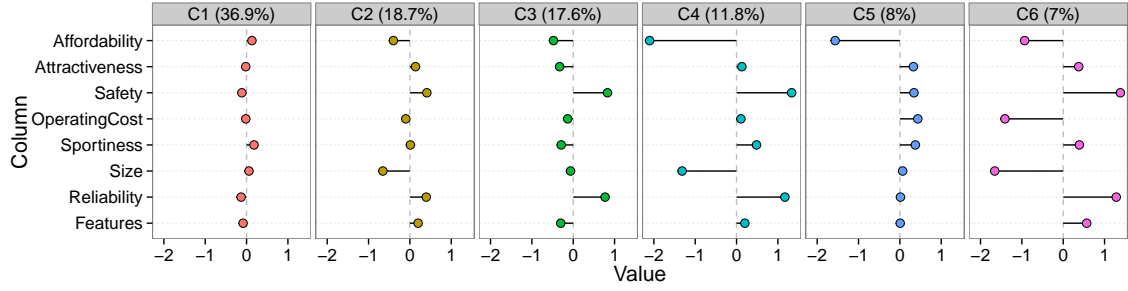


Figure 4: Attribute (column) cluster means detected in the cars data.

Segment C4 (11.8%), together with positive effects for Safety, Size and Reliability. The largest segment, C1 (36.9%), again has negligible effects.

The most interesting results can be found among the interactions, which is where respondents distinguish between different car manufacturers on the measured attributes. Figure 5 shows the biplots for the eight interaction segments. The car manufacturers are represented by points, and the attributes by arrows. The labels, points and arrows are shaded according to their goodness-of-fit, with well-fitting points being darker. The locations of the car brands are fixed across all biplots to make them easier to interpret. All car brands except Ford, with a fit of only 0.04, fit reasonably well – see Table 2. It is immediately apparent that the French manufacturers (Peugeot, Citroën and Renault) are judged to be similar, while the German brands Opel and Volkswagen are also located close together. The Swedish car manufacturer, Volvo, is somewhat isolated towards the right of the biplots. Fiat and Toyota are judged to be somewhat similar to the French and German brands respectively. Seat in turn are most similar to Toyota. By virtue of its low fit, Ford is hardly visible and lies near the origin.

The fit for the eight attributes vary per segment and is summarized in Table 3. Typically only a subset of items fit well in each segment, and only those with a fit larger than 0.5 are adorned with calibrated axes in Figure 5. For any manufacturer, the estimated cluster mean effects can be read off from the orthogonal projection of its representing point unto the biplot axes. For example, Volkswagen scores approximately 2 points above that predicted by the overall mean, row mean and column mean on Safety in Segment I8. Also, Volvo score about 3.3 rating points below the overall and marginal effects on Affordability in the same segment. The overall variance accounted for is 82.5% in two dimensions, with 63.2% and 19.2% attributed to dimension 1 and 2 respectively. Hence two dimensions are a reasonable choice.

Again, the largest Segment I1 (31.6%) represents very small effects, indicating that there is a significant proportion of respondents who did not discern between the car manufacturers based on the measured attributes. Predictably, this includes the 11 persons in Segments O4 and O5 discussed above. Segment I2 (16.6%) is characterized by large effects on Affordability, Size, Safety and Operating Cost. In particular, Affordability goes hand-in-hand with smaller cars

(i.e. higher scores on Size), while high Operating Cost is strongly associated with increased Safety. Volvo scores well on Safety and Size but badly on Operating Cost and Affordability. On the other end of the scale Seat, Fiat and Toyota are seen as producing more affordable, smaller cars which are less safe to drive.

The 15% of respondents in Segment I3 consider Affordability and Size by far the most important items. These attributes are also highly correlated: more affordable cars are the smaller ones. Safety and Reliability also correlates, and Attractiveness as well as Features, although to a lesser extent, also elicit relatively large effects. Seat scores well on Affordability but low on size, as does Fiat, but these cars are seen as unsafe and unreliable. Fiat is considered much more attractive than Seat, but less than the French cars and Volvo. Volvo are now seen as both safe and reliable, but also as the most attractive brand.

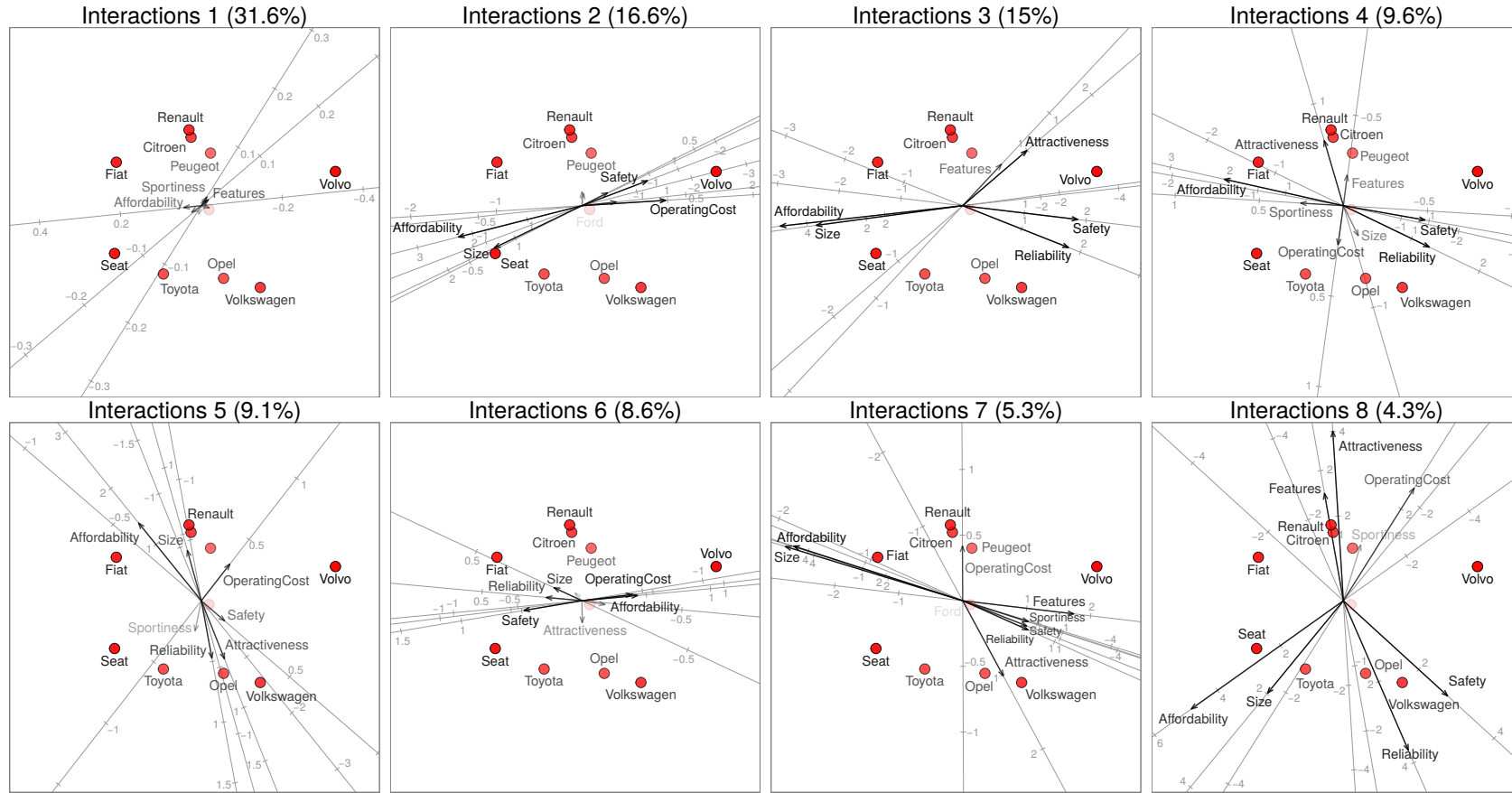


Figure 5: Biplots for the interaction clusters detected in the cars data. Each attribute is represented by a vector, and those which fit well also by calibrated axes. The car manufacturers are represented by points, and the orthogonal projections of these points onto the attribute axes give the estimated mean effects. The colours and labels are faded according to how well they fit into the display: solid colours fit well and transparent ones fit badly.

Segment I4 (9.6%) displays large effects on Affordability, Reliability, Safety and Attractiveness. As in Segment I3, cars cannot be affordable as well as safe and reliable at the same time. Attractiveness is more or less orthogonal to these polar opposites. People in this segment consider Fiat to be the most affordable, even more so than Seat. Volvo is followed by Volkswagen in Safety and Reliability but Volkswagen cannot compete with Volvo in terms of Attractiveness. The French cars are seen as the most attractive though.

In segment I5 (9.1%), Affordability again elicits the largest effects, although smaller than in some of the previous segments. In this case, Fiat, Renault and Citroën are seen as the most affordable cars while Volkswagen and Volvo again score badly in this category. Volkswagen, Opel and Toyota are seen as the most attractive. Segment I6 (8.6%) harbours even smaller effects, the largest being with respect to Safety, Affordability and Operating Cost. In the opinions of these respondents, better Affordability means higher Operating Costs and lower Safety. It is particularly interesting to see that these people do not believe that Volvo's are safe, but do think they are affordable. This contrasts with most other segments. Seat and Fiat are considered the safest cars in this case, but they are considered to be expensive too.

Size and Affordability are again highly correlated in segment I7 (5.3%), and responsible for some large effects. These attributes are strongly opposed to Features, Sportiness, Safety and Reliability. Fiat and Seat score best on Size and Affordability, but score low on Features. Volkswagen and Opel are seen as the most attractive cars. Finally, the smallest segment (Segment I8, 4.3%) considers Attractiveness and Features to be correlated, as well as Affordability and Size, and Reliability and Safety. The French cars score well on Attractiveness and Features, followed by Fiat and Volvo. Volkswagen are now seen as safer and more reliable than Volvo, with Opel also scoring well in these categories. Seat and Toyota score more than two rating points above the marginal effect on Affordability, but are considered to produce particularly small cars. Volvo and the French manufacturers produce much larger models. Volvo is the only manufacturer to combine attractive cars with high Safety and Reliability, but they are by far the most expensive.

4.2. List of Values

The list of values (LOV) is a well-known values system, developed in Kahle (1983), which asks respondents to indicate the importance of nine different values as a guiding principle in their lives. These items are: (1) a sense of belonging, (2) excitement, (3) warm relationships with others, (4) self-fulfillment, (5) being well-respected, (6) fun and enjoyment in life, (7) security, (8) self-respect and (9) a sense of accomplishment (Van Rosmalen *et al.* 2010). A labelled nine-point rating scale with endpoints 'very important' (category 1) to 'not important at all' (category 9) are used for all items. LOV data have been studied in single or multi-country contexts in various studies, including Beatty, Kahle, Homer, and Misra (1985); Grunert and Scherlorn (1990); Kamakura and Novak (1992); Brunsø, Scholderer, and Grunert (2004);

All segments	
Citroën	0.72
Fiat	0.87
Ford	0.04
Opel	0.63
Peugeot	0.53
Renault	0.82
Seat	0.92
Toyota	0.66
Volkswagen	0.75
Volvo	0.97

Table 2: Brand fit for the cars data across all clusters. Higher values indicate better fit, with a maximum of one and minimum of zero.

	Interaction segment							
	I1	I2	I3	I4	I5	I6	I7	I8
Affordability	0.52	0.96	0.94	0.93	0.72	0.83	0.95	0.80
Attractiveness	0.39	0.25	0.93	0.74	0.62	0.38	0.68	0.83
Safety	0.38	0.91	0.98	0.94	0.51	0.92	0.68	0.90
OperatingCost	0.46	0.94	0.08	0.64	0.66	0.90	0.53	0.57
Sportiness	0.51	0.22	0.03	0.56	0.28	0.35	0.73	0.26
Size	0.15	0.87	0.95	0.44	0.61	0.76	0.87	0.89
Reliability	0.07	0.62	0.97	0.98	0.69	0.71	0.83	0.84
Features	0.60	0.68	0.51	0.45	0.00	0.12	0.77	0.81

Table 3: Attribute fit for the cars data, for all eight interaction segments.

Chryssohoidis and Krystallis (2005); Lee, Soutar, and Louviere (2007); Sudbury and Simcock (2009).

The particular data set we analyse here originates from a commercial survey performed in 1996 (Van Herk 2000), and was also analysed by Van Rosmalen *et al.* (2010). As in many data sets collected using rating scales, there are some concern regarding differences in response style between respondents. Response styles are related to respondents exercising their freedom-of-choice with respect to how to the rating scale is used, irrespective of the item content (e.g. Baumgartner and Steenkamp 2001). For example, a person exhibiting an extreme response style may choose to use rating categories one and nine a majority of the time, while a person exhibiting midpoint scoring may favour mainly categories four, five and six. Crucially, response styles do not convey anything regarding the preference that a person may have for the items. Therefore response style effects should be accounted for separately in a model so that these effects are not confounded with the substantive information in the data.

Van Rosmalen *et al.* (2010) develop the so-called latent-class bilinear multinomial logit (LC-BML) model specifically to deal with situations where response styles are a concern. The

LC-BML model is a parametric finite mixture of multinomial logit models which models the response to all items jointly. Also, this model simultaneously segments respondents into two types of clusters, namely response style and substantive item segments. Similarly to LSBCLUST, the LC-BML model produces biplots describing the relationship between the values and the rating categories within each item segment. The coordinates of the rating categories are fixed across all biplots. The response styles are modelled as marginal effects for the rating categories.

A nonparametric equivalent of the LC-BML model can be formulated within the LSBCLUST framework. The data array is constructed by transforming each observation into an indicator matrix, with the rows representing the respective rating categories and the columns the value items. Each column contains a single one indicating which rating was used to answer that item. In effect we therefore consider the rating scale as one of the modes in our three-way data set. Choosing $\delta = (0, 1, 0, 0)$ fits a model containing only row or response style effects and interactions. Additionally, we use the option (a) $\mathbf{C}_1 \mathbf{D}'_u$ as a model for the interactions so that the coordinates for the rating categories are fixed across biplots. The resulting model is, except for the inclusion of demographic variables in the LC-BML model, equivalent to the LC-BML model. It has the distinct advantage of being much faster to compute, as least-squares estimation and crisp clustering are used instead of maximum likelihood and finite mixture models.

Our particular sample consists of 4514 respondents from five European countries (Great Britain, France, Germany, Italy and Spain). The demographic variables Education (Low or High) and Age (0–25, 25–39, 40–54 or 55+ years) are also available. There are 344 missing values in the data, spread among 141 respondents. In order to emulate the results in [Van Rosmalen *et al.* \(2010\)](#), we follow their convention of adding a specific rating category labelled ‘NA’ for the missing value. Also, all respondents who did not answer any item or who answered all items with the same rating were removed from the sample, as well as those who did not disclose either or both of the demographics Education and Age. [Van Rosmalen *et al.* \(2010\)](#) report that this resulted in a deletion of 8% of the original respondents.

For comparability, we use the same number of clusters as [Van Rosmalen *et al.* \(2010\)](#): 11 row and 5 interaction segments, with two-dimensional biplots. These were selected for the LC-BML model by using the Bayesian Information Criterion (BIC; [Schwarz 1978](#)). The number of starts used in LSBCLUST for the row and interaction effects were 1000 and 100 respectively. The clusters for the row effects (response styles) are summarized in Figure 6. For comparison the response styles detected by LC-BML are given in Figure 7 (see also Table 2 in [Van Rosmalen *et al.* \(2010\)](#)). These have been reordered such that they match more or less those detected by LSBCLUST.

There are definite similarities between the two sets of clusters. Quite clearly segments R1 (18.7%) and R1* (17.5%) correspond quite closely, indicating that slightly more than 1 in

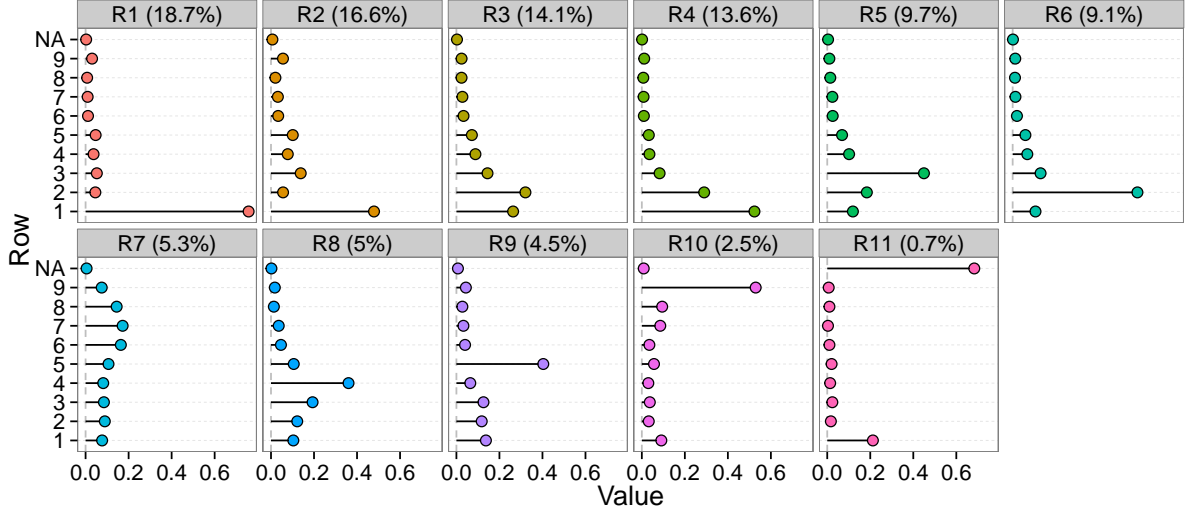


Figure 6: Response styles (row main effects) detected by LSBCLUST.

every 6 people tend to use rating category 1 (‘very important’) for 75% of their answers. This is an extreme form of an acquiescent response style, which favours positive responding (Baumgartner and Steenkamp 2001). Segments R2 (16.6%) and R2* (16.7%) agree closely and indicate that a less severe acquiescent response style is used by approximately 1 in every 6 persons. In addition, clusters R3 (14.1%) and R4 (13.6%) are also versions of acquiescence, but with less focus on category 1 only. We note that the similarity between R3 and R3* (24.2%) are not perfect, but R4 and R4* does correspond well. Indeed R3* was the largest segment detected by the LC-BML model. Among the remaining segments, there are also strong agreement between the LSBCLUST and LC-BML results, except for R8 (5%) and R9 (4.5%). These segments can be interpreted as imperfect forms of midpoint responding. We note that LSBCLUST tend to identify more balanced response styles while the LC-BML model identifies styles which give large points masses to specific ratings.

The interaction segments detected by LSBCLUST are shown as biplots in Figure 8. The overall fit is very high at 99.1% for two dimensions, with 88.5% attributed to the first dimension alone. The goodness-of-fit for the rating categories and value items are given in Table 4 and 5. All rating categories except the ‘NA’ category fit very well; likewise the fit for the items are very high in almost all cases.

Looking at the rating categories, the variation on the first dimension is associated almost exclusively with effects on rating category 1. Categories 1 and 2 are involved in the largest effects, but these categories are almost orthogonal. This indicates that a large effect on category 1 does not necessarily translate to a similarly large effect on category 2. The second dimension is associated with effects on category 2 and to a lesser extent categories 4 through 9. The directions of the axes for these latter categories are very similar, indicating that when a respondent uses one of these categories other persons in the same segment are likely to

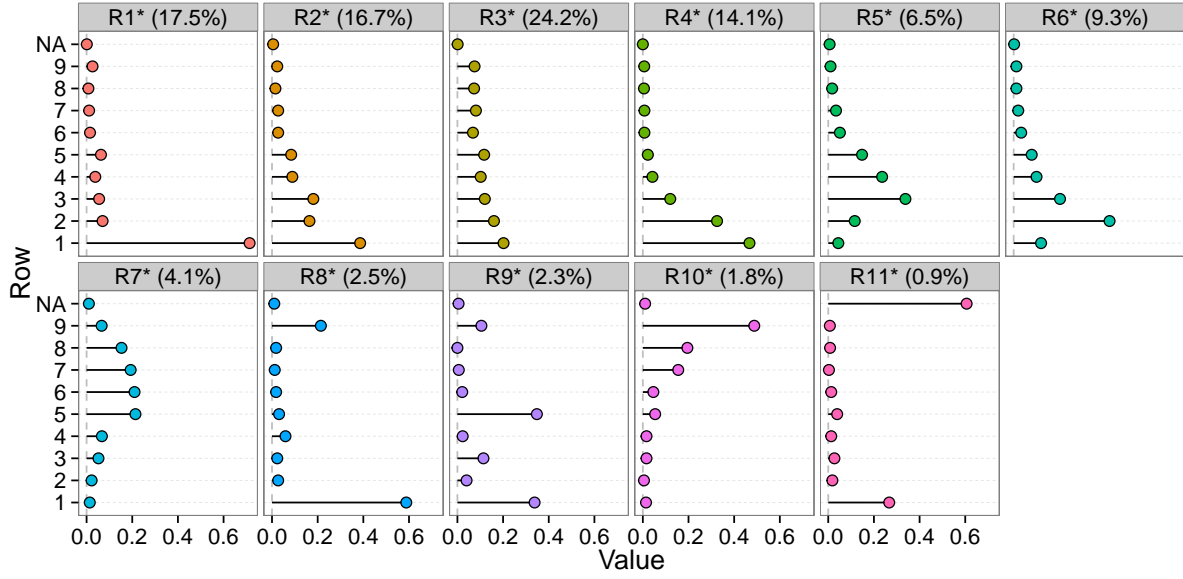


Figure 7: Response styles detected by the LC-BML model (see Table 2 in [Van Rosmalen et al. 2010](#)). The order of the segments have been adjusted so that it matches those of Figure 6 as closely as possible.

use one of these ratings too. This suggests that we should perhaps consider collapsing these categories into a single one. Effects on the ‘NA’ category is very small – being very close to the origin it is not labelled in the biplots.

Segment I1 (31.1%) contains respondents who attach most importance to a sense of belonging, warm relationships and self-respect, and find excitement the least important value. The effects in this segment are not particularly large, indicating that these persons do not consider the values to differ very much in importance. Segment I2 (20.8%) boasts much larger effects, with self-respect now considered the most important value followed by security and fun and enjoyment. The least important values in this segment are excitement, being respected and self-fulfilment. In I3 (20.3%), self-fulfilment and self-respect are most valued while belonging and excitement are not considered important. Segment I4 (15.4%) is characterized by the great importance attached to being respected and self-respect, while a sense of belonging, excitement and self-fulfilment are not considered important. Finally, in I5 (12.4%) the most importance is attached to fun and enjoyment, with self-respect also considered important. Least important in this segment is a sense of belonging and excitement.

Drawing comparisons between our segments and those detected by LC-BML in [Van Rosmalen et al. \(2010\)](#) proves to be a difficult task. This is not entirely unexpected as there are substantial differences between the methods. The most important difference is that LSBCLUST models the interactions directly whilst the LC-BML models the response probabilities. Hence the resulting sets of biplots are on different scales.

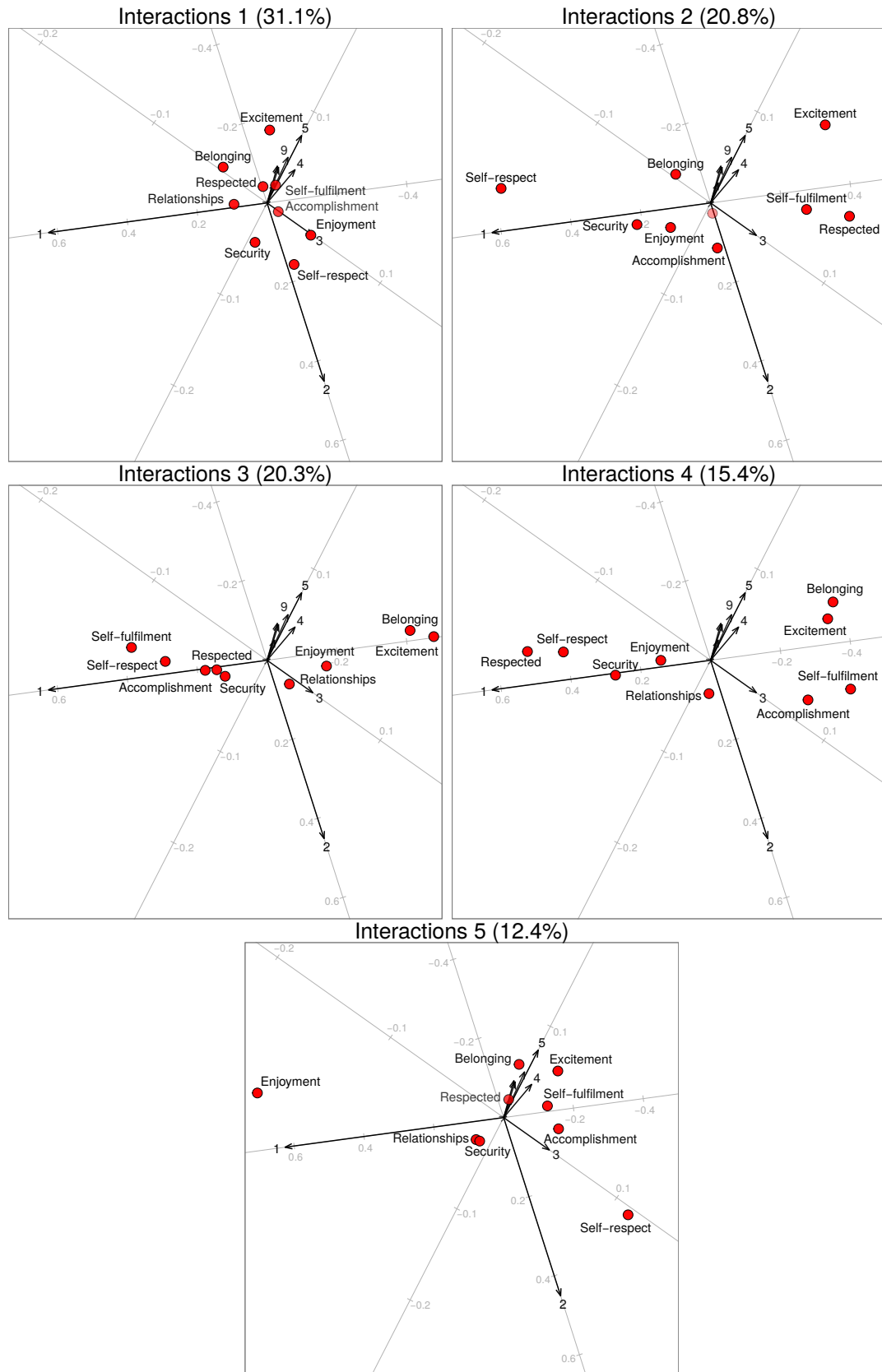


Figure 8: Biplots for the interaction segments detected in the LOV data. Rating categories are indicated as vectors and biplot axes, while the LOV items are indicated by labelled points. Points, lines and labels are shaded according to their goodness-of-fit (see Tables 4 and 5).

All segments	
1	1.00
2	1.00
3	0.90
4	0.94
5	0.98
6	0.91
7	0.92
8	0.85
9	0.86
NA	0.60

Table 4: Rating category fit for the LOV data across all interaction segments.

	Interaction segment				
	I1	I2	I3	I4	I5
Belonging	0.97	0.98	1.00	0.98	0.94
Excitement	0.97	0.99	1.00	1.00	0.98
Relationships	0.92	0.35	0.96	0.97	0.96
Self-fulfilment	0.80	0.99	1.00	1.00	0.95
Respected	0.91	0.99	0.98	1.00	0.73
Enjoyment	0.99	0.98	0.95	0.99	1.00
Security	0.97	1.00	0.99	1.00	0.96
Self-respect	0.95	1.00	1.00	1.00	0.98
Accomplishment	0.73	0.99	1.00	0.99	0.96

Table 5: Item fit for the LOV data across all interaction segments.

5. Conclusions

LSBCLUST is a modelling framework for three-way data, where one of the three ways is clustered over whilst the corresponding matrix slices are approximated by low-rank decompositions. The philosophy is to alternate between finding clusters of similar two-way matrices and doing low-rank matrix decompositions within each cluster. The clustering is done simultaneously with respect to up to four different aspects of these matrix slices, namely the overall mean responses, the row means, the column means, and the row-column interactions. These are the four elements of the biadditive (or bilinear) model used to approximate each of the matrix slices. Which of these terms are included in the model depends on the choice of identifiability constraints, as elegantly parametrized by δ . We show that in eight out of nine unique choices for δ , the combination of the bilinear model and least-squares loss allows the

four clustering problems to be treated independently. This important property greatly simplifies the complexity of the clustering problem, which also has positive implications for model selection and the interpretation of the results. The low-rank decompositions of the interaction cluster means lead to readily interpretable biplots which greatly aid in the interpretation of the results.

We argue that LSBCLUST is a useful and natural alternative to more traditional three-way matrix decomposition methods such as PARAFAC/CANDECOMP and TUCKALS3. Whereas these methods can be difficult to interpret, hard to fit and require careful study to understand properly, LSBCLUST uses a combination of well-known statistical methods in k -means cluster analysis, low-rank decompositions of two-way matrices and biplots. Since least-squares loss functions are used, the problems can be treated very efficiently in software. Such software implementing LSBCLUST has been developed in the form of an eponymous R (R Core Team 2014) package. The package, **lsbclust** (Schoonees 2015), is available for download from the Comprehensive R Archive Network (CRAN, <http://cran.r-project.org>).

There are some points that require further research. The treatment of missing values have not been discussed, apart from adding a dedicated catch-all missing category in the LOV example (Section 4.2). More refined techniques can be employed, such as iterative majorization (Kiers 1997), and should be investigated in the future. In terms of model selection, a wide variety of alternatives to the scree test can and should be investigated. There are a number of promising methods out there, including using multiple criteria and taking a vote to determine the most attractive choice. We note that the rank of the low-rank decomposition can also be considered as a model selection step. Furthermore, it would be possible to add case weights to the methodology. An advantage of case weights is that it allows a mechanism for implementing the bootstrap (e.g. Efron and Tibshirani 1994) to assess the variability of any given solution.

Finally, we have compared the results of LSBCLUST when $\delta = (0, 1, 0, 0)$ and the LC-BML model of Van Rosmalen *et al.* (2010) in the LOV example (Section 4.2). This required reformulating a two-way data set of categorical variables as three-way data by introducing dummy variables for each of the rating categories as the third way. Interestingly, there were obvious similarities between the response styles detected by both methods, but not so much among the interaction segments. The differences in the interaction clusters found can most likely be attributed to some important differences between the two methods: crisp clustering (LSBCLUST) versus probabilistic clustering (LC-BML), least-squares approximation of matrices (LSBCLUST) compared to generalized linear modelling of response probabilities (LC-BML), and the inclusion of covariates in the LC-BML. A definite advantage of LSBCLUST over LC-BML is that the results can be computed in a fraction of the time needed to run the Expectation-Maximization algorithm needed for the latter.

References

- Baumgartner H, Steenkamp JBEM (2001). "Response styles in marketing research: A cross-national investigation." *Journal of Marketing Research*, **38**(2), 143–156.
- Beatty SE, Kahle LR, Homer P, Misra S (1985). "Alternative measurement approaches to consumer values: the list of values and the Rokeach value survey." *Psychology & Marketing*, **2**(3), 181–200.
- Bijmolt TH, van de Velden M (2012). "Multiattribute perceptual mapping with idiosyncratic brand and attribute sets." *Marketing Letters*, **23**(3), 585–601.
- Brunsø K, Scholderer J, Grunert KG (2004). "Closing the gap between values and behavior: A means-end theory of lifestyle." *Journal of Business Research*, **57**(6), 665–670.
- Caliński T, Harabasz J (1974). "A dendrite method for cluster analysis." *Communications in Statistics - Theory and Methods*, **3**(1), 1–27.
- Carroll JD, Chang JJ (1970). "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition." *Psychometrika*, **35**(3), 283–319.
- Cattell RB (1966). "The scree test for the number of factors." *Multivariate Behavioral Research*, **1**(2), 245–276.
- Ceulemans E, Kiers HA (2006). "Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method." *British Journal of Mathematical and Statistical Psychology*, **59**(1), 133–150.
- Ceulemans E, Kiers HA (2009). "Discriminating between strong and weak structures in three-mode principal component analysis." *British Journal of Mathematical and Statistical Psychology*, **62**(3), 601–620.
- Ceulemans E, Timmerman ME, Kiers HA (2011). "The CHull procedure for selecting among multilevel component solutions." *Chemometrics and Intelligent Laboratory Systems*, **106**(1), 12–20.
- Chrysosoidis GM, Krystallis A (2005). "Organic consumers' personal values research: Testing and validating the list of values (LOV) scale and implementing a value-based segmentation task." *Food Quality and Preference*, **16**(7), 585–599.
- De Leeuw J (1994). "Block-relaxation algorithms in statistics." In *Information Systems and Data Analysis*, pp. 308–324. Springer.

- Denis JB, Gower JC (1994). "Asymptotic covariances for the parameters of biadditive models." *Utilitas Mathematica*, **46**, 193–205.
- Dolnicar S, Leisch F (2010). "Evaluation of structure and reproducibility of cluster solutions using the bootstrap." *Marketing Letters*, **21**(1), 83–101.
- Eckart C, Young G (1936). "The approximation of one matrix by another of lower rank." *Psychometrika*, **1**(3), 211–218.
- Efron B, Tibshirani RJ (1994). *An introduction to the bootstrap*, volume 57. CRC press.
- Everitt B, Landau S, Leese M, Stahl D (2011). "Cluster analysis." *John Wiley & Sons, Chichester*.
- Faber NKM, Bro R, Hopke PK (2003). "Recent developments in CANDECOMP/PARAFAC algorithms: a critical review." *Chemometrics and Intelligent Laboratory Systems*, **65**(1), 119–137.
- Gower JC (1975). "Generalized Procrustes analysis." *Psychometrika*, **40**(1), 33–51.
- Gower JC, Dijksterhuis GB (2004). *Procrustes problems*, volume 3. Oxford University Press Oxford.
- Gower JC, Hand DJ (1996). *Biplots*, volume 54. CRC Press.
- Gower JC, Lubbe SG, Le Roux NJ (2011). *Understanding biplots*. John Wiley & Sons.
- Greenacre M (2007). *Correspondence analysis in practice*. CRC Press.
- Grunert SC, Scherlorn G (1990). "Consumer values in West Germany underlying dimensions and cross-cultural comparison with North America." *Journal of Business Research*, **20**(2), 97–107.
- Hardy A (1996). "On the number of clusters." *Computational Statistics & Data Analysis*, **23**(1), 83–96.
- Harshman RA (1970). "Foundations of the Parafac procedure: Models and conditions for an explanatory multimodal factor analysis."
- Kahle LR (1983). *Social values and social change: Adaptation to life in America*. Praeger.
- Kamakura WA, Novak TP (1992). "Value-system segmentation: Exploring the meaning of LOV." *Journal of Consumer Research*, pp. 119–132.
- Kaufman L, Rousseeuw PJ (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

- Kiers HA (1997). “Weighted least squares fitting using ordinary least squares algorithms.” *Psychometrika*, **62**(2), 251–266.
- Kiers HA (2000). “Some procedures for displaying results from three-way methods.” *Journal of Chemometrics*, **14**(3), 151–170.
- Kiers HA, Krijnen WP (1991). “An efficient algorithm for PARAFAC of three-way data with large numbers of observation units.” *Psychometrika*, **56**(1), 147–152.
- Krijnen WP, Dijkstra TK, Stegeman A (2008). “On the non-existence of optimal solutions and the occurrence of degeneracy in the Candecomp/Parafac model.” *Psychometrika*, **73**(3), 431–439.
- Kroonenberg PM, De Leeuw J (1980). “Principal component analysis of three-mode data by means of alternating least squares algorithms.” *Psychometrika*, **45**(1), 69–97.
- Kruskal JB (1977). “Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics.” *Linear Algebra and Its Applications*, **18**(2), 95–138.
- Krzanowski WJ, Lai Y (1988). “A criterion for determining the number of groups in a data set using sum-of-squares clustering.” *Biometrics*, pp. 23–34.
- Lee JA, Soutar GN, Louviere J (2007). “Measuring values using best-worst scaling: The LOV example.” *Psychology & Marketing*, **24**(12), 1043–1058.
- Lorenzo-Seva U, Timmerman ME, Kiers HA (2011). “The Hull method for selecting the number of common factors.” *Multivariate Behavioral Research*, **46**(2), 340–364.
- Milligan GW, Cooper MC (1985). “An examination of procedures for determining the number of clusters in a data set.” *Psychometrika*, **50**(2), 159–179.
- R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Schepers J, Ceulemans E, Van Mechelen I (2008). “Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria.” *Journal of Classification*, **25**(1), 67–85.
- Schoonees PC (2015). *lsbclust: Least-squares bilinear clustering for three-way data*. R package version 1.0.1, URL <http://CRAN.R-project.org/package=lsbclust>.
- Schwarz G (1978). “Estimating the dimension of a model.” *Annals of Statistics*, **6**, 461–464.
- Sudbury L, Simcock P (2009). “Understanding older consumers through cognitive age and the list of values: A UK-based perspective.” *Psychology & Marketing*, **26**(1), 22–38.

- Sugar CA, James GM (2003). “Finding the number of clusters in a dataset.” *Journal of the American Statistical Association*, **98**(463).
- ten Berge JM, Sidiropoulos ND (2002). “On uniqueness in CANDECOMP/PARAFAC.” *Psychometrika*, **67**(3), 399–409.
- Tibshirani R, Walther G, Hastie T (2001). “Estimating the number of clusters in a data set via the gap statistic.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 411–423.
- Van Herk H (2000). *Equivalence in a cross-national context: Methodological & empirical issues in marketing research*. Ph.D. thesis, Tilburg University.
- Van Rosmalen J, van Herk H, Groenen PJF (2010). “Identifying response styles: A latent-class bilinear multinomial logit model.” *Journal of Marketing Research*, **47**(1), 157–172.

Appendix A: Orthogonality

Here we discuss the orthogonality of the decomposition (9), and consequently also (10). For the decomposition (9) to be orthogonal, it must be shown that all six cross-products occurring among the terms in (9) are zero. We treat each of these cross-products in turn.

1. For the cross-product between the interaction term and the row term, it holds that

$$\begin{aligned}
 & \text{tr} \left(\mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right) \left(\delta_2 \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \mathbf{1}_K' \right)' \\
 &= \delta_2 \text{tr} \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \mathbf{1}_K \left(\frac{1}{K} \mathbf{1}_K' \mathbf{X}_i' - \mathbf{a}_s' \right) \mathbf{J}_J^{(\delta_4)} \\
 &= 0.
 \end{aligned} \tag{30}$$

The last equality follows since when $\delta_2 = 1$, $\mathbf{J}_K^{(\delta_2)} \mathbf{1}_K = \mathbf{J}_K \mathbf{1}_K = \mathbf{0}$. When $\delta_2 = 0$, the equality is trivial.

2. For the cross-product between the interaction term and the column term, the result is analogous to the above, except that now the equality $\mathbf{1}_J' \mathbf{J}_J = \mathbf{0}'$ is used.
3. For the cross-product between the interaction term and the term for the overall mean, we have that

$$\begin{aligned}
 & \text{tr} \left(\mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \right) \left((\delta_1 \delta_3 + \delta_2 \delta_4 - \delta_1 \delta_2) \left(\frac{1}{JK} \mathbf{1}_J' \mathbf{X}_i \mathbf{1}_K - e_r \right) \mathbf{1}_J \mathbf{1}_K' \right)' \\
 &= (\delta_1 \delta_3 + \delta_2 \delta_4 - \delta_1 \delta_2) \left(\mathbf{1}_J' \mathbf{J}_J^{(\delta_1)} \left(\mathbf{X}_i - \mathbf{C}_u \mathbf{D}_u' \right) \mathbf{J}_K^{(\delta_2)} \mathbf{1}_K \right) \left(\frac{1}{JK} \mathbf{1}_J' \mathbf{X}_i \mathbf{1}_K - e_r \right) \\
 &= 0.
 \end{aligned} \tag{31}$$

This cross-product equals zero whenever at least one of the following is true: $\delta_1 = 1$, $\delta_2 = 1$ or $\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2 = 0$. But whenever both $\delta_1 = \delta_2 = 0$, $\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2 = 0$ irrespective of δ_3 and δ_4 . Hence the cross-product always equals zero.

4. For the cross-product between the row and column terms, we have

$$\begin{aligned} & \text{tr} \left(\delta_1 \mathbf{1}_J \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i - \mathbf{b}'_t \right) \mathbf{J}_K^{(\delta_3)} \right) \left(\delta_2 \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \mathbf{1}'_K \right)' \\ &= \delta_1 \delta_2 \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i - \mathbf{b}'_t \right) \mathbf{J}_K^{(\delta_3)} \mathbf{1}_K \left(\frac{1}{K} \mathbf{1}'_K \mathbf{X}_i - \mathbf{a}'_s \right) \mathbf{J}_J^{(\delta_4)} \mathbf{1}_J \\ &= \begin{cases} \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - \mathbf{b}'_t \mathbf{1}_K \right) \left(\frac{1}{K} \mathbf{1}'_K \mathbf{X}_i \mathbf{1}_J - \mathbf{a}'_s \mathbf{1}_J \right) & \text{if } \boldsymbol{\delta} = (1, 1, 0, 0)' \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (32)$$

Deducing when the cross-product equals zero uses the same concepts as above, but when $\boldsymbol{\delta} = (1, 1, 0, 0)'$ none of these apply and the cross-product is not necessarily equal to zero.

5. The cross-product between the column term and the term for the overall mean also does not necessarily equal zero. Here we can derive the following:

$$\begin{aligned} & \text{tr} \left(\delta_1 \mathbf{1}_J \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i - \mathbf{b}'_t \right) \mathbf{J}_K^{(\delta_3)} \right) \left((\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - e_r \right) \mathbf{1}_J \mathbf{1}'_K \right)' \\ &= J \delta_1 (\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2) \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i - \mathbf{b}'_t \right) \mathbf{J}_K^{(\delta_3)} \mathbf{1}_K \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - e_r \right) \\ &= \begin{cases} -J \left(\frac{1}{J} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - \mathbf{b}'_t \mathbf{1}_K \right) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - e_r \right) & \text{if } \boldsymbol{\delta} = (1, 1, 0, 0)' \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (33)$$

The last line follows from the fact that the crossproduct equals zero if $\delta_1 = 0$, if $\delta_1 = 1$ and $\delta_3 = 1$, and when $\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2 = 0$. Hence consideration must be given to the four cases $\delta_1 = 1, \delta_2 \in \{0, 1\}, \delta_3 = 0, \delta_4 \in \{0, 1\}$. It is easy to see that $\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2 = \delta_2\delta_4 - \delta_2 = 0$ in all these cases except when $\boldsymbol{\delta} = (1, 1, 0, 0)'$.

6. Analogously to the above, consider the cross-product between the row term and the term for the overall mean:

$$\begin{aligned} & \text{tr} \left(\delta_2 \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \mathbf{1}'_K \right) \left((\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - e_r \right) \mathbf{1}_J \mathbf{1}'_K \right)' \\ &= K \delta_2 (\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2) \mathbf{1}'_J \mathbf{J}_J^{(\delta_4)} \left(\frac{1}{K} \mathbf{X}_i \mathbf{1}_K - \mathbf{a}_s \right) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - e_r \right) \\ &= \begin{cases} -K \left(\frac{1}{K} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - \mathbf{1}'_J \mathbf{a}_s \right) \left(\frac{1}{JK} \mathbf{1}'_J \mathbf{X}_i \mathbf{1}_K - e_r \right) & \text{if } \boldsymbol{\delta} = (1, 1, 0, 0)' \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (34)$$

The expression equals zero when $\delta_2 = 0$, when $\delta_2 = 1$ and $\delta_4 = 1$ or when $\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2 = 0$. Considering the cases $\delta_1 \in \{0, 1\}, \delta_2 = 1, \delta_3 \in \{0, 1\}, \delta_4 = 1$ then, it can be

seen that $\delta_1\delta_3 + \delta_2\delta_4 - \delta_1\delta_2 = \delta_1\delta_3 - \delta_1 = 0$ except when $\boldsymbol{\delta} = (1, 1, 0, 0)'$.

Consequently the decomposition in (9) and (10) is valid for all $\boldsymbol{\delta}$, except for $\boldsymbol{\delta} = (1, 1, 0, 0)'$.

Appendix B: Degrees-of-freedom

The degrees-of-freedom (DF) associated with each of the four subproblems in (10) are given in Table 5. These are split according to the two sets of parameters that must be estimated in each case: the cluster memberships and the cluster means, or low-rank approximations of these. In this Appendix we give a short overview of how to calculate these quantities.

The number of parameters affiliated with the cluster memberships are somewhat ambiguous to quantify. For each person and C clusters, there are arguably $C - 1$ free parameters to find since this is the number of dummy coded variables needed to encode the clustering. However, only one of these C parameters are nonzero, and equal to one, since we are doing crisp clustering. So using $N(C - 1)$ DF for the cluster memberships is a conservative approach. Recall also that the inclusion of certain types of clusters depends on the choice of $\boldsymbol{\delta}$, and consequently also the number of parameters associated with the clustering.

The DF for the estimation of the mean effects not only depends on $\boldsymbol{\delta}$, but also on the choice between cases (a), (b) and (c). This choice affects the SVD used and hence the length and orthogonality restrictions imposed within the SVD itself. We illustrate the calculation here for case (a) only and leave cases (b) and (c) to be verified by the reader. Firstly, there are P singular values to be estimated in (15). Secondly, the estimation of \mathbf{UL} require in total JP parameters, but these are subject to length, orthogonality and, depending on δ_1 , sum-to-zero constraints. Each column of \mathbf{UL} must be of length one, orthogonal to the other columns and possibly have a zero sum, giving $P + \binom{P}{2} + \delta_1 P$ restrictions. Finally, \mathbf{VL} has UKP parameters but these are subject to the same length and orthogonality restrictions on the P columns. However, δ_2 now optionally applies sum-to-zero constraints within each block of \mathbf{D}_* . Hence there are $P + \binom{P}{2} + \delta_2 UP$ restrictions here.

The DF are therefore given by the total number of parameters $P + JP + UKP$ in $\mathbf{\Gamma L}$, \mathbf{UL} and \mathbf{VL} , minus all these constraints.

Appendix C: Biplot Diagnostics

Here we note fit diagnostics for cases (a) and (b) – case (c) has already been treated in

	Cluster membership	Mean effects
Overall means	$\delta^* N(R - 1)$	$\delta^* R$
Rows	$\delta_2 N(S - 1)$	$\delta_2 S(J - \delta_4)$
Columns	$\delta_1 N(T - 1)$	$\delta_1 T(K - \delta_3)$
Interactions (a)	$N(U - 1)$	$P(J + UK - P - \delta_1 - \delta_2 U)$
Interactions (b)	$N(U - 1)$	$P(UJ + K - P - \delta_1 U - \delta_2)$
Interactions (c)	$N(U - 1)$	$UP(J + K - P - \delta_1 - \delta_2)$

Table 6: Degrees-of-freedom associated with each of the four subproblems, split into the cluster membership parameters and estimates of the mean effects. Note that $\delta^* = \delta_1 \delta_3 + \delta_2 \delta_4 - \delta_1 \delta_2$.

Section 3.2. For case (a), we again have an orthogonal decomposition, based on (15), namely

$$\begin{aligned} \left(\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} \right) \left(\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} \right)' &= \left(\mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}'_* \right) \left(\mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}'_* \right)' \\ &\quad + \left(\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} - \mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}'_* \right) \left(\mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} - \mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}'_* \right)'. \end{aligned} \quad (35)$$

The row fit can therefore be defined as

$$\mathbf{r}_{\text{fit}} = \left[\text{diag } \mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}'_* \mathbf{D}_* \mathbf{C}'_1 \mathbf{J}_J^{(\delta_1)} \right] \left[\text{diag } \mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} \bar{\mathbf{X}}_{(C)}' \mathbf{J}_J^{(\delta_1)} \right]^{-1} \mathbf{1}_J \quad (36)$$

and the column fit as

$$\mathbf{c}_{\text{fit}} = \left[\text{diag } \mathbf{D}_* \mathbf{C}'_1 \mathbf{J}_J^{(\delta_1)} \mathbf{C}_1 \mathbf{D}'_* \right] \left[\text{diag } \bar{\mathbf{X}}_{(C)}' \mathbf{J}_J^{(\delta_1)} \bar{\mathbf{X}}_{(C)} \right]^{-1} \mathbf{1}_{UK}. \quad (37)$$

For case (b), a similar decomposition follows from (18) and we have

$$\mathbf{r}_{\text{fit}} = \left[\text{diag } \mathbf{C}_* \mathbf{D}'_1 \mathbf{J}_K^{(\delta_2)} \mathbf{D}_1 \mathbf{C}'_* \right] \left[\text{diag } \bar{\mathbf{X}}_{(R)} \mathbf{J}_K^{(\delta_2)} \bar{\mathbf{X}}_{(R)}' \right]^{-1} \mathbf{1}_{UJ} \quad (38)$$

for the rows, and similarly

$$\mathbf{c}_{\text{fit}} = \left[\text{diag } \mathbf{J}_K^{(\delta_2)} \mathbf{D}_1 \mathbf{C}'_* \mathbf{C}_* \mathbf{D}_1 \mathbf{J}_K^{(\delta_2)} \right] \left[\text{diag } \mathbf{J}_K^{(\delta_2)} \bar{\mathbf{X}}_{(R)}' \bar{\mathbf{X}}_{(R)} \mathbf{J}_K^{(\delta_2)} \right]^{-1} \mathbf{1}_K \quad (39)$$

for the columns.

Affiliation:

Pieter C. Schoonees, Patrick J.F. Groenen, Michel van de Velden
 Econometric Institute
 Erasmus School of Economics

Erasmus University Rotterdam

Burgemeester Oudlaan 50, 3062 PA Rotterdam

E-mail: schoonees@gmail.com; groenen@ese.eur.nl; vandevelden@ese.eur.nl