Econometric Institute Report
EI 2013-10

# Constrained dual scaling for detecting response styles in categorical data

Pieter C. Schoonees[1] Michel van de Velden Patrick J.F. Groenen

Econometric Institute, Erasmus University Rotterdam, The Netherlands

March 11, 2013

**Abstract**

Dual scaling is a multivariate exploratory method equivalent to correspondence analysis when analysing contingency tables. However, for the analysis of rating data different proposals appear in the dual scaling and correspondence analysis literature. It is shown here that a peculiarity of the dual scaling method can be exploited to detect differences in response styles. Response styles occur when respondents use rating scales differently for reasons not related to the questions, often biasing results. A spline-based constrained version of dual scaling is devised which can detect the presence of four prominent types of response styles, and is extended to allow for multiple response styles. An alternating nonnegative least squares algorithm is devised for estimating the parameters. The new method is appraised both by simulation studies and an empirical application.[2]

Keywords: response style; dual scaling; correspondence analysis; splines; nonnegative least squares; $K$-means.

[1]schoonees@ese.eur.nl

[2]We thank SURFsara Computing and Networking Services (www.surfsara.nl) for their support in using the Lisa Compute Cluster.

# 1 Introduction

A major issue in questionnaire-based research is the presence of response styles. A response style, sometimes also known as response bias or scale usage heterogeneity, can be described as systematic bias due to a respondent's tendency to respond to survey items regardless of its content (Van Rosmalen, Van Herk, & Groenen, 2010). Paraphrasing, a response style is the manner in which a person uses a rating scale, an example being extreme response style where the respondent, *for no substantial reason,* prefers to use the endpoints of the Likert scale more often than the intermediate rating categories.

Response styles can invalidate statistical analyses since they are completely confounded with the substantial information contained in the data and hence biases results in non-trivial ways (Baumgartner & Steenkamp, 2001). Advanced methods, such as the latent-class multinomial logit model of Van Rosmalen et al. (2010), the multidimensional ordinal IRT model of De Jong and Steenkamp (2010), or the ordinal regression model with heterogeneous thresholds of Johnson (2003), have been developed to deal with the data analysis when response style contamination is relevant. None of these appear to have achieved much popularity in practice.

Existing models often require a substantial investment of resources for its implementation, estimation and/or interpretation. As an alternative, the method presented in this paper results in a data set cleaned of the effects of response styles so that any analyses appropriate for the continuous nature of this cleaned data can be conducted. Furthermore, this method has three additional purposes, namely to (i) determine whether different response styles are present in categorical data; (ii) identify the respondents associated with each response style; and to (iii) classify the identified response styles into four different types. Software which implements the method in the `R` software environment (R Core Team, 2012) are available from the first author.

The proposed method is a variant of dual scaling (DS) for rating data (Nishisato, 1980a), also referred to as successive categories data in the DS literature. DS is an exploratory multivariate method, akin to correspondence analysis or CA (e.g. Greenacre, 2007). In the special case of rating data, DS however differs from CA in a manner that implicitly caters for response styles by including parameters for the Likert scale categories in an innovative way. These parameters allow for the detection of frequent (or infrequent) usage of certain ratings since the optimal scores assigned by DS to these parameters depend on how often each rating occurs in the data. The new method builds on this aspect of DS by including monotone spline functions to model the response styles

and by allowing for multiple response styles through latent classes.

The next section focuses on a closer discussion of response styles. Section 3 introduces spline functions for modelling response styles, explains the new methodology and details an alternating least squares algorithm for solving an extended version of the dual scaling problem. A simulation study is conducted in Section 4 to assess the strengths and weaknesses of the method. Finally, an application (Section 5) is presented.

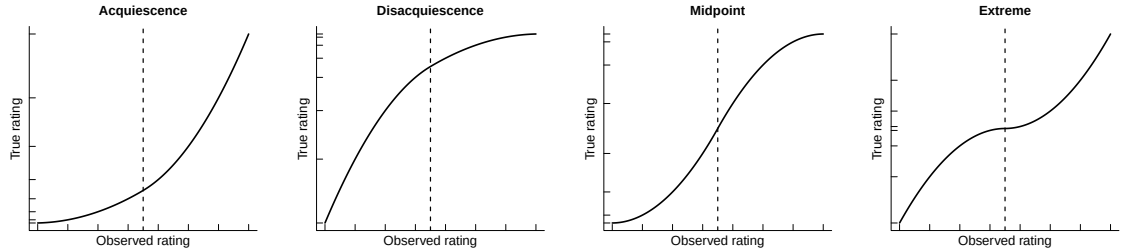
# 2 Overview of Response Styles

It is assumed that the process of formulating a response to a survey item requires the respondent to map a latent opinion, preference or some similar concept to a Likert scale. For example, the respondent may be asked how much she agrees with a certain statement using a scale with categories ranging from "1 – Totally Disagree" to "5 – Totally Agree." During the cognitive process of formulating the answer, the respondent first forms an opinion about the survey item and subsequently needs to decide how to transform or map this opinion to the presented rating scale (see for example Weijters & Baumgartner, 2012). The mathematical properties of this response mapping from the latent to the Likert scale determines whether a respondent exhibits a response style or not.

Specifically, a response style can be defined as a monotone nonlinear response mapping (Van de Velden, 2007). If this transformation is linear, no response style is present. Consequently, once a method is available to estimate response mappings the presence of response styles can be assessed by looking at the curvature properties of the estimated mappings. These steps are carried out in subsequent sections. In the case where Likert scales are used these transformations are step functions, but for the moment it is more intuitive to consider continuous transformations.

Four different response styles are considered here, as depicted in Figure 1. This figure shows different possible inverse mappings from the rating supplied by the respondent on the horizontal axis to the respondent's true latent opinion on the vertical axis. The inverse transformations are shown since these must be estimated from the observed data.

The different styles can be characterized by which parts of the latent opinion scale is stretched and which parts are shrunk. These are shown by the rug plots on the respected axes in Figure 1. The rug on the horizontal axis partitions the axis into intervals of equal length, with each interval receiving a rating on the Likert scale. Here a seven-point scale is employed. The rug on the vertical axis shows the effect that the response style transformation has on the intervals of equal length. Assuming that the respondent's true




(a) (b) (c) (d)

Figure 1: Examples of (inverse) response style functions mapping the true item content scale (vertical axis) into the observed measurement scale (horizontal axis).

latent opinion comes from a uniform distribution, these transformations characterize the following four response styles:

- *Acquiescence* (ARS) shrinks the lower part of the latent scale and stretches the upper part indicating that higher ratings are favoured (panel (a));
- *Disacquiescence* (DRS) in contrast favours lower ratings by stretching and shrinking the lower and upper parts of the latent scale respectively (panel (b));
- *Midpoint responding* (MRS) reflects a tendency to frequent the middle categories of the rating scale (panel (c)); and
- *Extreme responding* (ERS) in contrast means that the endpoints of the rating scale is used more often than the middle categories (panel (d)).

A critical concept is that the boundaries dividing the latent preference scale into the different rating categories, that is the tick marks on the vertical axes in Figure 1, determines which response style is present. If these boundaries are equally spaced, no response style is present. Any significant deviations however give cause to believe that a response style is present.

The methodology outlined in the next section makes use of these boundaries to provide an estimate of the response mappings of groups of individuals.

# 3 Methodology

Consider the situation where a set of $m$ objects or survey items are being rated on a $q$-point Likert scale, enumerated as 1 to $q$. Due to the ordinality this is sometimes known as successive categories data (Nishisato, 1980b, 1994). It is supposed that $n$ individuals are asked to rate the objects according to their preference. Objects may receive equal

ratings, and it is assumed that there exists a fixed but unknown preference structure for the set of objects, such as a population mean. Let $\mathbf{X}$ denote the $n \times m$ data matrix.

The next subsection discusses using dual scaling for analysing successive categories data in general, making use of the method's relationship with correspondence analysis. Monotone quadratic splines for modelling response styles are introduced in Section 3.2. Subsequently the dual scaling method is modified to utilise these splines together with latent classes to allow for multiple response styles. An alternating non-negative least squares algorithm is described for fitting the model in Section 3.4. Selecting the number of latent response style groups (Section 3.5) and creating a data set purged of the effects of response styles (Section 3.6) are also discussed.

## 3.1 Dual Scaling of Successive Categories Data

Dual scaling (DS) is a multivariate exploratory statistical technique which is equivalent to correspondence analysis (CA) when analysing contingency tables (Van de Velden, 2000). For such cases it is used to visualise departures from the independence assumption in the two-way contingency table in a low dimensional space, akin to principle components analysis (PCA) for continuous data (Nishisato, 1980a; Greenacre, 2007). However, for the successive categories data dealt with here there are important differences.

Both DS and CA deal with non-contingency table data by typically applying the standard procedure to a specific recoding of the data, designed to transform the data into a form that resembles a contingency table (Greenacre, 2007). This recoding requires the original data matrix $\mathbf{X}$ to be transformed before analysis, and for successive categories data in particular the recoding schemes differ in an important way. The usual CA method uses a doubling of columns (that is, adding an additional column to $\mathbf{X}$ for each object) to construct scales with "positive" and "negative" poles before applying ordinary CA (see Greenacre, 2007). However Nishisato (1980b) proposes the following alternative method. This involves augmenting rating scale category thresholds or boundaries to the data, which increases the number of columns from $m$ to $m + q - 1$, and then converting this to rank-orders. Although Nishisato's original DS formulation focuses on a so-called dominance matrix (see Nishisato, 1980a), it has been shown that DS applied to these rank-orders are equivalent to doubling the rows (instead of the columns) of the matrix of rankings before applying CA (Van de Velden, 2000; Torres & Greenacre, 2002).

The method is perhaps best illustrated by an example. Consider transforming the following data matrix $\mathbf{X}$, where three objects $A$, $B$ and $C$ are rated by $n = 4$ respondents

on a 5-point Likert scale (thus, $q = 5$). The first step requires augmenting 4 ($= q - 1$) columns to $\mathbf{X}$, one column for each of the boundaries between the pairs of adjacent ratings. Let the boundaries be called $b_1, \ldots, b_4$, where $b_1$ falls between ratings 1 and 2, and so forth up to $b_4$ which falls between categories 4 and 5. It suffices to assign scores midway between the rating categories to each boundary, to arrive at the augmented data matrix:

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} \mathtt{A} & \mathtt{B} & \mathtt{C} \end{array} \\ \begin{pmatrix} 4 & 3 & 1 \\ 2 & 2 & 5 \\ 3 & 2 & 2 \\ 1 & 5 & 4 \end{pmatrix} \end{array} \Rightarrow \mathbf{X}_{aug} = \begin{array}{c} \begin{array}{ccccccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{b}_1 & \mathtt{b}_2 & \mathtt{b}_3 & \mathtt{b}_4 \end{array} \\ \begin{pmatrix} 4 & 3 & 1 & 1.5 & 2.5 & 3.5 & 4.5 \\ 2 & 2 & 5 & 1.5 & 2.5 & 3.5 & 4.5 \\ 3 & 2 & 2 & 1.5 & 2.5 & 3.5 & 4.5 \\ 1 & 5 & 4 & 1.5 & 2.5 & 3.5 & 4.5 \end{pmatrix} \end{array}. \tag{1}$$

Secondly, each row is converted to rankings, starting with a lowest rank of 0 and a highest rank of 6 ($= m + q - 2$) in this case. For ties the total ranking assigned to the tied objects are distributed equally. This yields the following result for the example:

$$\mathbf{X}_{aug} \Rightarrow \mathbf{T} = \begin{array}{c} \begin{array}{ccccccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{b}_1 & \mathtt{b}_2 & \mathtt{b}_3 & \mathtt{b}_4 \end{array} \\ \begin{pmatrix} 5 & 3 & 0 & 1 & 2 & 4 & 6 \\ 1.5 & 1.5 & 6 & 0 & 3 & 4 & 5 \\ 4 & 1.5 & 1.5 & 0 & 3 & 5 & 6 \\ 0 & 6 & 4 & 1 & 2 & 3 & 5 \end{pmatrix} \end{array}. \tag{2}$$

Note that in general $\mathbf{T}$ has $n$ rows and $m + q - 1$ columns. DS also requires construction of the matrix $\mathbf{S}$ that would have resulted if $q$ was the lowest and not the highest rating of the Likert scale. This is easily achieved as

$$\mathbf{S} = (m + q - 2)\mathbf{1}\mathbf{1}^{'} - \mathbf{T}. \tag{3}$$

Using the CA formulation of DS of Van de Velden (2000), a row-doubled ratings matrix $\mathbf{F}_r : 2n \times (m + q - 1)$ is constructed as

$$\mathbf{F}_r = \begin{pmatrix} \mathbf{T} \\ \mathbf{S} \end{pmatrix}. \tag{4}$$

This matrix is subjected to CA, which assigns optimal scores in the vectors $\mathbf{a}$ and $\mathbf{b}$ to the rows and columns of $\mathbf{F}_r$ respectively (assuming a one-dimensional solution). This

is achieved by minimising a least squares criterion $L(\mathbf{a}, \mathbf{b})$ through the singular value decomposition (SVD) (Van de Velden, Groenen, & Poblome, 2009). In the present context $L$ is given by

$$L(\mathbf{a}, \mathbf{b}) = c\|\mathbf{F}_r - \frac{1}{2}(m + q - 2)(\mathbf{1}\mathbf{1}^{'} + \mathbf{a}\mathbf{b}^{'})\|^2 \tag{5}$$

where $c$ is a proportionality constant, $\mathbf{1}$ denotes vectors of ones of the appropriate lengths and $\frac{1}{2}(m + q - 2)\mathbf{1}\mathbf{1}^{'}$ centres the rankings in $\mathbf{F}_r$. For identifiability a constraint such as $\|\mathbf{a}\| = 1$ is imposed. The method is discussed in more detail in Section 3.3.

Note that an important consequence of the data recoding scheme is that the dual scaling procedure provides coordinates for the boundaries. The effect of the boundaries is to retain the information on how different the original ratings assigned to the objects were before the rankings were constructed. The coding scheme also imposes ordinality on the object and the boundary scores in $\mathbf{b}$ by constructing rankings.

The optimal scores assigned to the boundaries can be used to detect response styles since they estimate the thresholds of the response mapping of the group of respondents, as was discussed in relation to Figure 1. Intuitively optimal scores assigned to the boundaries work as follows. If a specific rating category $j$ is used very often, the boundaries $b_{j-1}$ and $b_j$ will often receive rankings which differ substantially since the category is often filled. Consequently, the optimal scores assigned will differ significantly, indicating that respondents use the category very often. The same reasoning illustrates that when rating $j$ is used very infrequently, the optimal scores for $b_{j-1}$ and $b_j$ will be very similar. Therefore, when a group of respondents have the same response mapping, the method will be able to tell which type that mapping is.

In Section 3.3 latent classes will be introduced for the boundary scores which allows for multiple response styles. First, however, using monotone quadratic splines with the dual scaling method is discussed.

## 3.2 Modelling Response Styles by Monotone Quadratic Splines

From Figure 1 it is evident that the four response styles considered can be completely described in terms of its curvature properties. By dividing the horizontal axes into two equal lower and upper parts, the four response styles are characterized by either concavity or convexity in the lower and upper parts of its domain. This is summarised in Table 1.

For inferential and response style classification purposes it will prove useful to parameterize the response style transformations considered here. The family of monotone quadratic splines with a single interior knot is ideal for this purpose as it combines two

| **Response style** | **Lower Curvature** | **Upper Curvature** |
|---|---|---|
| No Response Style | None | None |
| Acquiescence | Convex | Convex |
| Disacquiescence | Concave | Concave |
| Extreme Responding | Concave | Convex |
| Midpoint Responding | Convex | Concave |

Table 1: Curvature properties of the four response styles.

quadratic polynomial functions in the adjacent intervals of the domain, subject to continuity and differentiability restrictions at the interior knot. These splines are either concave, convex or linear in the lower and upper halves of the domain and therefore reproduce all the curves described in Figure 1 and Table 1.

The splines have three non-constant basis functions (the so-called $I$-spline basis) derived by appropriately integrating the basis functions of the $M$-spline basis (see Ramsay, 1988). A quadratic monotone spline with a single interior knot $t \in [L, U]$ and intercept $\mu$ is of the form

$$f(x) = \mu + \sum_{i=1}^{3} \alpha_i M_i(x \mid t). \tag{6}$$

In the proposed model $t = L + 0.5(U - L)$ is chosen to lie halfway between the lower and upper limits $L$ and $U$ respectively. Monotonicity requires that $\alpha_i \geq 0$ for $i = 1, 2, 3$. The basis functions $M_1, M_2$ and $M_3$ are constructed to ensure continuity and first-order differentiability at $t$, and their formulae are as follows (Ramsay, 1988):

$$\begin{aligned}
M_1(x \mid t) &= \begin{cases} \frac{2t(x-L)-(x^2-L^2)}{(t-L)^2}, & \text{if } L \leq x < t; \\ 1, & \text{if } t \leq x \leq U; \end{cases} \\
M_2(x \mid t) &= \begin{cases} \frac{(x-L)^2}{(t-L)(U-L)}, & \text{if } L \leq x < t; \\ \frac{t-L}{U-L} + \frac{2U(x-t)-(x^2-t^2)}{(U-t)(U-L)}, & \text{if } t \leq x \leq U; \end{cases} \\
M_3(x \mid t) &= \begin{cases} 0, & \text{if } L \leq x < t; \\ \frac{(x-t)^2}{(U-t)^2}, & \text{if } t \leq x < U; \end{cases}
\end{aligned} \tag{7}$$

The spline functions are built into the column scores $\mathbf{b}$ in (5) by using the $(q-1) \times 4$ design matrix $\mathbf{M}$ to collect the basis functions corresponding to $\mu, \alpha_1, \alpha_2$ and $\alpha_3$ respectively. The basis functions are evaluated at the midpoints between rating categories, for

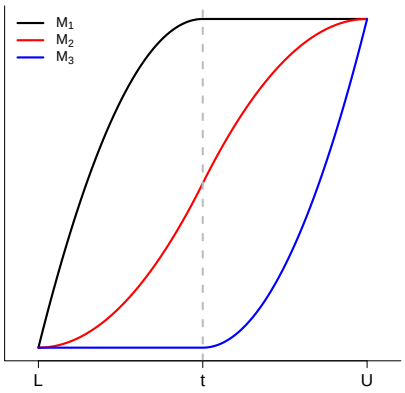



Figure 2: The three $I$-spline basis functions for quadratic monotone splines with a single interior knot $t$.

example at 1.5, 2.5 up to 6.5 for a 7-point Likert scale. Hence $\mathbf{b}$ can be written as

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{M}\boldsymbol{\alpha} \end{pmatrix} \tag{8}$$

with $\mathbf{b}_1$ the $m$-vector of unrestricted object scores and $\mathbf{b}_2$ the $(q-1)$-vector of spline-restricted boundary scores. The spline parameters are collected in $\boldsymbol{\alpha} = (\mu, \alpha_1, \alpha_2, \alpha_3)'$.

The basis functions $M_1, M_2$ and $M_3$ in (7), as depicted in Figure 2, are piecewise quadratic, with only two of them nonconstant in each of the intervals $[L,\, t)$ and $[t,\, U]$. This is convenient because it means the second derivative of $f$, and hence the curvature, depends only on two parameters in each interval. Rescaling without loss of generality so that $L = 0$ and $U = 1$, the curvature of $f$ (not necessarily defined at $t = 1/2$) is given by

$$\frac{d^2}{dx^2} f(x) = \begin{cases} -8\alpha_1 + 4\alpha_2, & \text{if } 0 \leq x < 1/2; \\ -4\alpha_2 + 8\alpha_3, & \text{if } 1/2 < x \leq 1; \end{cases} \tag{9}$$

The function $f(x)$ is either convex, concave or linear in a given interval depending on whether the second derivative (9) is positive, negative or zero respectively, which does not depend on $x$. In fact, assuming that $\alpha_1$ and $\alpha_3$ are larger than zero, the curvature can be measured solely in terms of the ratios $\alpha_2/\alpha_1$ and $\alpha_2/\alpha_3$, referred to henceforth as the curvature parameters. For example, the requirement for convexity in both the lower

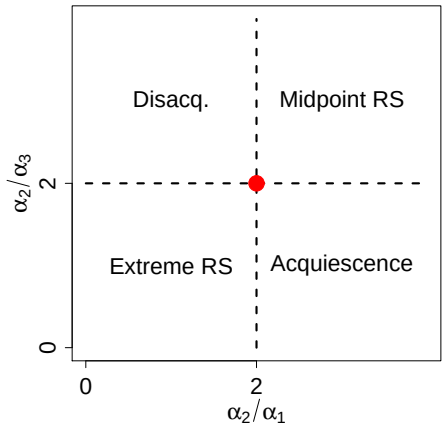



Figure 3: Classifying response styles graphically using the curvature properties of monotone quadratic splines.

and upper domain is

$$\frac{d^2}{dx^2}f(x) > 0 \Leftrightarrow \begin{cases} \frac{\alpha_2}{\alpha_1} > 2, & \text{if } L \leq x < t; \\ \frac{\alpha_2}{\alpha_3} < 2, & \text{if } t < x < U. \end{cases} \tag{10}$$

Problems can however occur with this analysis when either or both of $\alpha_1$ and $\alpha_3$ are zero. In such cases a continuity adjustment may need to be made.

It is possible to rewrite Table 1 wholly in terms of the curvature parameters, but more importantly using the curvature parameters it is possible to visualize the curvature of an estimated response style in a single plot. Figure 3 illustrates the situation by plotting $\alpha_2/\alpha_3$ against $\alpha_2/\alpha_1$, as well as incorporating the response style classification regions derived from Table 1. When both curvature parameters equal two, no response style is present. Due to the fact that both curvature parameters has the range $[0, \infty)$, a more symmetric plot is arrived at by using the base-2 logarithmic transform and centring – this is illustrated in Section 5.

## 3.3 Dual Scaling Method for Multiple Response Styles

To allow for multiple response styles, I see that suppose that each of the $n$ individuals belongs to one of $K$ response style groups, the exact membership being unknown. Let the $n \times K$ matrix $\mathbf{G}$ contain as columns the group indicator vectors $\{\mathbf{g}_k\}_{k=1}^K$, each indicating which individuals belong to that specific group. The column scores $\{\mathbf{b}_k\}_{k=1}^K$ are of the

same form as $\mathbf{b}$ in Equation (8), but are now group-specific by replacing $\mathbf{b}_2$ with $\mathbf{b}_{2k} = \mathbf{M}\boldsymbol{\alpha}_k$. This allows for the different groups to have different response mappings by letting the spline parameters $\boldsymbol{\alpha}_k = (\mu_k, \alpha_{1k}, \alpha_{2k}, \alpha_{3k})'$ vary between groups. The object scores $\mathbf{b}_1$ and the row scores $\mathbf{a}$ remain fixed across all response style groups.

The group membership $\mathbf{G}$ needs to be estimated, together with the $2n$-vector $\mathbf{a}$ of optimal scores for the individuals and the column score vectors $\mathbf{b}_k$ of length $(m+q-1)$ contained in the $K$ columns of $\mathbf{B}$. It is required for monotonicity that $\alpha_{ik} \geq 0$ for all $i$ and $k$. The loss function in Equation (5) must be adjusted to allow for the multiple response styles as well as for the spline restrictions. This constrained dual scaling method for the detection of response styles can be formulated as

$$\begin{aligned} &\min_{\mathbf{a},\mathbf{B},\mathbf{G}} L(\mathbf{a},\mathbf{B},\mathbf{G}) \\ \text{subject to } \mathbf{b}_k = &\begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_{2k} \end{pmatrix} \text{ and } \alpha_{ik} \geq 0,\ i = 1,2,3,\ k = 1,2,\ldots,K. \end{aligned} \tag{11}$$

The adjusted loss function (compare Eq. (5)) is

$$L(\mathbf{a},\mathbf{B},\mathbf{G}) = c\|\mathbf{F}_r - \frac{1}{2}(m+q-2)(\mathbf{1}\mathbf{1}' + \sum_{k=1}^{K} \mathbf{D}_{\mathbf{g}_k}\mathbf{a}\mathbf{b}_k')\|^2. \tag{12}$$

Again, $c$ is a proportionality constant, and the diagonal matrices $\mathbf{D}_{\mathbf{g}_k}$ are contructed as

$$\mathbf{D}_{\mathbf{g}_k} = \begin{pmatrix} \text{diag}(\mathbf{g}_k) & 0 \\ 0 & \text{diag}(\mathbf{g}_k) \end{pmatrix}. \tag{13}$$

Hence, through using the $\{\mathbf{D}_{\mathbf{g}_k}\}_{k=1}^{K}$ in (12), individuals are associated with the corresponding $\mathbf{b}_k$ for their group. Experience suggests that $L$ typically decreases as more groups are added (that is when $K$ increases). Therefore, when considering how the value of $L$ changes for different values of $K$ in a scree plot, it is convenient to standardise these values to the unit interval $[0,1]$.

An algorithm for minimising $L$ is discussed in the next section.

## 3.4 An Alternating Nonnegative Least Squares Algorithm

Solving the optimisation problem in (11) requires finding $\mathbf{a}, \mathbf{B}$ and $\mathbf{G}$ under the appropriate restrictions. The approach discussed here alternates over two steps:

1. The algorithm combines alternating least squares (ALS) and nonnegative least

squares (NNLS; Lawson & Hanson, 1974) to approximate the optimal $\mathbf{a}$ and $\mathbf{B}$ for a given group membership matrix $\mathbf{G}$. This involves fixing the value of $\mathbf{a}$, estimating the optimal $\mathbf{B}$ with NNLS, and then updating $\mathbf{a}$ by ordinary least squares (OLS) based on the estimate of $\mathbf{B}$. This ALS process is repeated for a given $\mathbf{G}$ until numerical convergence is observed.

2. For fixed $\mathbf{a}$ and $\mathbf{B}$, $\mathbf{G}$ is updated by a $K$-means type algorithm given the values determined for $\mathbf{a}$ and $\mathbf{B}$. This step simply allocates each individual sequentially to the group which minimises the loss function.

The algorithm alternates between steps one and two until the loss function $L$ changes by less than a small positive constant. Note that starting values for both $\mathbf{a}$ and $\mathbf{G}$ are required. Also, such block-relaxation algorithms may suffer from local minima, and therefore multiple random starts are required.

The optimisation process is described in more detail in Algorithm 1. The formulation is for a single starting configuration of $\mathbf{G}$, and needs to be repeated for multiple such configurations. Parameters that need to be specified include $n_a$, the number of (random) starts used for $\mathbf{a}$, the maximum number of iterations $\text{maxit}_a$ and $\text{maxit}_G$ for the ALS and $K$-means phases respectively, and also the numerical tolerances $\epsilon_1 > 0$ and $\epsilon_2 > 0$ for these two steps.

## 3.5 Selecting the Number of Response Style Groups

To select the number of groups $K$, a scree plot of the loss function for different values of $K$ can be used. The aim is to choose the smallest $K$ such that larger values do not significantly reduce the loss. This method was introduced by Cattell (1966) and has been widely adopted. The dual scaling method also separates individuals based on the shape of the response transformations and rating frequencies in the groups. This supplementary information can be helpful for choosing $K$ in cases where the scree plot is not conclusive. This will be illustrated in the empirical application of Section 5.

## 3.6 Purging Response Styles

Once the estimates $\hat{\mathbf{a}}, \hat{\mathbf{B}}$ and $\hat{\mathbf{G}}$ have been obtained, these can be used to create a version of the original data $\mathbf{X}$ which is purged of response styles. All that is needed is to use the splines estimated for each response style group to assign optimal scores to the rating scale. Then these scores are substituted in $\mathbf{X}$ by replacing every rating with the appropriate optimal score.

**Algorithm 1** Alternating Nonnegative Least Squares Algorithm

**set** $i = 0$, $h = 0$ and $n_a$, $\text{maxit}_a$, $\text{maxit}_G$, $\epsilon_1 > 0$ and $\epsilon_2 > 0$
**initialise** $\mathbf{G}_0$, set $\mathbf{F}_r^* = \mathbf{F}_r - \frac{1}{2}(m+q-2)\mathbf{1}\mathbf{1}'$
**while** $L_{h-1} - L_h > \epsilon_2$ and $h \leq \text{maxit}_G$ **do**
  **construct** $\mathbf{D}^h_{\mathbf{g}_k}$ from $\mathbf{G}_h$ according to Equation (13)
  **for all** $j = 1, 2, \ldots, n_a$ **do** (*iterate over different starts for* $\mathbf{a}$)
    **if** $i = 0$ and $h = 0$, generate a starting configuration $\mathbf{a}_j$ for $\mathbf{a}$
    **while** $L_{i-1,j} - L_{ij} > \epsilon_1$ and $i \leq \text{maxit}_a$ **do**
      **update** (*indices* $i$ *and* $h$ *are omitted for readability*)
        $w_{kj} \leftarrow (\mathbf{a}_j' \mathbf{D}^h_{\mathbf{g}_k} \mathbf{a}_j)^{-1/2}$ for all $k$
        $(\mathbf{v}_{1kj}, \mathbf{v}_{2kj})' \leftarrow 2w_{kj}^{-1}(\mathbf{F}_r^*)' \mathbf{D}^h_{\mathbf{g}_k} \mathbf{a}_j$ for all $k$
        $\mathbf{b}_{1j} \leftarrow (\mathbf{a}_j' \mathbf{a}_j)^{-1} \sum_{k=1}^K w_{kj} \mathbf{v}_{1kj}$
        $\boldsymbol{\alpha}_{kj} \leftarrow \arg\min_{\boldsymbol{\alpha}_{kj}} \|w_{kj}\mathbf{M}\boldsymbol{\alpha}_{kj} - v_{2kj}\|^2$ s.t. $\alpha_{1kj}, \alpha_{2kj}, \alpha_{3kj} \geq 0$ for all $k$
        $\mathbf{b}_{2kj} \leftarrow \mathbf{M}\boldsymbol{\alpha}_{kj}$ for all $k$ so that $\mathbf{b}_{kj} = (\mathbf{b}_{1j}, \mathbf{b}_{2kj})'$
        $\mathbf{a}_j \leftarrow \frac{2}{m+q-2}(\sum_{k=1}^K \mathbf{b}_{kj}' \mathbf{b}_{kj} \mathbf{D}^h_{\mathbf{g}_k})^{-1} \sum_{k=1}^K \mathbf{D}^h_{\mathbf{g}_k} \mathbf{F}_r^* \mathbf{b}_{kj}$
        $i \leftarrow i + 1$
      **calculate** $L_{ij} = L(\mathbf{a}_j, \mathbf{B}_j, \mathbf{G}_h)$
    **end while**
  **end for**
  **if** $n_a > 1$, set $(\mathbf{a}_1, \mathbf{B}_1) \leftarrow \arg\min_{(\mathbf{a}_j, \mathbf{B}_j)} L_{ij}$ and $n_a \leftarrow 1$
  **update** $h \leftarrow h + 1$ and $\mathbf{G}_{h-1}$ to $\mathbf{G}_h$ by reassigning each individual to the group which minimises $L$
  **calculate** $L_h = L(\mathbf{a}_1, \mathbf{B}_1, \mathbf{G}_h)$
**end while**
**return** $\hat{\mathbf{a}} = \mathbf{a}_1, \hat{\mathbf{B}} = \mathbf{B}_1$ and $\hat{\mathbf{G}} = \mathbf{G}_h$, and repeat for different starting values $\mathbf{G}_0$

Determining the optimal scores for the ratings requires constructing a design matrix from the spline basis functions evaluated at the rating categories 1 to $q$. Let this matrix be $\mathbf{M}^*$. The optimal scores are simply determined as $\mathbf{M}^*\boldsymbol{\alpha}_k$. In Section 4.3 a simulation experiment is conducted to assess how accurately this method can reproduce a known underlying correlation structure from contaminated data.

# 4 Simulation Results

## 4.1 Simulation Model

The simulated data was generated in a three-step procedure. First, the true underlying preference structure for the $m$ objects is obtained by simulating $m$ random numbers from a $U(0,1)$-distribution. These are gathered into the $m$-vector $\boldsymbol{\mu}$. Second, individual preferences are generated by simulating $n$ times from each of $m$ truncated normal distributions respectively centred at the elements of $\boldsymbol{\mu}$. The individual preferences are given by $\boldsymbol{\delta}_i = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i$, with $\boldsymbol{\varepsilon}_i, i = 1, \ldots, n$, representing the individuals deviation from the mean.

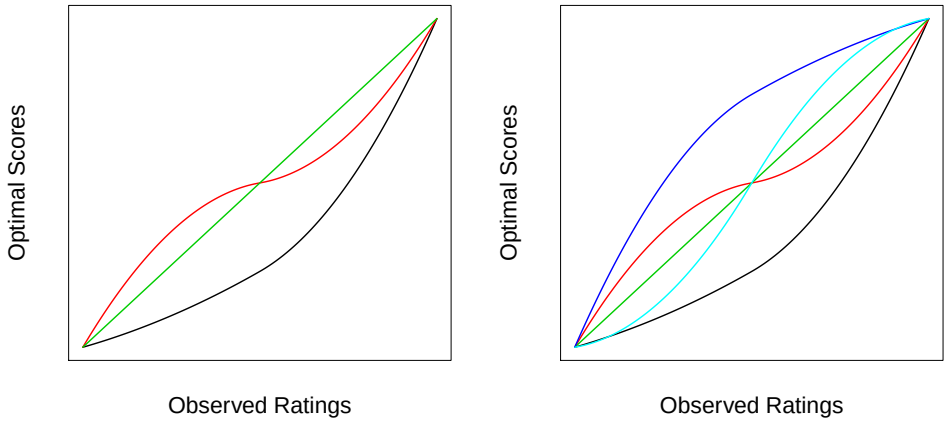
Truncation is done at 0 and 1 so that response styles can be defined on the closed interval $[0, 1]$. Note that the use of truncation avoids overflow problems at the lower and upper ends of the response style mapping, and hence improves on the original approach of Van de Velden (2007). The truncated normal draws are done independently and with error variance $\sigma^2$, which is an important parameter because it determines how pronounced the multi-modality of the mixture of truncated normals over $[0, 1]$ is.

The resultant true preferences are randomly divided into different response style groups. Finally, these data are discretised to $m$ categorical variables with $q$-point Likert-scales, according to the cut points on $[0, 1]$ implied by the chosen $K$ response styles. These response styles are parameterised to come from the family of monotone quadratic splines outlined in Section 3.2.

In the simulations, choices must be made regarding the following: the number of objects $m$, the number of rating categories $q$, the underlying standard deviation $\sigma$, the number of response styles $K$, as well as their shapes defined by $\boldsymbol{\alpha}_k, k = 1, \ldots, K$, the sample size $n$ and how this is divided among the $K$ groups, namely $n_k, k = 1, \ldots, K$.

## 4.2 Assessing Classification Performance

The first simulation study assesses the classification accuracy of the dual scaling method. It is assumed in this experiment that the number of groups $K$ is known beforehand.




(a) $K = 3$ (b) $K = 5$

Figure 4: Response styles used in the simulation study.

For each of the experimental conditions, 50 simulated data sets were constructed and the dual scaling method applied. For each data set estimation was based on 15 random starts for $\mathbf{G}$, and for each of these starts the ALS procedure was initialised from 50 different random configurations for the row scores $\mathbf{a}$.

The 108 experimental conditions consisted of the following. The number of objects $m$ was varied over 10, 20 and 30 items. The rating scales employed were either $q = 5$ or 7-point scales. Sample sizes of $n = 200$, 1000 and 5000 respectively were used. The number of groups $K$ were either 3 or 5. For each of these $K$, it was assumed that one of the groups has a linear response mapping (that is, a group with no response style). The additional $K - 1$ groups exhibited response styles through nonlinear mappings. For $K = 3$, these additional groups were acquiescence and extreme responding, since Baumgartner and Steenkamp (2001) found that these are most prevalent in survey data. For $K = 5$, groups for disacquiescence and midpoint responding were also added. The corresponding spline functions used to simulate from are shown in Figure 4. The sample of $n$ respondents was assigned to the groups by allocating either 20%, 50% or 80% of respondents equally among the $K - 1$ response style groups. These percentages represent the amount of contamination in the simulated data. The remaining percentage of respondents was assigned to the group exhibiting no response style. The latent standard deviation $\sigma$ was fixed at 0.1 for all experiments.

To assess the classification performance of the method, the adjusted Rand index as well as the percentage correctly classified (the so-called hit rate) were computed. The

adjusted Rand index (ARI) of Hubert and Arabie (1985) assesses the similarity between two partitions, adjusted for chance correspondences between these partitions. The upper limit of the ARI is one, and indicates perfect agreement. An ARI of zero indicates that the method does not improve on random assignment, with all positive values indicating an improvement. Negative ARI values are also possible, and indicate poorer performance than random assignment. The ARI is in general lower than the hit rate, and can be considered as a more objective measure of performance.

For each of the 108 experimental conditions, Tables 2 and 3 show the average values over the 50 simulated data sets. It is apparent that the sample size $n$ does not have a large influence on the ARI and hit rate. The number of groups $K$ is very important for performance when the contamination percentage is low (20%). This is because for $K = 5$ groups the 20% of contaminated data points must be divided into 4 groups instead of 2 when $K = 3$, which results in groups with very low proportions $n_k/n$ of the total sample. The low performance here is somewhat compensated for by using more items, such as $m = 30$, but for $K = 5$ groups even more items are needed. In general, using more items increases the classification accuracy. Using a larger number of rating categories $q$ also increases performance, but mostly so with fewer groups ($K = 3$). The method improves on random assignment – especially in cases with higher response style prevalence and 20 or more items the improvement is substantial.

| | | | $q = 5$ | | | | | | | | | $q = 7$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $n = 200$ | | | $n = 1000$ | | | $n = 5000$ | | | $n = 200$ | | | $n = 1000$ | | | $n = 5000$ | | |
| | RS% | $m =$ | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 |
| | 20% | | 0.28 | 0.40 | 0.61 | 0.30 | 0.42 | 0.62 | 0.29 | 0.40 | 0.62 | 0.31 | 0.48 | 0.74 | 0.29 | 0.48 | 0.80 | 0.30 | 0.48 | 0.80 |
| K = 3 | 50% | | 0.59 | 0.80 | 0.90 | 0.57 | 0.80 | 0.91 | 0.58 | 0.80 | 0.91 | 0.62 | 0.85 | 0.93 | 0.64 | 0.86 | 0.94 | 0.62 | 0.85 | 0.94 |
| | 80% | | 0.73 | 0.90 | 0.93 | 0.72 | 0.89 | 0.95 | 0.75 | 0.89 | 0.95 | 0.75 | 0.91 | 0.96 | 0.76 | 0.90 | 0.96 | 0.76 | 0.91 | 0.96 |
| | 20% | | 0.16 | 0.22 | 0.33 | 0.16 | 0.22 | 0.34 | 0.16 | 0.21 | 0.34 | 0.17 | 0.24 | 0.35 | 0.17 | 0.25 | 0.36 | 0.18 | 0.25 | 0.36 |
| K = 5 | 50% | | 0.42 | 0.65 | 0.82 | 0.42 | 0.65 | 0.81 | 0.42 | 0.65 | 0.82 | 0.44 | 0.67 | 0.86 | 0.44 | 0.66 | 0.84 | 0.44 | 0.66 | 0.85 |
| | 80% | | 0.70 | 0.85 | 0.93 | 0.70 | 0.86 | 0.93 | 0.71 | 0.86 | 0.93 | 0.73 | 0.88 | 0.94 | 0.73 | 0.88 | 0.95 | 0.73 | 0.88 | 0.95 |

Table 2: Average adjusted Rand index for 50 simulations at the different parameter settings.

| | | | $q = 5$ | | | | | | | | | $q = 7$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $n = 200$ | | | $n = 1000$ | | | $n = 5000$ | | | $n = 200$ | | | $n = 1000$ | | | $n = 5000$ | | |
| | RS% | $m =$ | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 |
| | 20% | | 0.66 | 0.76 | 0.87 | 0.67 | 0.77 | 0.87 | 0.67 | 0.76 | 0.88 | 0.69 | 0.81 | 0.92 | 0.67 | 0.81 | 0.94 | 0.68 | 0.81 | 0.94 |
| $K = 3$ | 50% | | 0.84 | 0.93 | 0.97 | 0.83 | 0.93 | 0.97 | 0.84 | 0.93 | 0.97 | 0.85 | 0.95 | 0.98 | 0.86 | 0.95 | 0.98 | 0.86 | 0.95 | 0.98 |
| | 80% | | 0.87 | 0.96 | 0.97 | 0.87 | 0.95 | 0.98 | 0.89 | 0.95 | 0.98 | 0.88 | 0.96 | 0.98 | 0.89 | 0.96 | 0.99 | 0.89 | 0.96 | 0.98 |
| | 20% | | 0.50 | 0.56 | 0.68 | 0.50 | 0.57 | 0.70 | 0.49 | 0.56 | 0.69 | 0.51 | 0.60 | 0.70 | 0.50 | 0.61 | 0.71 | 0.52 | 0.61 | 0.72 |
| $K = 5$ | 50% | | 0.72 | 0.86 | 0.93 | 0.71 | 0.86 | 0.93 | 0.71 | 0.86 | 0.93 | 0.73 | 0.87 | 0.95 | 0.74 | 0.87 | 0.94 | 0.74 | 0.87 | 0.94 |
| | 80% | | 0.84 | 0.93 | 0.97 | 0.84 | 0.94 | 0.97 | 0.85 | 0.94 | 0.97 | 0.86 | 0.95 | 0.98 | 0.86 | 0.95 | 0.98 | 0.86 | 0.95 | 0.98 |

Table 3: Average hit rates for 50 simulations at the different parameter settings.

## 4.3 Recovering Underlying Structure through Data Cleaning

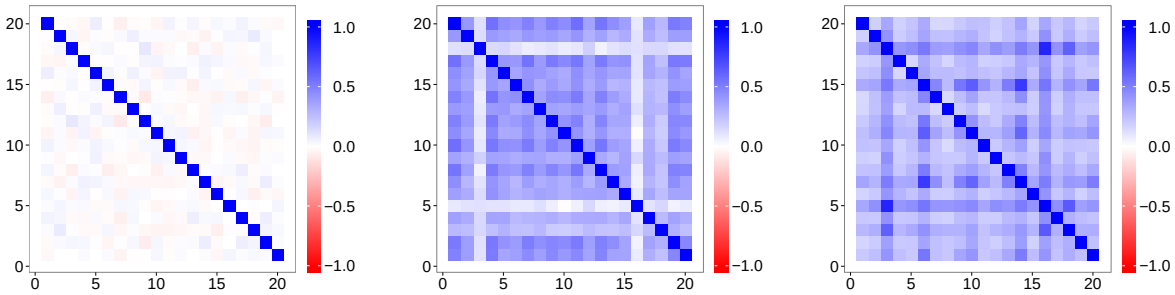
The simulation model of Section 4.1 assumes that, given the expected value of the object scores $m$, the objects are independently distributed as truncated normal distributions. Although the true correlation matrix between the objects thus is the identity matrix $\mathbf{I}$, the observed correlations after the response style contamination is often inflated. To show improvement, the cleaned data derived as in Section 3.6 should have correlations resembling independence more closely. A visual example is given in Figure 5 for simulated data ($m = 20, K = 3$ similar to the conditions used in Tables 2 and 3), where the colours indicate the magnitude of the Pearson correlations. It is evident that the response styles artificially inflate the correlations. When $q = 7$, the cleaned data to some extent succeeds in removing the spurious correlations, but when $q = 5$ the situation is not much improved.

The conditions under which the cleaned data can be expected to provide a better representation of the underlying correlation matrix was studied further through simulations. For the different values of $K, n, q$, and the proportion of response style contamination used in Section 4.2, 50 simulated data sets were constructed and cleaned through the dual scaling method. Here $m = 20$ was fixed for simplicity. For each of these data sets, the root mean square error (RMSE) between $\mathbf{I}$ and the empirical Pearson correlation matrix for the contaminated data was calculated, where

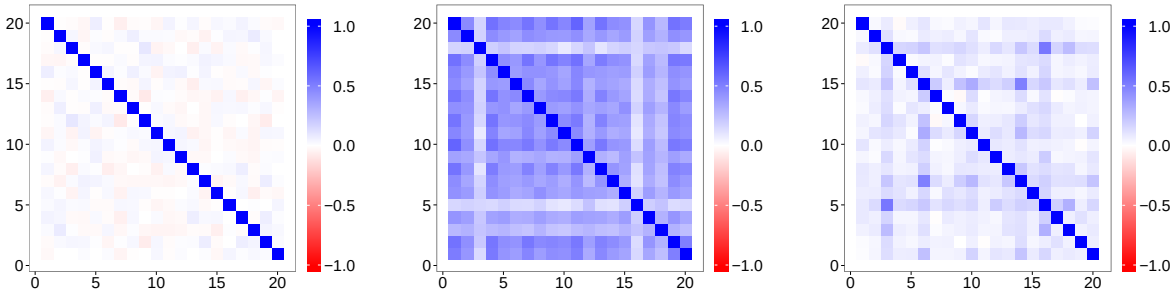
$$\text{RMSE}(\mathbf{V}, \mathbf{W}) = \sqrt{\sum_i \sum_j (v_{ij} - w_{ij})^2} \tag{14}$$

for commensurable matrices $\mathbf{V}$ and $\mathbf{W}$. Similarly, the RMSE comparing $\mathbf{I}$ with the empirical Pearson correlations of the cleaned data can be computed. A reduction in the RMSE when using the cleaned data as opposed to the contaminated data indicates that the cleaned data has a correlation structure which matches the true correlation structure more closely.

A two-sample Wilcoxon test, also known as the Mann-Whitney test, (e.g. Rice, 2007) was used to test the null hypothesis that the RMSE is equal for the contaminated and cleaned data against the one-sided alternative that the RMSE for the contaminated data is greater than that of the cleaned data. The results are quite clear: when $q = 7$ the null hypothesis is always rejected ($p < 0.001$) in favour of the alternative, whilst when $q = 5$ the null hypothesis cannot be rejected even once (all $p > 0.2$). It can therefore be deduced that when a sufficient number of rating categories $q$ are used, the correlation




(a) *True preferences* $q = 5$ (b) *Observed ratings* $q = 5$ (c) *Cleaned data* $q = 5$




(d) *True preferences* $q = 7$ (e) *Observed ratings* $q = 7$ (f) *Cleaned data* $q = 7$

Figure 5: The effect of response styles on the underlying uncorrelated objects: estimated Pearson correlations before and after contamination, as well as after cleaning the data. The number of items is $q = 5$ for (a) – (c) and $q = 7$ for (d) – (f) .

structure of the cleaned data is more representative of the true underlying structure of the data.

A related question concerns the performance of the method in the presence of a nontrivial correlation structure. To impose such a structure whilst retaining truncated normal marginal distributions for the objects, a copula is used (note that the truncated multivariate normal distribution does not guarantee truncated normal marginals). A copula is a multivariate distribution function $C(u_1, u_2, \ldots, u_m)$ with uniform marginals (Hofert & Mächler, 2011). According to Sklar's theorem (Sklar, 1959; Hofert & Mächler, 2011) a multivariate distribution function $F$ with marginals $\{F_j\}_{j=1}^m$ can be constructed as

$$F(x_1, x_2, \ldots, x_m) = C(F_1(x_1), F_2(x_2), \ldots, F_m(x_m)). \tag{15}$$

The marginal truncated normal distributions can be achieved by the inverse probability integral transform. The dependence structure between the variables is solely determined by the copula. Here two independent Clayton copula (Clayton, 1978) functions will be used to impose a correlation structure in terms of Kendall's $\tau$, a well-known measure of rank correlation (see Kendall, 1938; Hofert & Mächler, 2011). The structure induced here for $m = 20$ is as follows: the first 10 objects are correlated with $\tau = 0.2$, independent of the other 10 objects which are correlated with $\tau = 0.35$. These $\tau$ values amount to Pearson correlations of approximately $\rho = 0.3$ and $\rho = 0.5$ respectively (an approximate relationship is $\rho \approx \sin(\tau\pi/2)$ - see Kendall and Gibbons (1990)). It is also possible to introduce negative correlations by using $1-U$ instead of $U$ in the inverse probability integral transform. In the application here these reversals are made randomly with differing probability $\gamma$. The theoretical correlations given by Kendall's $\tau$ for one such copula is illustrated in Figure 6.

The difference in RMSE can again be used to evaluate the effect of the data cleaning on the correlation structure, now using Kendall's $\tau$ since the Clayton copula's use this measure directly. A simulation study was conducted for $m = 20$ objects with the other parameters varying as before. For each combination of the parameters, the RMSE was calculated for 50 randomly generated data sets according to the copula model described above. Then for each data set the constrained dual scaling model was fit as before, and a cleaned data set constructed. The difference in the RMSE for the contaminated data as compared to the cleaned data was recorded.

Table 4 presents the average reduction in RMSE as a result of cleaning the data with the dual scaling procedure. As before the two-sample Wilcoxon test was performed. Significant improvements were found in all cases except those printed in italic in Table 4.

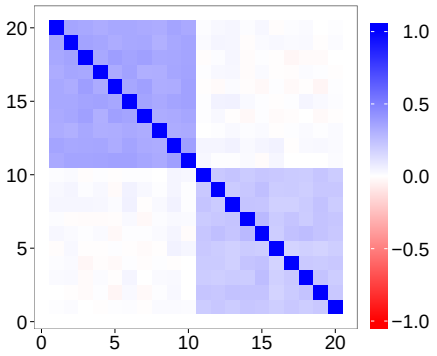



Figure 6: An example of the correlation structure imposed by the Clayton copula's, in terms of Kendall's $\tau$.

It is apparent that the cleaned data improves the RMSE in all cases, except where both $q$ and $K$ are small and the proportion of contamination is moderate (50%) to large (80%). Except for these circumstances, the constrained dual scaling method improves the estimation of the true correlation structure by removing the response styles effects.

| | | $q = 5$ | | | | | | | | | $q = 7$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $n = 200$ | | | $n = 1000$ | | | $n = 5000$ | | | $n = 200$ | | | $n = 1000$ | | | $n = 5000$ | | |
| | RS% $\gamma =$ | 0.5 | 0.75 | 1.0 | 0.5 | 0.75 | 1.0 | 0.5 | 0.75 | 1.0 | 0.5 | 0.75 | 1.0 | 0.5 | 0.75 | 1.0 | 0.5 | 0.75 | 1.0 |
| $K = 3$ | 20% | 0.08 | 0.09 | 0.33 | 0.05 | 0.03 | 0.35 | *0.02* | 0.01 | 0.36 | 0.67 | 0.71 | 0.45 | 0.63 | 0.77 | 0.44 | 0.69 | 0.74 | 0.48 |
| | 50% | *-0.10* | *-0.09* | *-0.07* | *-0.09* | *-0.14* | *-0.24* | *-0.07* | *-0.02* | *-0.15* | 0.64 | 0.70 | 0.83 | 0.70 | 0.70 | 0.86 | 0.64 | 0.69 | 0.87 |
| | 80% | *-0.41* | *-0.37* | *-0.44* | *-0.34* | *-0.41* | *-0.47* | *-0.38* | *-0.43* | *-0.46* | 0.60 | 0.65 | 0.81 | 0.64 | 0.66 | 0.79 | 0.61 | 0.66 | 0.8 |
| $K = 5$ | 20% | 0.09 | 0.19 | 0.50 | 0.14 | 0.19 | 0.55 | 0.14 | 0.15 | 0.54 | 0.75 | 0.85 | 0.47 | 0.70 | 0.82 | 0.48 | 0.70 | 0.79 | 0.49 |
| | 50% | 0.12 | 0.15 | 0.18 | 0.12 | 0.14 | 0.21 | 0.13 | 0.14 | 0.26 | 0.71 | 0.75 | 0.93 | 0.70 | 0.76 | 0.94 | 0.70 | 0.76 | 0.92 |
| | 80% | 0.10 | 0.12 | 0.07 | 0.07 | 0.11 | 0.12 | 0.08 | 0.11 | 0.10 | 0.70 | 0.72 | 0.85 | 0.68 | 0.72 | 0.85 | 0.68 | 0.72 | 0.85 |

Table 4: Average proportional improvement in the RMSE when comparing the cleaned to the contaminated data. A two-sample Wilcoxon test for no difference in the RMSE was performed, against the alternative hypothesis that the cleaned data significantly reduces the RMSE. Significant improvements (at the 95% level) was observed for all tests *except* those shown in italic print.

# 5 Application

To illustrate the method applied to empirical data, consider data obtained from an unnamed multinational food and beverage conglomerate regarding an investigation of product perceptions for 20 similar products. The products include in-house products as well as those of competitors. Data were collected from $n = 268$ panelists, who supplied Likert scores for 7 attributes of these products. Each product is rated on all 7 items or questions, so that there are 140 items collected in a data matrix with 268 rows and $m = 140$ columns. The Likert scale ranges from 1 ("low") to 9 ("high"), and hence $q = 9$.

The first step is to select $K$ by inspecting the loss function through a scree plot. Consideration is also given to the curvature properties of the splines as well as how well the method separates groups of panelists who exhibit different distributions of rating scale use. It is expected that once spurious clusters are added at least two of the estimated response curves will be very similar, and/or that two groups will on aggregate use the rating scale in a very similar fashion. For each of $K = 1, 2, \ldots, 8$ groups, the algorithm was run from 50 different random starts for the grouping matrix $\mathbf{G}$, where appropriate. Also, 50 random starts for the alternating least squares (ALS) part of the algorithm was used.

Figure 7 shows the resulting (rescaled) scree plot. There does not seem to be a clear "elbow" in the plot, although it is apparent $K = 3, 4$ and 5 are the options requiring closer scrutiny. As $K$ increases beyond 5 not much improvement in the loss function is observed.

The response mappings for the solutions $K = 1, \ldots, 8$ are displayed in Figure 8. In these plots the horizontal axis contains the original rating scale, while the vertical axis denotes the optimal scores assigned to the Likert scale. The area of the bubbles superimposed on the transformation plots indicate how often each rating category is used, aiding in the interpretation. A first observation is that (strictly, almost) all the detected response mappings have the characteristic convex shape of acquiescence. This means that all panelists have a tendency to use positive ratings frequently. The groups differ with respect to the intensity of the acquiescence.

Furthermore, the range of optimal scores that is assigned to each group, namely $\sum_{i=1}^{3} \alpha_{ik}$ in terms of the spline parameters set out in Sections 3.2 and 3.3, depends on the within-group variability of rating scale use. Groups where individual panelists' rating scale use show more variability from the group's aggregate rating scale use are assigned optimal scores with a wider range. Hence the method treats such groups, i.e. groups

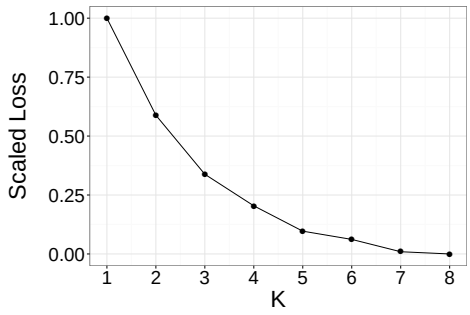
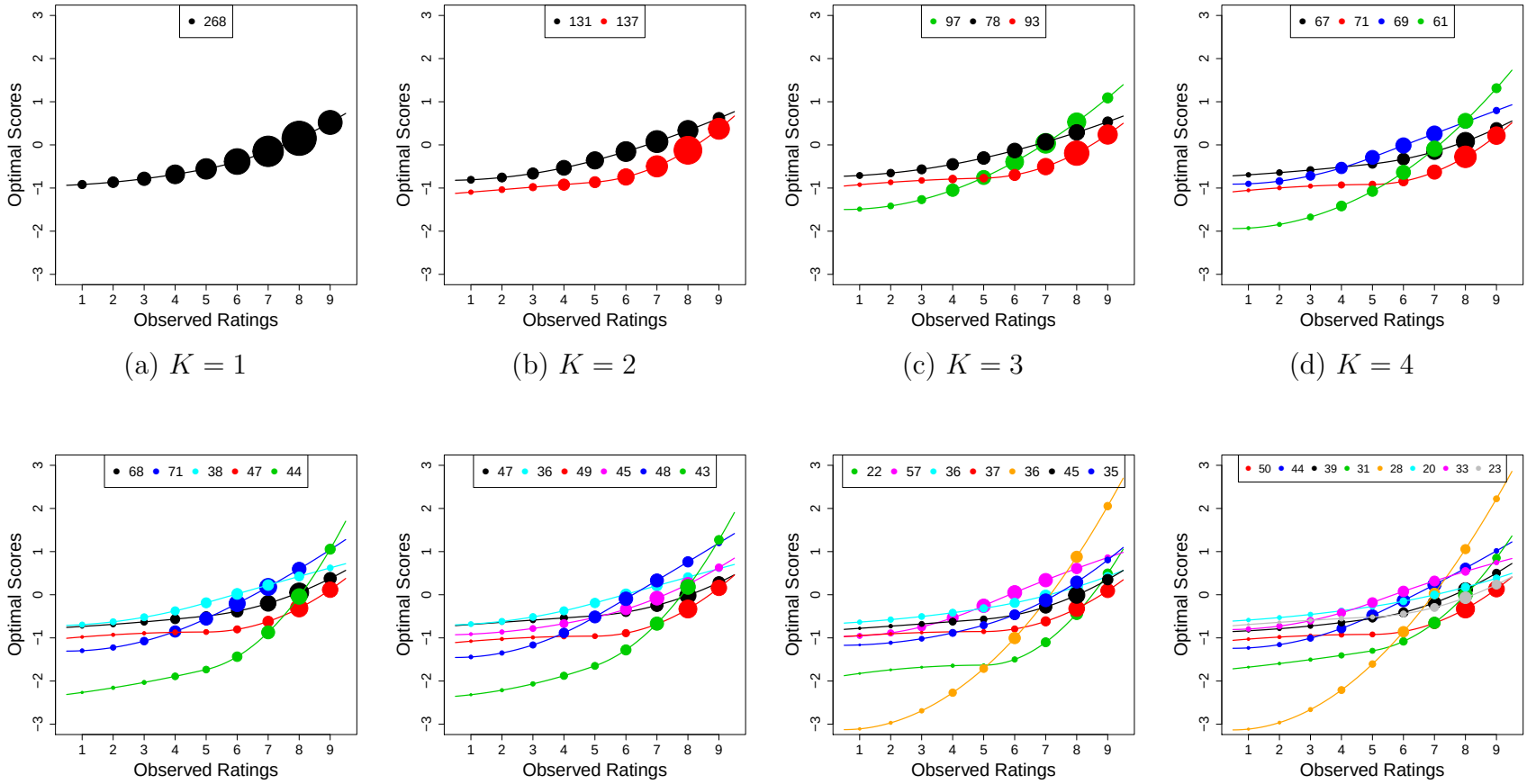



Figure 7: Scree plot for the panelist's ratings.

containing more individualistic respondents, as more informative as opposed to groups with more uniform response behaviour.

A closer look at the distribution of the rating scale use in the identified groups reveal that all groups in the solutions $K = 3, 4$ and 5 show visually different distributions, except the black and green groups when $K = 5$. The relative frequencies with which each rating is used in each of the groups when $K = 4$ and 5 is shown in the barplots in Figure 9. It is obvious that the black and green groups when $K = 5$ have very similar aggregate behaviour. This is however not immediately apparent from the spline functions displayed in Figure 8, which assign different optimal scores to these groups.

A more formal comparison can also be made by using the Kullback-Leibler (KL; e.g. Lehmann & Casella, 1998) divergence between the distributions of different groups – this is also known as entropy distance. It is an asymmetric measure of the dissimilarity between two density functions, the reference density $f$ and another density $g$, defined as $E_f[\log(f(X)/g(X))]$. When $f = g$, the entropy is zero. Assessing the pairwise KL divergence for all pairs of groups (and using both $f$ and $g$ as reference) show that indeed the above mentioned two groups when $K = 5$ diverge the least among all pairs – see Table 5. Since the method is designed to detect groups with different aggregate rating scale use it can be concluded that the addition of a fifth group is spurious and therefore $K = 4$ is selected.



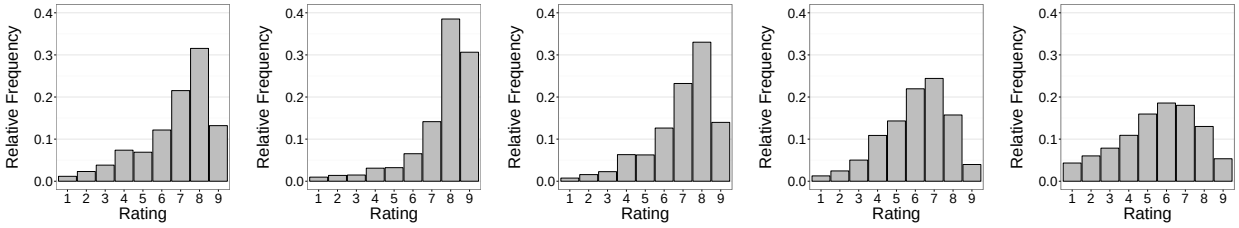

(a) $K = 1$ (b) $K = 2$ (c) $K = 3$ (d) $K = 4$

(e) $K = 5$ (f) $K = 6$ (g) $K = 7$ (h) $K = 8$

Figure 8: The estimated response mappings for $K = 1, \ldots, 8$. The area of the bubbles are proportional to how often that particular rating is used. The group sizes are also shown in a legend.

| | *Black* | *Red* | *Green* | *Blue* | *Cyan* |
|---|---|---|---|---|---|
| *Black* | - | 0.158 | 0.009 | 0.187 | 0.234 |
| *Red* | 0.161 | - | 0.138 | 0.699 | 0.701 |
| *Green* | 0.008 | 0.134 | - | 0.224 | 0.297 |
| *Blue* | 0.166 | 0.606 | 0.202 | - | 0.053 |
| *Cyan* | 0.231 | 0.680 | 0.317 | 0.065 | - |

Table 5: The Kullback-Leibler divergence between the groups when $K = 5$, based on the rating scale use per group. The distributions of the groups in the rows are treated as the respective reference distributions, $f$.

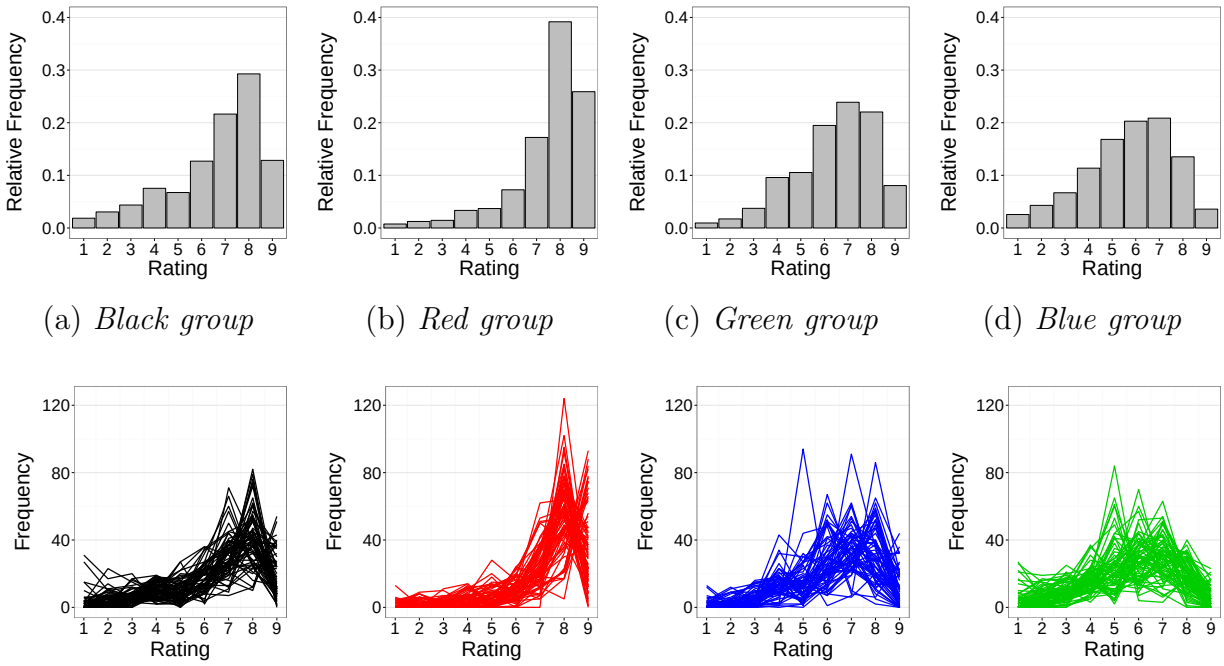



(a) *Black group* (b) *Red group* (c) *Green group* (d) *Blue group* (e) *Cyan group*

Figure 9: Relative aggregate frequencies of rating scale use in the identified groups when $K = 5$.

Consider the results for $K = 4$ groups. These four groups consist of 67, 71, 69 and 61 panelists respectively. The rating scale usage of these groups are displayed in Figure 10 panels (a) – (d). Figure 11 displays the optimal scores assigned to the ratings in the different groups as well as their curvature chart. The curvature chart includes an approximate 95% confidence ellipse constructed for the parameter estimates of 5000 data sets simulated under the assumption that no response styles exist. Any group falling outside this band therefore has a significantly nonlinear response mapping and hence a response style.

The first group (black in previous plots) represents acquiescence as mainly ratings 6 to 9 are used by panelists. There is a slight boundary effect, as also with the other groups, in that a 9 is used less often than an 8. Because the categories 6 to 9 are frequently used, the optimal scores assigned to these are close to zero. The most meaningful optimal scores are assigned to the lower categories since when these are used it contains more information for this group of panelists. Overall the information provided by this group is low since the range of optimal scores assigned is very narrow. This is because the group members display low variability with respect to their rating scale use. This is evident from Figure 10 (e), which plots the frequency with which each rating is used per individual.




(a) *Black group* (b) *Red group* (c) *Green group* (d) *Blue group*

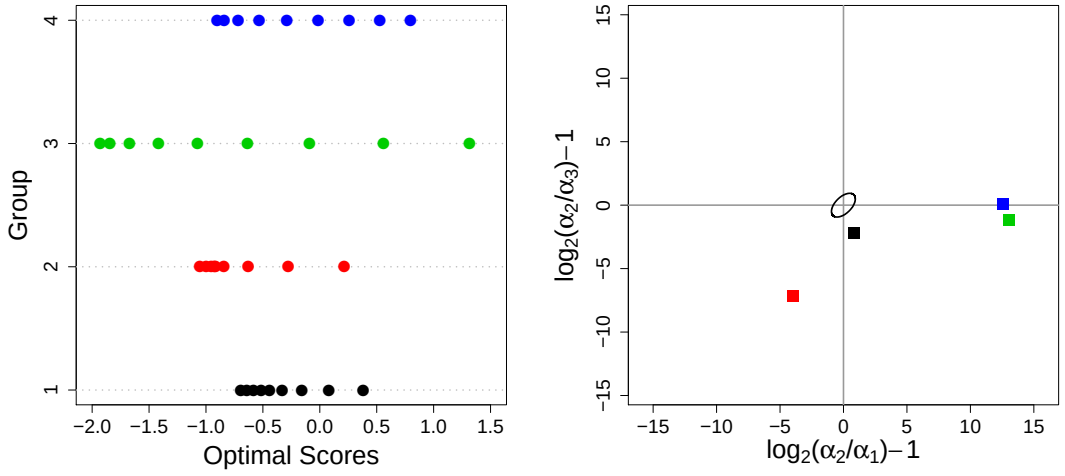
(e) *Black group* (f) *Red group* (g) *Green group* (h) *Blue group*

Figure 10: (a) - (d): Relative frequencies of rating scale use for the chosen solution $K = 4$; and (e) - (h) Variability of rating scale use within these groups, with each line representing a single individual.

The second group (red) represent a more extreme acquiescence where 7 to 9 are often used. The range of assigned optimal scores, and hence information, is similarly narrow, but shifted further to the left since the upper categories are used even more frequently. Since the response mapping is concave in the lower part of the domain there is a slight deviation from acquiescence towards an extreme response style.

The green and blue groups both exhibit a mix of acquiescence and midpoint responding. This is evident from the relative frequencies in Figure 10 and the curvature chart in Figure 11 (b). In these groups generally ratings 4 to 8 are preferred. Based on the range of optimal scores assigned to them these consist of the panelists providing the most information. Especially the green group is endowed with the most meaningful spread of optimal scores, and can be seen in Figure 10 (g) to exhibit the most within-group variation.

Finally, consider the optimal scores assigned to the items as displayed in Figure 12. It is evident that Product R, and to a lesser extent Products N, D, E and F, received the lowest ratings. In contrast, Product P was the best performing one. By using a cleaned




(a) (b)

Figure 11: (a) Optimal scores assigned to the $K = 4$ response style groups, from rating 1 (left) to rating 9 (right). (b) Curvature plot similar to Figure 3 for the four groups, with the axes now transformed to obtain a more symmetrical plot. The ellipse in the centre is an approximate 95% confidence ellipse for no response style.

data set constructed by replacing the ratings by optimal scores further analyses can be conducted which are less influenced by the presence of the response styles.

# 6 Conclusions

A method that relies on the properties of dual scaling for successive category data to detect response styles in categorical data was presented. It combines newly suggested spline models for four main types of response styles with the original dual scaling method to construct optimal scores for the boundaries between rating categories. These optimal scores are sensitive to the presence of response styles. The method was adapted to allow for multiple response style groups by utilizing a $k$-means type procedure, which is combined with a constrained alternating least squares algorithm using nonnegative least squares to fit the model.

Both the ability of the method to detect reponse styles and the improvement in correlation structure that results from a cleaned data set where ratings are replaced by optimal scores were studied. It was found that using 30 or more items and a rating scale of 7 or

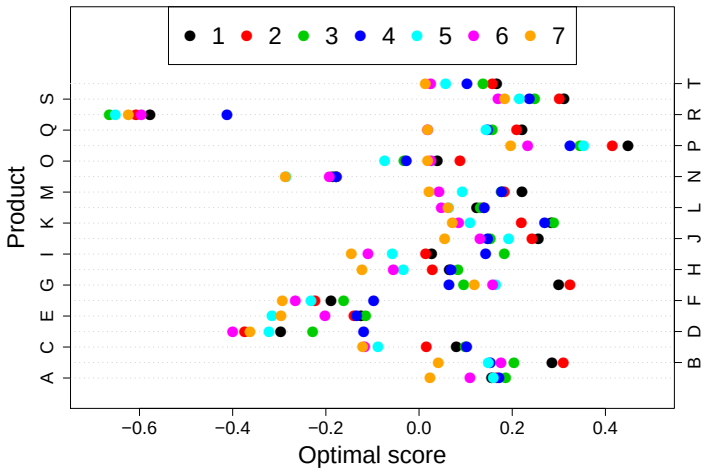



Figure 12: Optimal scores for each of the 7 questions, separated by product and with similar items depicted by the same colours.

more categories yields great improvements in the classification of individuals to different response style groups. When fewer rating categories are used other factors become important, such as the extent to which response styles are present in the data. Also, when using a 7-point scale or more, the resulting cleaned data provide a more accurate description of the true substantial content in the data, after accounting for different repsonse styles. The use of the method to identify respondents who provide similar amounts of information in their repsonses to a survey was illustrated on an empirical data set.

The number of response style groups to retain was selected on the grounds of a scree plot of the loss function, combined with the distribution of rating scale use in the different response style groups. It remains to be seen whether a more formal selection method can be derived. Other grounds for further research include alternatives for or additional restriction to the spline functions, and whether more freedom is needed by allowing for differences between the $m$ object scores in different groups.

# References


Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151.

De Jong, M. G., & Steenkamp, J. B. E. M. (2010). Finite mixture multilevel multidimensional ordinal irt models for large scale cross-cultural research. *Psychometrika*, *75*(1), 3–32.

Greenacre, M. J. (2007). *Correspondence analysis in practice*. Chapman & Hall, Boca Raton.

Hofert, M., & Mächler, M. (2011). Nested archimedean copulas meet R: The nacopula package. *Journal of Statistical Software*, *39*(9), 1–20. Retrieved from `http://www.jstatsoft.org/v39/i09/`.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563–583.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods*. Edward Arnold, London.

Lawson, C. L., & Hanson, R. J. (1974). *Solving least squares problems*. Prentice Hall, Upper Saddle River.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (Vol. 31). Springer, New York.

Nishisato, S. (1980a). *Analysis of categorical data: Dual scaling and its applications*. University of Toronto Press, Toronto.

Nishisato, S. (1980b). Dual scaling of successive categories data. *Japanese Psychological Research*, *22*(3), 134–143.

Nishisato, S. (1994). *Elements of dual scaling: an introduction to practical data analysis*.

L. Erlbaum Associates, Hillsdale.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0.)

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 425–441.

Rice, J. A. (2007). *Mathematical statistics and data analysis*. Duxbury, Belmont.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'institut de Statistique de l'Université de Paris*, *8*(1), 11.

Torres, A., & Greenacre, M. (2002). Dual scaling and correspondence analysis of preferences, paired comparisons and ratings. *International Journal of Research in Marketing*, *19*(4), 401–405.

Van de Velden, M. (2000). Innovations in multivariate statistical analysis. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), (Vol. 36, p. 87-99). Kluwer Academic Publishers, Dordrecht.

Van de Velden, M. (2007). *Detecting response styles by using dual scaling of successive categories* (Econometric Institute Report). Erasmus University Rotterdam, Rotterdam.

Van de Velden, M., Groenen, P. J. F., & Poblome, J. (2009). Seriation by constrained correspondence analysis: A simulation study. *Computational Statistics & Data Analysis*, *53*(8), 3129–3138.

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, *47*(1), 157–172.

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747.