2019 Doctoral Thesis

Clustering and visualization for enhancing interpretation of categorical data

Graduate School of Culture and Information Science,
Doshisha University

Mariko Takagishi

Supervisor Prof. Hiroshi Yadohisa

Submitted

Abstract

Large-scale categorical data are often obtained in various fields. As an interpretation of large-scale data tends to be complicated, methods to capture the latent structure in data, such as a cluster analysis and a visualization method are often used to make data more interpretable.

However, there are some situations where these methods failed to capture the latent structure which is interpretable. Therefore in this paper, two problems that often occur in large-scale categorical data analysis is considered, new methods to address these issues are proposed.

In Chapter 2, a problem of response style often contained in ordinal categorical data is considered. A response style is defined as a respondent's systematic response tendencies irrespective of the item content. For example, some respondents may tend to select categories at the ends of the scale, which is called an "extreme response style". A cluster of respondents with an "extreme response style", can be mistakenly identified as an item based cluster. To address this issue, I, van de Velden and Yadohisa propose a new method to cluster respondents based on their indicated preferences for a set of items while simultaneously correcting for response style bias, which we call Correcting and Clustering Response Style (CCRS). Specifically, we assume the existence of response functions that can be used to model response styles. We then simultaneously estimate these response functions and perform a cluster analysis based on the corrected preference data. A simulation study is performed to evaluate the proposed method by comparing the accuracy of clustering with the existing methods. In addition, we apply our CCRS to empirical data from four diffrent countries concerning social values, and show using CCRS, we can get a result which seems more interpretable than the one by existing method, in the sense that results by existing methods seem to only indicate individual's response style information.
In Chapter 3, enhancing an interpretation of visualization method on categorical data is considered. When categorical data are large scale, Multiple Correspondence Analysis (MCA) is often used to visualise the data structure by reducing the dimension of data. In general, incorporating external information on MCA biplot can be useful to enhance the interpretation. In this chapter, only categorical variables are considered as the external information. Then the aim is set to visually interpret how associations among the categorical variables differ with respect to external information class. The naive approach to achive our objective is to get the average of quantification for each class, and plot them as well as other categories. However with this approach, when there are heterogeneous tendencies within a class, all of them cannot be interpreted in the MCA biplot. Therefore,

I and van de Velden propose Multiple Set Cluster CA (MSCCA), to address the issue. Specifically, we find clusters for different classes of data, and then simultaneously estimate quantifications for categories and clusters from each class in common low dimensional space. By doing this, we can visualize heterogeneous tendencies in each class in a single biplot. By a simulation study, we investigate how the selection of external information variable affects the accuracy of biplot and clustering. In addition, we apply MSCCA to empirical data set about accidents, and show MSCCA yields a biplot which visualizes heterogeneous tendencies in each class, which helps characterize the external information class, compared to the existing methods.

By proposing these two new methods, we can expect that large-scale categorical data which has not been easily interpreted can be more interpretable, and this can help finding new knowledge via data analysis.

Contents

Chapter 1

Introduction

Large-scale categorical data are often obtained in the social sciences, biomedical, and marketing research (Agresti, 2013). For interpretation of large-scale data, it is useful to capture the latent structure in data. Methods to achive this objective include a cluster analysis such as k-means, a method to identify group of individuals having similar tendencies, and a visualization method such as Multiple Correspondence Analysis (MCA), a method to visualize the latent structure of categorical data by reducing the dimension of data.

However, with these methods, sometimes it is difficult to interpret the result. For example, ordinal categorical data are often affected by response style, here which is defined as an individual-specific response tendency irrespective of item contents. If data contains response style bias, cluster analysis may yield clusters of respondents with similar response styles, which is not of interest of the analysis. For example, some respondents may tend to select categories at the ends of the scale, which is called an "extreme response style". A cluster of respondents with an "extreme response style", can be mistakenly identified as an item based cluster.

Another example of failing to obtain interpretable result is in visualization method of categorical data. To visualize categorical data, Multiple Correspondence Analysis (MCA) is often used. In MCA, the external information on individuals (e.g. gender and nationality) is often incorporated to enhance the interpretation of MCA biplot. Using external information, it can be interesting to know how associations among the categorical variables differ with respect to external information class. However, tendencies that many individuals have in common in each class are only interpretable. That is, when there are heterogeneous tendencies within a class, all of them are cannot be interpreted in the MCA biplot.

Therefore, in this paper, these issues to enhance the interpretation of categorical data are addressed. In Chapter 2, the response style bias problem is considered, and a new method proposed by , to cluster respondents based on their indicated preferences for a set of items while simultaneously correcting for response style bias, is mentioned. In Chapter 3, I consider the second problem in MCA biplot, and propose a new visualization method by extending MCA. By the proposed method, I can visualize heterogeneous tendencies in

each external information class in a single biplot.

Both methods are evaluated by conducting simulation study and applying empirical data set. In empirical data example, by comparing the result of proposed method with the one by existing methods, I show how interpretation of result by data analysis is enhanced by our proposed methods.

Chapter 2

Correcting and clustering response style biased categorical data

2.1 Problem of response style in ordinal categorical data

In cluster analysis, respondents are allocated to groups of similar observations (MacQueen, 1967). In many applications, respondents are clustered based on ordinal categorical data, when cluster structure is assumed to exist in data. In this section, among ordinal categorical data, we mainly consider preference data, which is often measured in questionnaires in which respondents indicate their preference using a rating scale, e.g., a Likert scale, where respondents make selections from a set of predetermined preference categories. Clustering respondents relative to their answers may be useful to identify latent clustering structures.

Questionnaire-based preference data may be affected by so-called response styles. The response styles have been defined in several ways depending on the context. Baumgartner and Steenkamp (2001) mentioned that

> *response styles may be defined as tendencies to respond systematically to questionnaire items on some basis other than what the items were specifically designed to measure (Baumgartner & Steenkamp, 2001, p.143).*

Response styles discussed in Baumgartner and Steenkamp (2001) and commonly seen in the literature can be categorized as follows: tendencies to respond based on contents but not based on what the item intended to measure (e.g., socially desirable responding), and tendencies to respond irrespective of item content. Baumgartner and Steenkamp (2001) mainly focused on the latter category of response styles.

Moreover, the latter category can be further divided into two types: tendencies to select specific categories irrespective of content (e.g., tendencies to select only categories at the ends of the scale), and others (e.g., tendencies to respond carelessly, or nonpurposefully).

In this paper, we focus on the first type of response styles in the latter category. That is, in this paper, response styles are defined as respondent's systematic response tendencies

selecting specific categories irrespective of item content, such as extreme response style and a midpoint response style, a tendency to only select the middle of the scale. We focus on this type of response styles in this paper, because these are commonly seen in practice, it is rather simple to quantify such response styles from responses, and thus many statistical methods have been proposed for this type of response styles (e.g., van Rosmalen, Van Herk, & Groenen, 2010; Schoonees, van de Velden, & Groenen, 2015; Böckenholt & Meiser, 2017). In this paper, we refer to data in which observations are affected by these response styles as "response-style-biased data".

Response styles are related to various factors, including culture (Cheung & Rensvold, 2000; Meisenberg & Williams, 2008), education (Meisenberg & Williams, 2008), gender (Austin, Deary, & Egan, 2006; Weijters, Geuens, & Schillewaert, 2010), and age (Stukovsý, Palat, & Sedlakova, 1982). In cross-cultural surveys, typically several of the above-mentioned factors are present and response style bias is considered particularly significant (Baumgartner & Steenkamp, 2001). Moors (2012) and Cheung and Rensvold (2000) showed that response styles can lead to incorrect conclusions. Biases due to response styles can be considered as "systematic error", rather than "random error" (Baumgartner & Steenkamp, 2001). Therefore, to perform a meaningful data analysis, such systematic errors must be considered.

In practice, if data contains response style bias, cluster analysis may yield clusters of respondents with similar response styles ("response-style-based clusters"), rather than clusters with similar item preferences ("content-based clusters"). For example, assume that in a survey one group of respondents tends to select midpoint categories, while another group tends to favor endpoint categories, regardless of their preferences. Applying cluster analysis to the resulting data may extract clusters of respondents who have selected midpoint and endpoint categories. However, these clusters only reflect their response styles and any content-based structure in the data remains undetected.

Several methods have been proposed to detect and control for response style bias. The previous works can be divided into two types: probabilistic or non-probabilistic method. Many of former methods are proposed within the Item Response Theory (IRT) framework, Böckenholt and Meiser (2017) reviewed two types of IRT models designed to handle response styles: threshold-based models such as polytomous Rasch models and their mixture extensions (Rost, 1991; von Davier & Yamamoto, 2007), and an item response (IR) tree model (Böckenholt, 2012, 2017), which can be used to distinguish the effects of the judgment processes associated with content and response style. Plieninger and Meiser (2014) also validated several IR tree methods using an empirical dataset. In other IRT related research involving response styles, IRT and mixture IRT models have further been applied to correct for response style by adjusting parameters representing the response styles (Austin et al., 2006; Bolt & Johnson, 2009; Meiser & Machunsky, 2008; Morren, Gelissen, & Vermunt, 2012).

The other probabilistic method proposed in non-IRT framework was proposed by van Rosmalen et al. (2010). The primary objective of their latent-class bilinear multinomial logit model was to investigate how response style and item content (and background

variables, if relevant) affect responses in a low-dimensional space.

In many probabilistic models, probabilities for selecting each category are modeled, and these probabilities are then used to identify the presence of response-styles. However, this requires many assumptions (e.g. the distribution on data), and tends to need relatively large sample sizes for the parameter estimation (e.g., Finch & French, 2012, p. 177).

On the other hand, as non-probabilistic model, Schoonees et al. (2015) proposed constrained dual scaling (CDS), which was designed to detect several, typically more than two, types of response styles and, compared to other studies, focuses more on correcting the response style bias. While other probabilistic models control for response styles by adjusting parameters related to the probabilities for selecting specific ratings, in CDS the correction is done by transforming the original value.

In this paper, we focus on non-probabilistic method, because in Schoonees et al. (2015), the accuracy of correction was investigated using a simulation, while other papers tend to examine the correction only by the empirical study. Then, we consider the application of k-means cluster analysis to CDS-corrected data and refer to this as "CDS tandem analysis".

CDS is an extension of dual scaling for preference data (Nishisato, 1980), which involves dimension reduction. Specifically, Schoonees et al. (2015) formulated a constrained dual scaling approach that yields parameters that can be interpreted as response styles. To estimate the parameters in CDS, dimension reduction is applied. In particular, a one-dimensional solution is required to estimate the response styles. However, the use of dimension reduction implies a loss of respondent-specific information that may complicate the retrieval of accurate content-based clusters. In other words, CDS can remove respondents' differences that may be useful for content-based clustering.

To address these problems, we propose a new method for correcting and clustering response-style-biased data. Throughout this paper, we refer to our new method as CCRS. To achieve our objective, we first focus on correction of response styles, and introduce a framework to detect, and correct for, response styles by generalizing the definition of response styles used in CDS. In this way, we obtain a new correction method that does not require dimension reduction and that includes CDS as a special case. Next, we consider content-based clustering of the corrected data. However, rather than performing these steps sequentially, we propose to simultaneously correct for respondent-specific response styles and apply content-based clustering to the corrected data. By this simultaneous approach, we avoid a potential problem associated with the CDS tandem analysis, where the response style correction removes information relevant for the content-based cluster analysis. Note that, although in this paper we only consider content-based clustering, our new correction method can be used in combination with other data analysis methods as well.

The remainder of this chapter is organized as follows. In Section 2.2, we formalize the idea of response functions to identify and correct for response styles. In Section 2.3, we introduce our CCRS method, briefly describe CDS to show how it is different from CCRS as a correction method. Also, several characteristics and properties of CCRS are considered.

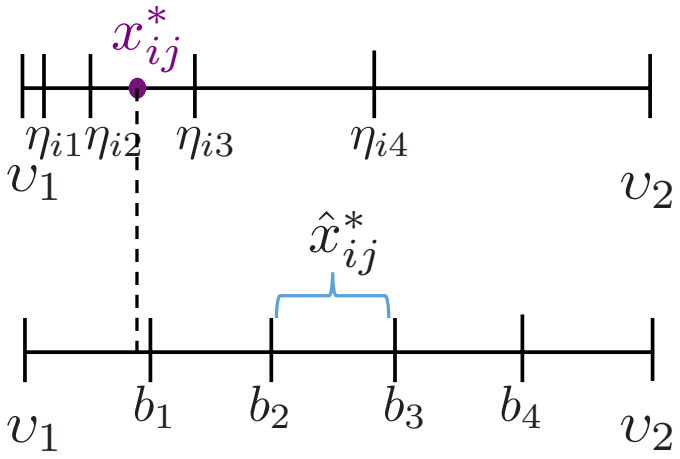

Figure 2.1: Response style bias: On the upper scale $[\upsilon_1, \upsilon_2] \subset \mathbb{R}$, respondent-specific boundaries are shown, while on the lower scale (equal-spaced) reference boundaries are shown. x^*_{ij} indicates the true preference, and $\hat{x}^*_{ij} \in (b_2, b_3]$ is the estimation of x^*_{ij} on a scale with reference boundaries b_ℓ, when $x_{ij} = 3$ is obtained. The set of $\eta_{i\ell}\,(\ell = 1, \ldots, q-1)$ on the upper scale represents a response style in which the fourth and fifth categories are more likely to be selected.

We evaluate the proposed method and compare its performance to existing methods using a simulation study and an empirical example in Sections 2.4 and 2.5, respectively.

2.2 Formalizing response functions

To describe the proposed methodology, a new framework is first introduced to formalize the concept of a "response function". Herein, response styles and corrected values are defined more rigorously than in previous studies by van de Velden (2008) and Schoonees et al. (2015). This framework can be used more generally when dealing with preference data possibly contaminated by response style effects. The relationship between our framework and CDS is elaborated on in Section 2.3.4.

2.2.1 Category boundaries in preference data

Response style problems occur when the interpretation of the preference categories differs for different respondents. For example, with 5-point scale data, if a respondent has an acquiescence response style, that is, a tendency to agree with items regardless of item content, the third category indicates a low preference of the respondent for that item, even though that category is the midpoint of the scale.

To express this formally, let $x_{ij} \in \{1, \ldots, q\}$ denote the q scale preference data provided by the ith respondent for the jth item, $(i = 1, \ldots, n\,;\, j = 1, \ldots, m)$. Suppose the observed

preference data x_{ij} are related to the true preference data $x_{ij}^* \in \mathbb{R}$ as follows:

$$x_{ij} = \sum_{\ell=1}^{q} \ell \, I\{\eta_{i(\ell-1)} < x_{ij}^* \leq \eta_{i\ell}\}$$

where $I\{\cdot\}$ is an indicator function, and $\eta_{i\ell}\,(\ell = 0, \ldots, q)$ are respondent-specific boundaries. We refer to the set of boundaries $b_\ell\,(\ell = 0, \ldots, q)$, which are equal for all respondents and are spaced equally, as reference boundaries. In this paper, we consider a bounded interval, that is, $\eta_{i0} = b_0 = \upsilon_1$ and $\eta_{iq} = b_q = \upsilon_2$.

Using these notations, "response-style-biased data" are data for which the true preferences x_{ij}^* are categorized based on equally-spaced reference boundaries b_ℓ even though each respondent has respondent-specific boundaries $\eta_{i\ell}$. This process is illustrated in Figure 2.1.

In Figure 2.1, respondent i has true preference x_{ij}^* and boundaries $\eta_{i\ell}\,(\ell = 1, \ldots, q-1)$ as shown on the upper scale. The aim is to "estimate" x_{ij}^* from x_{ij}. In this example, the observed preference is $x_{ij} = 3$. If we ignore the possibility that each respondent has different boundaries and simply assume that the reference boundaries are used as shown on the lower scale in Figure 2.1, a rough estimation of x_{ij}^*, say $\hat{x}_{ij}^*$, would be far from the true one x_{ij}^*. This indicates that depending on the unobservable respondent-specific boundaries, we obtain a bias from the true x_{ij}^*.

2.2.2 response functions

To correct for response-style-biased data, we introduce a definition of a response function in more rigorous way than previous studies as follows.

Definition 2.2.1. *Response function*
Suppose reference boundaries $b_\ell\,(\ell = 1, \ldots, q-1)$ and respondent-specific boundaries $\eta_{i\ell}\,(\ell = 1, \ldots, q-1)$ are given. Let both boundaries be monotonically increasing for ℓ. Then

$$\phi_i : b_\ell \longmapsto \eta_{i\ell}, \quad (\ell = 1, \ldots, q-1),$$

is defined as the response function for respondent i.

From this definition, it follows that ϕ_i is a monotonically increasing function. In addition, I assume that the response function is continuous. For later purposes, it is useful to specifically define the response function corresponding to the absence of a response style:

Definition 2.2.2. *No response style*
If $\eta_{i\ell} = b_\ell\,(\ell = 1, \ldots, q-1)$, we say that respondent i has no response style.

If ϕ_i is known for all respondents, we can use it to correct response-style-biased data, and to interpret respondent's response styles.

Definition 2.2.3. *Correcting preference data using the response functions*
Given q scale preference data x_{ij} with reference boundaries $b_1, \cdots, b_{q-1}$, and a response function ϕ_i, when $x_{ij} = \ell$, the corrected value of x_{ij} is

$$y_{ij} = \phi_i(\tau(\ell)), \qquad \text{where} \;\; \tau(\ell) \in (b_{\ell-1}, b_\ell].$$

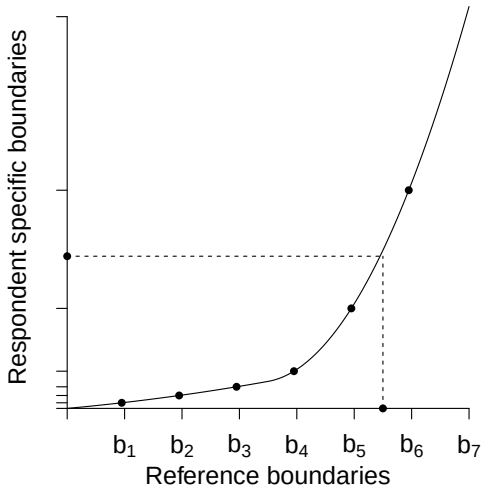

Figure 2.2: Example depicting how the observed value, $x_{ij} = 6$, corresponds to the corrected value. The solid line indicates the response function, ϕ_i. The horizontal axis represents the reference boundary (scale), while the vertical axis represents the respondent-specific boundary.

This definition indicates that the corrected value of x_{ij} is defined as the product of the transformation of some value between $b_{\ell-1}$ and b_ℓ, $\tau(\ell)$, according to ϕ_i. In this paper, as in CDS, we fix $\tau(\ell) = (b_\ell + b_{\ell-1})/2$. As this definition implies, in this paper the estimated value of x^*_{ij} from x_{ij} using ϕ is considered as a corrected value.

Figure 2.2 illustrates how a response function can be used to correct for response style bias. Suppose that we want to know x^*_{ij} when the observed rating is $x_{ij} = 6$ on a 7-point scale. In this case, the argument of ϕ_i can be any value in the interval $(b_5, b_6]$. Following Definition 2.2.3, we use the midpoint of the interval, and call it the representative value of category 6. If we set $b_\ell = \ell, (\ell = 1, \ldots, q-1)$, 5.5 (i.e., the point on the horizontal axis in Figure 2.2) will be the argument of ϕ_i. Assuming that the true response function is continuous, the output value of the response function corresponding to the representative value of category (i.e., the point on the vertical axis in Figure 2.2), can be read (i.e., interpolated) of the vertical axis. The resulting value, y_{ij} in this case, is the corrected value.

Response functions can be used to interpret the respondents' response styles. Figure 2.3 shows examples of typical response functions corresponding to respondents who have no, acquiescence, disacquiescence (a tendency to disagree), midpoint, or extreme response styles.

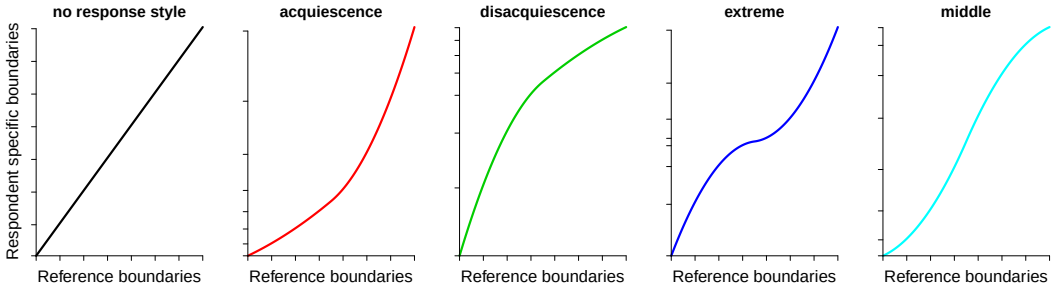

Figure 2.3: Response functions. The horizontal axis represents the reference boundary (scale), while the vertical axis represents the respondent-specific boundary.

2.3 Correcting and clustering preference data in the presence of response style bias

Based on the ideas and definitions introduced in Section 2.2, we consider estimation of respondent-specific response functions. Moreover, we show that the estimated response functions can be used to correct for response style bias and, at the same time, to find clusters of respondents based on their corrected item preferences. In this paper, these response tendencies shown in Figure 2.3 are considered as response styles, and it is assumed that there are no respondents having response-style-like preference (e.g, there are no respondents whose true responses agree with all items). In addition, it is assumed that categories in all items to be applied to CCRS have the same direction (e.g., a category indicating "agree" has a high number in all items).

2.3.1 Modeling response functions

To estimate a response function, data that represent respondent-specific boundaries are required. Here, similar to dual scaling and CDS, we code the preference data as "rank-ordered boundary data". This means that the indicated item preferences and the reference boundaries are converted to rank-orders for each respondent. The obtained boundary rankings reflect respondents' tendencies to select certain rating categories.

Suppose that q scale preference data $\boldsymbol{X} = (x_{ij})\,(i = 1, \ldots, n\,;\; j = 1, \ldots, m)$ are given with the reference boundaries $b_1, b_2, \cdots b_{q-1}$. Then, the rank-ordered boundary data $f_{i\ell}$, $(\ell = 1, \ldots, q-1)$ can be obtained as follows.

$$f_{i\ell} = \sum_{t=1}^{m+q-1} \left(I\{\xi_{it} < b_\ell\} + \frac{1}{2} I\{\xi_{it} = b_\ell\}\right) - \frac{1}{2}$$

$$\text{where} \quad \xi_{it} = \begin{cases} \frac{b_\ell + b_{\ell-1}}{2} & (t = 1, \ldots, m,\ x_{it} = \ell) \\ b_{t-m} & (t = m+1, \ldots, m+q-1) \end{cases} \tag{2.3.1}$$

For $t = 1, \ldots, m$, ξ_{it} indicate the representative values of a category, in our case, $(b_\ell + b_{\ell-1})/2$. On the other hand, for $t = m+1, \ldots, m+q-1$, ξ_{it} indicate reference boundaries.

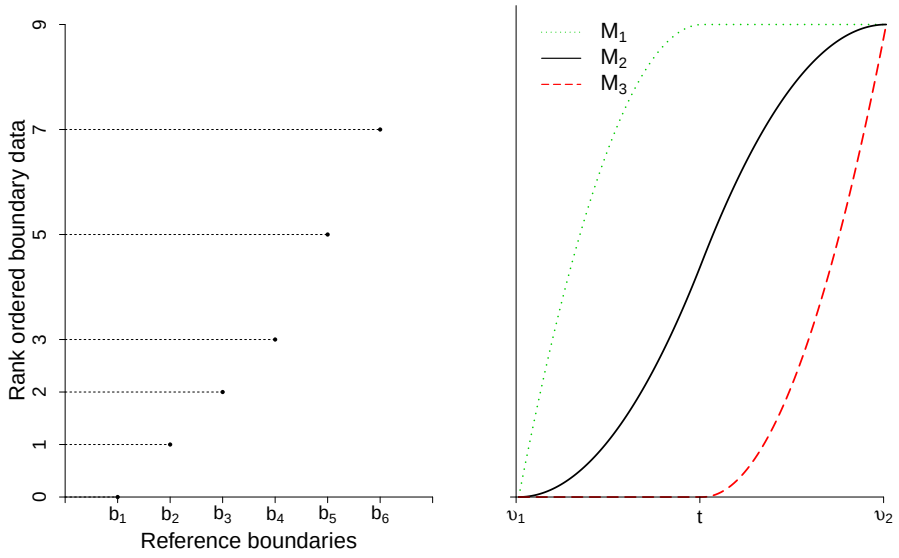

Figure 2.4: (Left) An example of rank-ordered boundary data. The horizontal axis corresponds to reference boundaries, the vertical axis shows $f_{i\ell}$ values corresponding to each boundary. Each dot represents, $f_{i1}, \ldots, f_{i6}$. (Right) Three I-spline basis functions. $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ are shown with solid, dot and dashed line, respectively.

The same idea was used in CDS (constrained dual scaling) (Schoonees et al., 2015), in which the use of this idea followed from dual scaling for successive data (Nishisato, 1980).

To illustrate how this works in practice, consider 7-point scale preference data, $\boldsymbol{x}_i = (5, 6, 7)$, is given. Using Equation (2.3.1), we obtain $\boldsymbol{\xi}_i = (4.5, 5.5, 6.5, 1, 2, 3, 4, 5, 6)$, where $\boldsymbol{\xi}_i = (\xi_{it}), (t = 1, \ldots, m+q-1)$. Then, sorting and converting these to rank-orders (starting from 0) yields

$$\begin{array}{rccccccccc} \boldsymbol{\xi}_i^{sorted} = & \big(1 & 2 & 3 & 4 & 4.5 & 5 & 5.5 & 6 & 6.5\big) \\ \text{rank}: & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & 4 & \mathbf{5} & 6 & \mathbf{7} & 8 \end{array}$$

Since $\xi_{i4} = 1, \xi_{i5} = 2, \xi_{i6} = 3, \xi_{i7} = 4, \xi_{i8} = 5, \xi_{i9} = 6$ corresponds to rank $0, 1, 2, 3, 5$ and 7, respectively, we get $\boldsymbol{f}_i = (0, 1, 2, 3, 5, 7)$. Figure 2.4 (left) plots the $\boldsymbol{f}_i = (f_{i\ell})\,(\ell = 1, \ldots, 6)$ against these reference boundaries. Using this converted $\boldsymbol{f}_i$, we see that respondent i demonstrates an acquiescence response style. For example, for $f_{i1}, \ldots, f_{i4}$, the values increase one by one, which indicates that respondent i does not select categories between the first and fourth reference boundaries frequently (i.e., the respondent does not often assign a rating smaller than 4). On the other hand, there is a large gap between f_{i4} and f_{i6}, which indicates that categories between the fourth and sixth reference boundaries are often selected.

Using $f_{i\ell}$, we consider a model for response functions corresponding to Definition 2.2.1, using I-Spline basis functions. Let $\bar{f}_{i\ell} = f_{i\ell}/p$, where $p = m + q - 1$, so that $\bar{f}_{i\ell} \in [0, 1]$.

Also, from here on, we use $b_\ell = \ell/q$, $(\ell = 1, \ldots, q-1)$. In CCRS, $\bar{f}_{i\ell}$ is approximated as

$$\begin{aligned} &\bar{f}_{i\ell} \approx \phi_i^{CCRS}(\ell/q), \quad (i = 1, \ldots, n\,;\, \ell = 1, \ldots, q-1) \\ &\text{where} \quad \phi_i^{CCRS}(x) = \sum_{r=1}^{3} \beta_{ir} \mathcal{I}_r(x) \\ &\text{s.t.} \quad \sum_{r=1}^{3} \beta_{ir} = 1, \quad \beta_{ir} \geq 0 \quad (r = 1, 2, 3) \end{aligned} \tag{2.3.2}$$

Here, $\mathcal{I}_r$ $(r = 1, 2, 3)$ are I-Spline basis functions, and β_{i1}, β_{i2} and β_{i3} are the coefficients of $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{I}_3$, respectively. $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{I}_3$ are defined by

$$\begin{aligned} \mathcal{I}_1(x) &= \begin{cases} \frac{2t(x-\upsilon_1)-(x^2-\upsilon_1^2)}{(t-\upsilon_1)^2} & (\upsilon_1 \leq x < t) \\ 1 & (t \leq x \leq \upsilon_2) \end{cases} \\ \mathcal{I}_2(x) &= \begin{cases} \frac{(x-\upsilon_1)^2}{(t-\upsilon_1)(\upsilon_2-\upsilon_1)} & (\upsilon_1 \leq x < t) \\ \frac{(t-\upsilon_1)}{(\upsilon_2-\upsilon_1)} + \frac{2U(x-\upsilon_1)-(x^2-t^2)}{(\upsilon_2-t)(\upsilon_2-\upsilon_1)} & (t \leq x \leq \upsilon_2) \end{cases} \\ \mathcal{I}_3(x) &= \begin{cases} 0 & (\upsilon_1 \leq x < t) \\ \frac{(x-t)^2}{(\upsilon_2-t)^2} & (t \leq x \leq \upsilon_2) \end{cases} \end{aligned} \tag{2.3.3}$$

and $x \in [\upsilon_1, \upsilon_2]$, $t = (\upsilon_1 + \upsilon_2)/2$. Note that in this definition of I-spline functions, similar to Schoonees et al. (2015), we fix the number of order is 2, and use a single knot at the median of the given interval, as recommended by Ramsay (1988); Ramsay and Abrahamowicz (1989). For more general definition and its property, see, for example, Ramsay (1988).

In CCRS, we use $\upsilon_1 = 0$, $\upsilon_2 = 1$. Nonnegative conditions, $\beta_{ir} \geq 0\,(r = 1, 2, 3)$, are required for ϕ_i to be a monotone increasing function. See Section 2.3.5 for a more detailed justification of the rationale underlying the scaling of $[\upsilon_1, \upsilon_2]$, $f_{i\ell}$ and b_ℓ to $[0, 1]$ as well as the advantages of adding the constraint $\sum_{r=1}^{3} \beta_{ir} = 1$.

By using three I-spline basis functions (as shown in Figure 2.4, right), we can handle the five types of response styles shown in Figure 2.3. Further, in this model, only β_{i1}, β_{i2} and β_{i3} need to be considered to interpret the response styles. For example, a greater β_{i3} value indicates a stronger tendency to select high categories because it results in more weight being placed on $\mathcal{I}_3$, which alters the shape of function to be more similar to the shape of the response function corresponding to the acquiescence response style (shown in Figure 2.3).

Now we can define a new correction method. Using the model defined in Equation (2.3.2), the response function can be estimated by "smoothing" via the constrained least squares method. In other words, given a $q \times 1$ vector $\bar{\boldsymbol{f}}_i = (\bar{f}_{i\ell})$ and a $(q-1) \times 3$ matrix, $\boldsymbol{\mathcal{I}} = (\mathcal{I}_r(\ell/q))$ $(\ell = 1, \ldots, q-1\,;\, r = 1, 2, 3)$, $\boldsymbol{\beta}_i$ is obtained by minimizing

$$\sum_{i=1}^{n} \|\bar{\boldsymbol{f}}_i - \boldsymbol{\mathcal{I}}\boldsymbol{\beta}_i\|^2, \quad \text{s.t.} \quad \sum_{r=1}^{3} \beta_{ir} = 1, \quad \beta_{ir} \geq 0 \tag{2.3.4}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \beta_{i3})$. Using the estimated value of $\hat{\boldsymbol{\beta}}_i$, we can construct the "estimated" response function (see Definition 2.2.3), $\hat{\phi}(x) = \sum_{r=1}^{3} \hat{\beta}_{ir} \mathcal{I}_r(x)$. By transforming

all responses in the preference data $\boldsymbol{X}$ using $\hat{\phi}(x)$, we obtain a $(n \times m)$ "corrected data" matrix, where response style bias is removed. Note that our new correction method can be considered as a special case of the framework introduced in Section 2.2.

In order to cluster respondents based on content in corrected data matrix, content-based clustering, such as k-means clustering, can be applied to the corrected data. We shall refer to this type of analysis as CCRS tandem.

Sequentially applying two methods (smoothing and clustering) may not yield optimal results for the correction and content-based clustering as the criteria of correction and clustering are optimized separately (e.g., Arabie, 1994). Therefore, we propose a method to conduct these two procedures simultaneously.

2.3.2 CCRS: Correcting and clustering response-style-biased data

Simultaneous smoothing and clustering can be achieved by simply adding the two minimization criteria (e.g., Hwang, Dillon, & Takane, 2006). Let K be the number of content-based clusters. Then we define the objective function of CCRS as follows;

$$\psi(\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U} \mid \bar{\boldsymbol{F}},\, \boldsymbol{Z},\, \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3) = \lambda \sum_{i=1}^{n} \|\bar{\boldsymbol{f}}_i - \boldsymbol{\mathcal{I}}\boldsymbol{\beta}_i\|^2 + (1-\lambda) \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik} \|\boldsymbol{Z}_i \widetilde{\boldsymbol{\mathcal{I}}} \boldsymbol{\beta}_i - \boldsymbol{g}_k\|^2 \tag{2.3.5}$$

$$\text{s.t.} \quad \sum_{r=1}^{3} \beta_{ir} = 1, \quad \beta_{ir} \geq 0 \quad (r = 1, 2, 3\,;\, i = 1, \ldots, n)$$

where $\boldsymbol{B} = (\boldsymbol{\beta}_i)$, $\boldsymbol{G} = (\boldsymbol{g}_k)$, $\boldsymbol{U} = (u_{ik})$, $\bar{\boldsymbol{F}} = (\bar{\boldsymbol{f}}_i)$, $(i = 1, \ldots, n\,;\, k = 1, \ldots, K)$, and, $\boldsymbol{Z} = (\boldsymbol{Z}_i)$, $\boldsymbol{Z}_i = (z_{ij\ell})$ $(j = 1, \ldots, m\,;\, \ell = 1, \ldots, q)$. The first term in equation (2.3.5) is the smoothing term, and the second term is the content-based clustering term. Note that $\lambda \in [0, 1]$ weighs these two terms and needs to be determined prior to the analysis.

In the content-based clustering term, k-means clustering is performed on the corrected data, namely, $\boldsymbol{Z}_i \widetilde{\boldsymbol{\mathcal{I}}} \boldsymbol{\beta}_i = (\hat{y}_{ij})$ $(i = 1, \ldots, n\,;\, j = 1, \ldots, m)$. Specifically, the $q \times 1$ vector $\boldsymbol{z}_{ij} = (z_{ij\ell})\,(\ell = 1, \ldots, q)$ is a dummy vector that takes $z_{ij\ell} = 1$ if respondent i selects category ℓ for the jth item; otherwise, $z_{ij\ell} = 0$. $q \times 3$ matrix $\widetilde{\boldsymbol{\mathcal{I}}} = (\mathcal{I}_r(\tau(\ell)))$ $(\ell = 1, \ldots, q\,;\, r = 1, 2, 3)$ is a basis function matrix; however, unlike $\boldsymbol{\mathcal{I}}$, it takes the middle points of the boundaries as arguments to construct the corrected data in Definition 2.2.3. The $K \times 1$ vector $\boldsymbol{u}_i = (u_{ik})\,(k = 1, \ldots, K)$ is an indicator vector for the content-based cluster, where $u_{ik} = 1$ if respondent i belongs to the kth content-based cluster; otherwise, $u_{ik} = 0$. $\boldsymbol{G}$ is the $K \times m$ content-based cluster centroid matrix.

Choosing an appropriate value for λ is a complicated task as there is no clear criterion that can be used. In Section 2.4, we show how different values of λ affect the clustering results and, in Section 2.5, we propose a pragmatic approach to determine λ and K at the same time.

Technically both CCRS and the correction method defined in Equation (2.3.4) can be applied to any ordinal categorical data, if the data are assumed to be contaminated by the effect of response styles.

2.3.3 CCRS parameter estimation

Algorithm to estimate CCRS parameters

To obtain parameters $\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U}$, two operations, i.e., estimation of the response functions (estimation of $\boldsymbol{B}$) and content-based clustering (estimation of $\boldsymbol{G}$ and $\boldsymbol{U}$), are performed sequentially. For fixed $\boldsymbol{B}$, minimizing Equation (2.3.5) reduces to k-means clustering of the (response style corrected) data $\boldsymbol{Z}_i\widetilde{\boldsymbol{\mathcal{I}}}\boldsymbol{\beta}_i$ $(i = 1, \ldots, n)$. On the other hand, when $\boldsymbol{G}$ and $\boldsymbol{U}$ are fixed, solving for $\boldsymbol{B}$ is less trivial as this appears in both terms in Equation (2.3.5). However, minimizing Equation (2.3.5) with respect to $\boldsymbol{B}$ can be reduced to a simple constrained least squares problem as follows;

Proposition 2.3.1. *The objective function of* CCRS (2.3.5) *can be written as follows.*

$$\psi(\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U} \mid \bar{\boldsymbol{F}},\, \boldsymbol{Z},\, \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3) = \sum_{i=1}^{n} \left\| \begin{pmatrix} \sqrt{\lambda}\bar{\boldsymbol{f}}_i \\ (\sqrt{1-\lambda})\boldsymbol{G}'\boldsymbol{u}_i \end{pmatrix} - \begin{pmatrix} \sqrt{\lambda}\boldsymbol{\mathcal{I}} \\ (\sqrt{1-\lambda})\boldsymbol{Z}_i\widetilde{\boldsymbol{\mathcal{I}}} \end{pmatrix} \boldsymbol{\beta}_i \right\|^2$$

Proof.

$$\begin{aligned}
\psi(\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U} \mid \bar{\boldsymbol{F}},\, \boldsymbol{Z},\, \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3) &= \lambda \sum_{i=1}^{n} \|\bar{\boldsymbol{f}}_i - \boldsymbol{\mathcal{I}}\boldsymbol{\beta}_i\|^2 + (1-\lambda) \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik} \|\boldsymbol{Z}_i\widetilde{\boldsymbol{\mathcal{I}}}\boldsymbol{\beta}_i - \boldsymbol{g}_k\|^2 \\
&= \sum_{i=1}^{n} \|\sqrt{\lambda}(\bar{\boldsymbol{f}}_i - \boldsymbol{\mathcal{I}}\boldsymbol{\beta}_i)\|^2 + \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik} \|\sqrt{1-\lambda}(\boldsymbol{Z}_i\widetilde{\boldsymbol{\mathcal{I}}}\boldsymbol{\beta}_i - \boldsymbol{g}_k)\|^2
\end{aligned}$$

Note for any vector $\boldsymbol{a}' = (\boldsymbol{a}_1', \boldsymbol{a}_2')'$, $\boldsymbol{b}' = (\boldsymbol{b}_1', \boldsymbol{b}_2')'$, it can be shown

$$\|\boldsymbol{a}_1 - \boldsymbol{b}_1\|^2 + \|\boldsymbol{a}_2 - \boldsymbol{b}_2\|^2 = \left\| \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \right\|^2.$$

Using this and $\boldsymbol{g}_k = \boldsymbol{G}'\boldsymbol{u}_i$, the proposition can be verified immediately. □

Using this property, parameters in CCRS are estimated based on the following algorithm.

Step 1: Initialization. Set λ and a convergence criterion ε, randomly choose an initial value for $\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U}$, and set the number of iterations w to $w = 1$.

Step 2: Response function estimation. For fixed $\boldsymbol{G}, \boldsymbol{U}$, update $\boldsymbol{B}$ in such a way that Equation (2.3.1) is minimized with the constraint in Equation (2.3.2) (Haskell & Hanson, 1981).

Step 3: Content-based clustering. For fixed $\boldsymbol{B}$, update $\boldsymbol{G}, \boldsymbol{U}$ using the following formula.

$$\boldsymbol{g}_k = \frac{\sum_{i=1}^{n} u_{ik}(\boldsymbol{Z}_i\widetilde{\boldsymbol{\mathcal{I}}}\boldsymbol{\beta}_i)}{\sum_{i=1}^{n} u_{ik}}$$

$$u_{ik} = \begin{cases} 1 & (k = \underset{s \in \{1,\ldots,K\}}{\arg\min} \|(\sqrt{1-\lambda}(\boldsymbol{g}_s - \boldsymbol{Z}_i\widetilde{\boldsymbol{\mathcal{I}}}\boldsymbol{\beta}_i)\|^2 \\ 0 & (\text{others}) \end{cases} \qquad (i = 1, \ldots, n\,;\, k = 1, \ldots, K)$$

Step 4: Convergence test Compute $\psi^{(w)}$, the value of the objective function (2.3.5) using updated parameters and, for $w > 1$, if $\psi^{(w)} - \psi^{(w-1)} < \varepsilon$, terminate; otherwise, let $w = w + 1$ and return to Step 2.

Convergence of the algorithm is guaranteed because the objective function (2.3.5) is monotonically decreasing in subsequent steps. Note that in Step 1 of the algorithm, Initial values for $\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U}$ need to be selected. This can be done randomly, e.g., by randomly generating values from uniform distribution. Alternatively, one could consider initial values for $\boldsymbol{B}, \boldsymbol{G}, \boldsymbol{U}$ by solving $\boldsymbol{\beta}_i$ $(i = 1, \ldots, n)$ for the first term of Equation (2.3.5), that is, the optimal fitting of the response functions to the boundary data, and applying k-means to corrected data $\boldsymbol{Z}_i \widetilde{\boldsymbol{\mathcal{I}}} \boldsymbol{\beta}_i$ $(i = 1, \ldots, n)$ to obtain initial values for $\boldsymbol{G}, \boldsymbol{U}$. We shall refer to this type of initialization as CCRS tandem initialization.

Problem of local minimum in CCRS

In parameter estimation of CCRS, we apply k-means type algorithm, which is well-known for causing a serious local minimum problem. Though we proposed using "CCRS tandem initialization" above, this does not guarantees the global minimum. The commonly used approach to tackle with this problem is to run algorithm many times with different randomly generated initial values, and select the estimates which yields the minimum value of objective function among estimates obtained by each run.

Figure 2.5 shows that the value of optimized CCRS objective function over the number of algorithm runs. Note that this is monotone non-increasing because the initial value is fixed at each t th time ($t = 1, \ldots, T$, ; $T = 1, \ldots, 100$ in this Figure). That is, for example, the 1st, 2nd and 3rd initial values are the same both when the number of initial values is 3 ($T = 3$ at the horizontal axis), and when the number of initial values is 10 ($T = 10$ at the horizontal axis).

This suggests that with $\lambda = 0.2$, the result of CCRS parameter estimation is unstable, because until around $T = 40$, the optimized value is frequently decreased. On the other hand, with $\lambda = 0.8$, the optimized value of CCRS objective function does not change over 100 runs, except the first three runs. That is, this figure suggests that in this case, the estimation result does not change whether the number of runs is 4 or 100. This should be because with $\lambda = 0.2$, the weight on k-means term is bigger than the smoothing term, and thus the estimation result tends to be unstable similarly to k-means algorithm.

2.3.4 Correcting preference data in the presence of response style bias by CDS

Schoonees et al. (2015) used constrained dual scaling (CDS) to estimate a response function defined similarly as in Section 2.2. In dual scaling, which is equivalent to correspondence analysis when analyzing contingency tables (van de Velden, 2000), category quantifications are obtained such that the quantifications best capture variance in the data in low dimensional space. For the analysis of preference data, dual scaling aims to

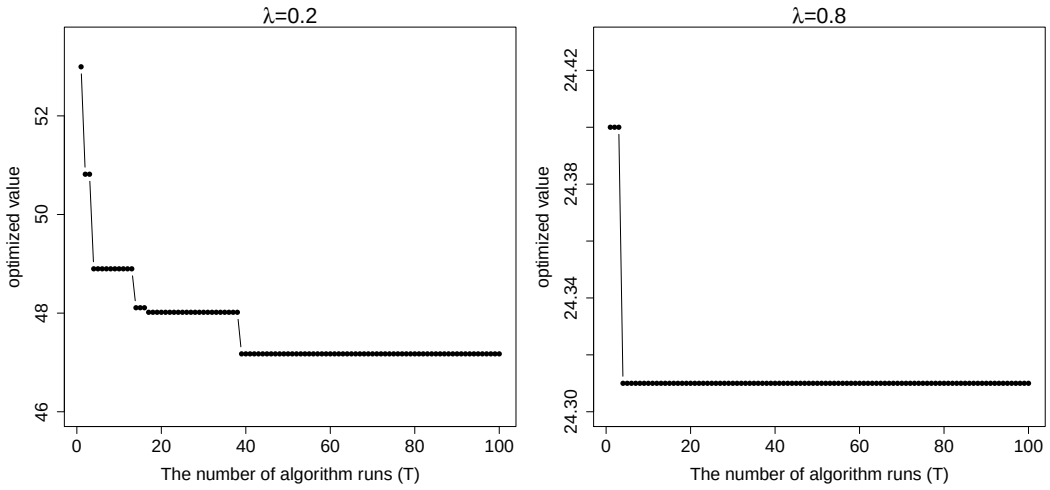

Figure 2.5: The graph of optimized value of CCRS objective function over the number of algorithm runs T with different initial values. That is, the horizontal axis T indicates how many times the algorithm runs with different random initial values ($t = 1, \ldots, T$), and the vertical axis indicates the minimum value of objective function among all T runs. In this numerical example, we fixed the initial values at the tth time of run for each $t = 1, \ldots, T$, for all $T = 1, \ldots, 100$, so that the randomness of initial values can be removed to investigate the stability of the algorithm. The artificial data used in this numerical example are with $n = 300$, $m = 10$, $D = 3$, $K = 3$ and $q = 5$. How to generate the artificial data is explained in later Section 2.4.1.

quantify respondents, items and boundaries. In particular, in CDS, one-dimensional quantifications for respondents and boundaries are obtained to model monotonically increasing response functions for clusters of respondents. Response style bias can then be corrected for in a manner similar to that described in Section 2.2. A sequential analysis where we first correct for response style effects using CDS, after which k-means is applied to the corrected data, can be seen as an alternative to the CCRS approach. We refer to such an approach as CDS tandem analysis.

As CDS is based on dual scaling, there are several restrictions. To explain this in detail, let v_i and $w_{d\ell}$ denote quantified values by CDS for respondent i, and the ℓth boundary for the dth response-style-based cluster ($d = 1, \ldots, D$), respectively. In addition, suppose that a respondent i belongs to the dth response-style-based cluster. In CDS, $w_{d\ell} = \phi_d^{CDS}(\ell)$, where ϕ_d^{CDS} is the CDS response function for the dth response-style-based cluster. Then, ϕ_d^{CDS} approximates the rank ordered boundary data $f_{i\ell}$ as

$$\tilde{f}_{i\ell} \approx v_i \phi_d^{CDS}(\ell), \quad (i = 1, \ldots, n\,;\, \ell = 1, \ldots, q-1) \tag{2.3.6}$$

$$\text{where} \quad \phi_d^{CDS}(x) = \mu_d + \sum_{r=1}^{3} \alpha_{dr} \mathcal{I}_r(x)$$

$$\text{s.t.} \quad \alpha_{dr} \geq 0, \quad (r = 1, 2, 3)$$

and $\tilde{f}_{i\ell} = f_{i\ell} - p/2$, where $p = m + q - 1$. For the spline basis function $\mathcal{I}_r$ in CDS, υ_1 and υ_2 are set to 0 and q (rather than 0 and 1 as is the case in CCRS) respectively. For more details, see Schoonees et al. (2015).

Comparing Equation (2.3.6) with Equation (2.3.2), it is clear that CDS only estimates response functions for response-style-based clusters $d = 1, \ldots, D$. Hence, due to the one-dimensional approximation only one parameter v_i ($i = 1, \ldots, n$) in Equation (2.3.6) is respondent-specific. Therefore, estimating response functions in CDS could incur a significant loss of respondent-specific information.

Note that, by setting $D = n$ and fixing the cluster indicator, CDS may be used to estimate respondent-specific $\boldsymbol{\alpha}_d$ ($d = 1, \ldots, n$) values. However, in practice, this process only yields degenerate solutions in which the parameters are zero or close to zero due to the one-dimensional reduction.

2.3.5 Properties and interpretation of CCRS

In addition to yielding content based clusters, CCRS can provide several insights into response styles. In particular, the constraint, $\sum_{r=1}^{3} \beta_{ir} = 1$, the lack of a constant term, and the scaling of the range of $f_{i\ell}$ and boundaries b_ℓ to $[0, 1]$ are useful for two reasons: first, these constraints restrict the corrected data to $[0, 1]$ for all respondents and items. Second, these constraints facilitate a straightforward visualization of response styles.

The range of corrected data

Proposition 2.3.2. *Let*

$$\phi_i(x) = \sum_{r=1}^{3} \beta_{ir} \mathcal{I}_r(x), \quad x \in [0,1]$$
$$\text{where} \quad \beta_{ir} \geq 0 \quad (r = 1,2,3\,;\, i = 1,\ldots,n)$$

be a monotone response function of respondent i. Imposing the constraint $\sum_{r=1}^{3} \beta_{ir} = 1$ is equivalent to imposing

$$\phi_i(0) = 0, \quad \phi_i(1) = 1.$$

Equivalently,

$$\hat{y}_{ij} \in [0,1]$$
$$\text{where} \quad \boldsymbol{Z}_i \widetilde{\boldsymbol{\mathcal{I}}} \boldsymbol{\beta}_i = (\hat{y}_{ij}) \ (i = 1,\ldots,n\,;\, j = 1,\ldots,m)$$

Proof. The proposition follows immediately from Equation (2.3.3) by setting $\upsilon_1 = 0$ and $\upsilon_2 = 1$. □

In other words, the constraint $\sum_{r=1}^{3} \beta_{ir} = 1$ implies a constraint on the range of ϕ_i, and, as a result, a constraint on the range of the corrected data, $\hat{y}_{ij}$. This is useful for avoiding excessive values for β_{ir}. If respondent i could receive a very large β_{ir} for some r, the corrected data $\hat{y}_{ij}$ $(j = 1,\ldots,m)$ would also become quite big, and as a result, respondent i would be considered as an outlier in the cluster analysis. However, large values for β_{ir} do not necessarily indicate that a respondent i is an outlier with respect to item preferences, even though the observation could be considered to be an outlier with respect to response styles. Thus, the summation constraints prevents the corrected values to be affected by strong response style effects.

Visualization of response styles

Constraining $\beta_{ir}\,(r = 1,2,3)$ to a sum of 1 allows for a simple visualization of these coefficients. Such a visualization can be used to interpret the respondent-specific response tendencies. In particular, by combining a scatterplot of the respondent-specific estimates of β_{i1} against β_{i3}, we obtain a visualization of the estimated response functions. Figure 2.6 illustrates this for an example dataset. Note that respondents having no response style (Definition 2.2.2) can be expressed as the single cross point in this plot, as indicated in the following proposition.

Proposition 2.3.3. *Let*

$$\phi_i(x) = \sum_{r=1}^{3} \beta_{ir} \mathcal{I}_r(x), \quad x \in [0,1]$$

be the true response function of respondent i, and suppose $\beta_{ir} \geq 0\,(r = 1,2,3)$, $\sum_{r=1}^{3} \beta_{ir} = 1$. Then respondent i has no response style, if and only if

$$\beta_{i1} = \beta_{i3} = 0.25, \quad \beta_{i2} = 0.5.$$

Proof. First, we show that having no response style $\Longrightarrow \beta_{i1} = \beta_{i3} = 0.25$. From Definition 2.2.2, having no response style means having the identity function as response function. In that case,

$$\frac{\partial^2}{\partial x^2}\phi_i(x) = 0.$$

On the other hand, when $\upsilon_1 = 0$ and $\upsilon_2 = 1$, from Equation (2.3.3) it follows that

$$\frac{\partial^2}{\partial x^2}\phi_i(x) = \begin{cases} -8\beta_1 + 4\beta_2 & (0 \leq x < 1/2) \\ -4\beta_2 + 8\beta_3 & (1/2 \leq x \leq 1) \end{cases}$$

Therefore having no response style implies $\beta_1 = 2\beta_2$ for $0 \leq x < 1/2$, and $\beta_3 = 2\beta_2$ for $1/2 \leq x \leq 1$. Since $\beta_{ir}\,(r = 1, 2, 3)$ is common for all $x \in [0, 1]$,

$$2\beta_{i1} = \beta_{i2} = 2\beta_{i3}.$$

From the constraint $\sum_{r=1}^{3} \beta_{ir} = 1$, the result immediately follows.

Next, to proof that $\beta_{i1} = \beta_{i3} = 0.25 \Longrightarrow$ having no response style, note that from the constraint $\sum_{r=1}^{3} \beta_{ir} = 1$ it immediately follows that $\beta_{i2} = 0.5$. Then, substituting $\beta_{i1} = \beta_{i3} = 0.25$ and $\beta_{i2} = 0.5$ into Equation (2.3.2) yields

$$\frac{\partial}{\partial x}\phi_i(x) = 1, \;\; \frac{\partial^2}{\partial x^2}\phi_i(x) = 0, \quad x \in [0, 1].$$

Hence, $\phi_i(x)$ is an identity function.

□

From this proposition, it follows that the purple cross in Figure 2.6, with coordinates $(0.25, 0.25)$, corresponds to no response style. The black points close to this purple point also correspond to respondents who do not have clear response style and deviations from this point indicate the presence of response styles.

2.4 Simulation study of CCRS

We conducted a simulation study to evaluate the performance of CCRS. In this simulation study, we investigated two things:

- the accuracy of correction comparing our CCRS correction defined in Equation (2.3.4) with CDS correction.
- the accuracy of content-based clustering comparing our CCRS in Equation (2.3.5) with k-means and CDS tandem.

Note that in CDS tandem, preference data are first corrected using CDS. Then, k-means is applied to the corrected data.

To assess the performance of the methods, we consider two scenarios. In scenario I, we assume that there are two kinds of underlying clustering structures: content and response-style-based clusters. In scenario II, only an content-based clustering structure is assumed. By considering these two scenarios, data are generated corresponding to situations that are assumed to underlie, either implicitly or explicitly, both the CDS and the CCRS methods.

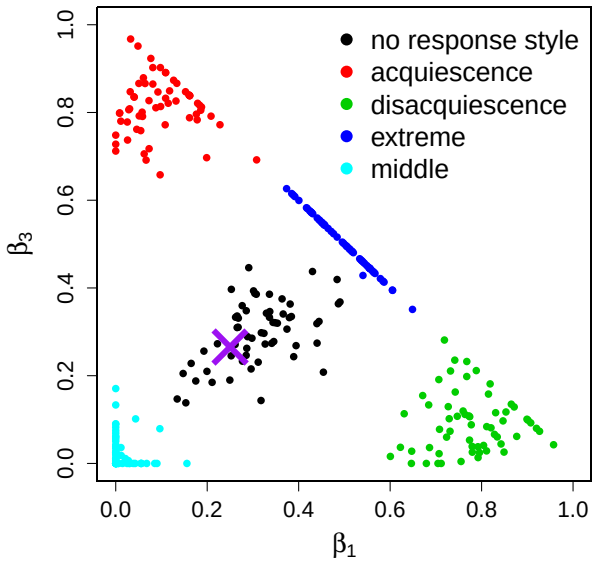

Figure 2.6: A scatterplot of β_{i1} and β_{i3} for each respondent ($i = 1, \ldots, n$) estimated by CCRS using $\lambda = 0.8$ for a simulated data set with $n = 300$, $m = 20$, $K = 2$, $q = 7$. The colors correspond to the true response styles. The way to generate these data is explained in Section 2.4.

2.4.1 Data generation

The data generation process can be divided into two steps: (i) generation of true preferences $x^*_{ij} \in \mathbb{R}$ and (ii) mapping of the true preferences to q scale data $x_{ij} \in \{1, \ldots, q\}$. Content-based clusters and, for scenario I only, response-style-based clusters, are induced in steps (i) and (ii), respectively.

(i) Generation of the true preferences

As we want a subset of items to be related to the clustering structure, the m items are divided into two groups: items related to the clustering structure and "noisy" items that are unrelated to the clusters.

In addition, the cluster-related items are divided further into three groups with different means of true preferences to ensure that the content-based clusters do not resemble either of the response-style-based clusters shown in Figure 2.7 (left). To see why this is useful, consider a situation in which all cluster-related items have one common cluster center. The corresponding content-based cluster could then be considered a response-style-based cluster corresponding to acquiescence, disacquiescence, or midpoint depending on the mean (e.g., if the means for all cluster-related items are high, it could be seen as an acquiescence response-style-based cluster). Furthermore, if all items have two centers only, the resulting cluster could be considered a response-style-based cluster corresponding to the extreme response style (e.g., if the means for the two item groups is extremely either high or low). Thus, both possibilities are avoided by dividing the cluster-related items into three groups

and generating preferences from the item groups depending on the means.

Next, for the items related to the clustering structures, true preference data are generated as $x^*_{ij} \sim N(\xi_{jk}, 0.1)$, where $\xi_{jk} \sim N(v_{sk}, 0.01)$ for $k = 1, \ldots, K$. Here, v_{sk} for all $s = 1, 2, 3$ and $k = 1, \ldots, K$ are randomly selected from the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ under the conditions that $v_{sk} \neq v_{s\ell}$ for $k, \ell = 1, \ldots, K$, $s = 1, 2, 3$ and $k \neq \ell$, and similarly, $v_{sk} \neq v_{tk}$ for all $s \neq t$. In other words, x^*_{ij} is generated based on the cluster mean, v_{sk}, which differs depending on the three cluster-related item groups, s, and the content-based clusters, k. For noisy items, i.e. the items unrelated to the clusters, the true preference data are generated as $x^*_{ij} \sim U(0, 1)$.

(ii) **Categorization of true preference data into q scale data**

The true preference data are categorized based on respondent-specific boundaries. Note that this step differs between scenarios I and II.

First, for scenario I, we assume that there exist D response-style-based clusters $d = 1, \ldots, D$. If respondent i belongs to the dth response-style-based cluster, x_{ij} is obtained according to the following rule.

$$x_{ij} = \sum_{\ell=1}^{q} \ell \, I\{\eta_{d(\ell-1)} < x^*_{ij} \leq \eta_{d\ell}\}.$$

Here, $\eta_{d\ell}$ are the dth response-style-based cluster specific boundaries obtained using true response functions, as shown in Figure 2.7 (left).

On the other hand, for scenario II, x_{ij} is obtained according to the following rule.

$$x_{ij} = \sum_{\ell=1}^{q} \ell \, I\{\eta_{i(\ell-1)} < x^*_{ij} \leq \eta_{i\ell}\}.$$

In this case, $\eta_{i\ell}$ are generated using $\beta_{i1}, \beta_{i2}, \beta_{i3}$, which were randomly generated from $U(0, 1)$ and scaled so that $\sum_{r=1}^{3} \beta_{ir} = 1$.

Figure 2.7 (right) shows an example of how the artificial data described above is generated. Here, preference data with $K = 2$, $m = 20$ are assumed. The horizontal axis indicates the item index, where an item set with $m = 20$, $\{1, \ldots, 20\}$ is divided such that items $\{1, \ldots, 4\}$, $\{5, \ldots, 8\}$, and $\{9, \ldots, 12\}$ are related to clusters whereas items $\{13, \ldots, 20\}$ are noisy items. The figure shows that the mean preferences (v_{sk}) are $v_{11} = 0.3$, $v_{12} = 0.1$, $v_{21} = 0.7$, $v_{22} = 0.5$, $v_{31} = 0.9$, and $v_{32} = 0.7$; note that the v_{sk} values were determined randomly under the condition that there is no overlap between both the cluster-related item groups and the content-based clusters.

2.4.2 Simulation study design

We consider a full factorial design with $n = 300$, 600, $m = 20$, 30, $q = 5$, 7, and $K = 2, \ldots, 5$, to assess the performance of the methods in different settings. To evaluate the correction accuracy, we only evaluate $m = 20$ case, because unlike k-means clustering, the correlation among variables do not contribute to the performance of both correction methods. In addition, for scenario I, in which response-style-based clusters exist, we

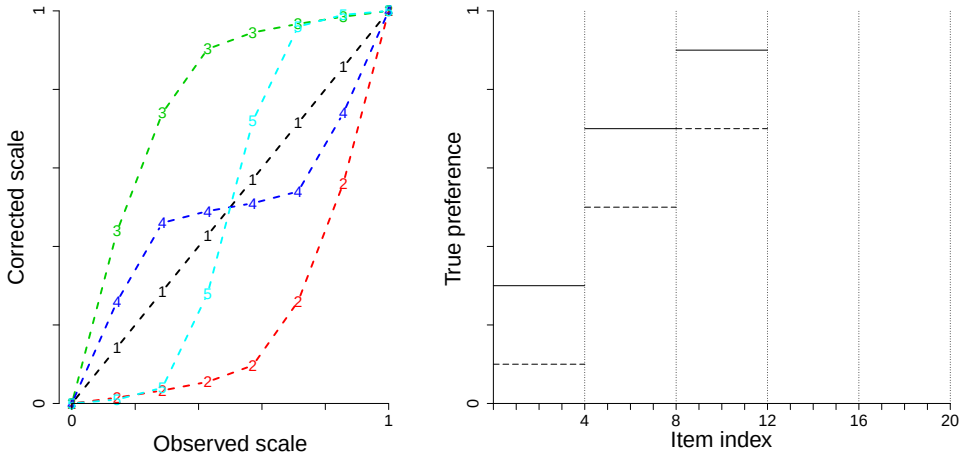

Figure 2.7: (Left) True response functions used in simulation. Lines 1 through 5 correspond to respondents having no response style, acquiescence, disacquiescence, extreme, and midpoint response styles, respectively. (Right) Example image of the generation of artificial data. Item indices are on the horizontal axis and corresponding true preferences, x^*_{ij}, on a scale of $[0, 1]$, are on the vertical axis. The mean preferences, v_{sk} $(s = 1, 2, 3, k = 1, 2)$, for the two clusters are indicated by the solid and dotted lines.

generated data with $D = 3$ and 5, where D corresponds to the number of response-style-based clusters. For $D = 3$, the considered response styles are acquiescence, midpoint, and no response style. For $D = 5$, disacquiescence and extreme response styles are added. In this simulation, we assume that the true number of content and response-style-based clusters K and D are known.

For each combination of parameters in our simulation, we randomly generated 100 different data sets. For each data set we apply all methods and assess, both for the content and response-style-based clustering.

Evaluation

To evaluate the accuracy of correction, the median of sum of squared error (SSE) between $\hat{y}_{ij}$, the corrected value, and x^*_{ij}, the corresponding true preference, is calculated. Here we use the median instead of mean of SSE, because the variance of SSE by CDS tends to be large in some situations.

To evaluate the content-based clustering, the accuracies of CCRS, CDS tandem, and k-means clustering are compared. On the other hand, for the response-style-based clustering, the accuracy of CDS is compared to that of CCRS with k-means clustering applied to the estimated $\boldsymbol{\beta}_i$ values $(i = 1, \ldots, n)$. The Adjusted Rand Index (ARI) is used to evaluate the retrieval of the underlying structure (Hubert & Arabie, 1985). The ARI assesses the similarity between two cluster allocations (a true and estimated cluster allocation, in this case). It takes a value of 1 for a perfect recovery, and this value decreases as performance worsens.

Selecting the number of response-style-based clusters for CDS

CDS requires a choice for the number of response-style-based clusters D. In scenario Ⅱ, no response style clusters exist and we therefore need to find an estimate for this. Schoonees et al. (2015) use a scree plot of the optimized objective function over different D. However, this approach cannot be used when we want to compare the results from different methods. Therefore, both in our simulation and empirical study, we use the Krzanowski-Lai cluster index (KL index) (Krzanowski & Lai, 1988) to determine the number of clusters. The KL index is based on an idea similar as the scree plot, but also takes into account the number of variables.

In the simulation study, we selected different D depending on different n and K values by first running a small simulation. For example, for $n = 300$ and $K = 2$ case, 10 data sets were simulated for each combination of $m = 20, 30$ and $q = 5, 7$. The KL index was calculated for the results obtained for each generated dataset. Among the resulting $2 \times 2 \times 10 = 40$ KL indexes, the most frequently selected D value was used as the number of response-style-based clusters for all settings with $n = 300$ and $K = 2$. This process was performed for all different n and K values. The result of this procedure was that $D = 3$ was selected for $n = 300$ and $K = 2, \ldots, 5$, and $D = 4$ was selected for $n = 600$ and $K = 2, \ldots, 5$. The resulting D are used to estimate CDS, for both evaluation of correction and clustering accuracy.

Other setting

Concerning the choice of λ in CCRS, we considered values of $0.2, 0.5, 0.8$ and compared the results of each λ. In addition, all methods require some type of initialization. For CDS we use the defaults from the "cds" package (Schoonees, 2016) in R (R Core Team, 2017); for k-means, we use 100 random starts; for the CCRS method, we consider the CCRS tandem initialization as well as 49 random starts.

2.4.3 Correction accuracy

The SSE results for correction of response styles are shown in Figure 2.8 for scenario I, and Figure 2.9 for scenario Ⅱ. As it can be seen, the accuracy of CCRS correction method is stable and higher than the one by CDS in all cases, regardless of the presence of content-based cluster structure. While CDS seems to have the difficulty of correcting for response styles especially when $D = 3$ and $q = 7$, our CCRS correction performs similarly well to all cases.

There are two possible reasons for this improvement of CCRS correction. At first, CDS conducts one dimensional reduction, which causes a serious information loss. In addition, in CDS, since the range of corrected values is $[0, \infty]$, the corrected value can be inflated if there exists the strong response style effects. On the other hand, the range of corrected values in CCRS correction is $[0, 1]$, due to the constraint as shown in Proposition 2.3.2. Therefore the scale of corrected value is not affected by the strong response style effect.

These results suggest that by generalizing the concept of response functions introduced

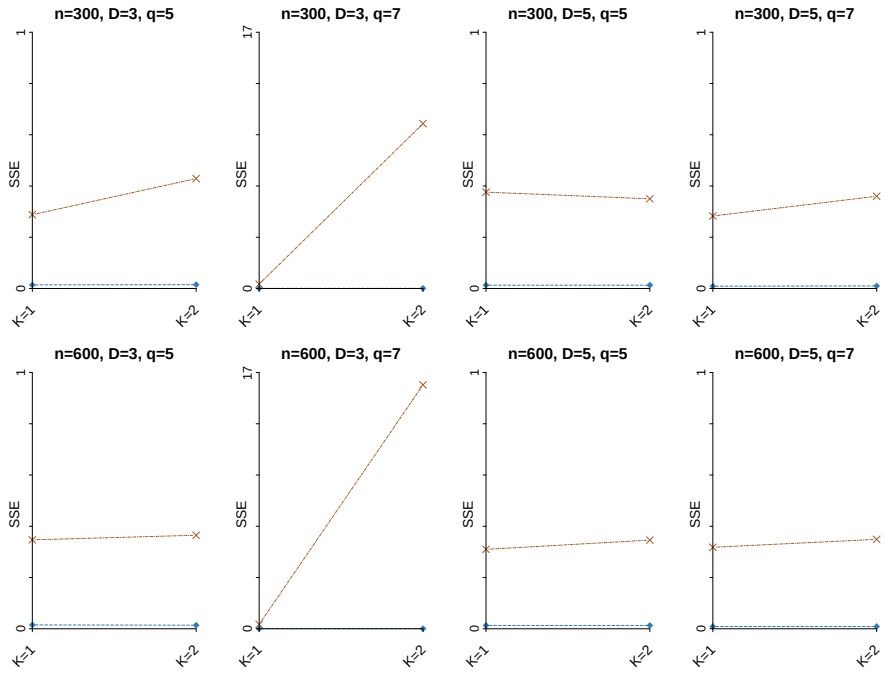

Figure 2.8: Parallel plot showing median SSE for correction for scenario I (presence of content and response-style-based clusters), $m = 20$ and different parameter settings and methods. The orange dashed line with a cross indicates SSE by CDS correction, while the blue dot line with a square indicates SSE by CCRS correction.

by Schoonees et al. (2015), we can obtain a new correction method which yields more accurate and stable correction performance.

2.4.4 Clustering accuracy

Scenario I : clustered response styles

Content-based clusters retrieval

The ARI results for the content-based clusters (content ARI) are shown in Figures 2.10 and 2.11. As can be seen, the k-means results are poor, possibly due to the presence of response style bias. However, CDS tandem, which does correct for response style bias, also demonstrates poor results. Apparently, the joint, but uncorrelated presence of content and response clusters, makes it difficult for CDS to detect the true content clustering structure.

CCRS tandem and 0.8 appear to work well compared to all other methods. A general tendency of the content ARI obtained by CCRS is that greater q and n, and smaller m, K and D values yield better results. Larger q values may yield good results as the estimation of the response function improves when there are more rating categories and hence more boundaries. Note also that the performance of CCRS does not appear to be strongly affected by an increase (from $D = 3$ to $D = 5$) in the number of response styles.

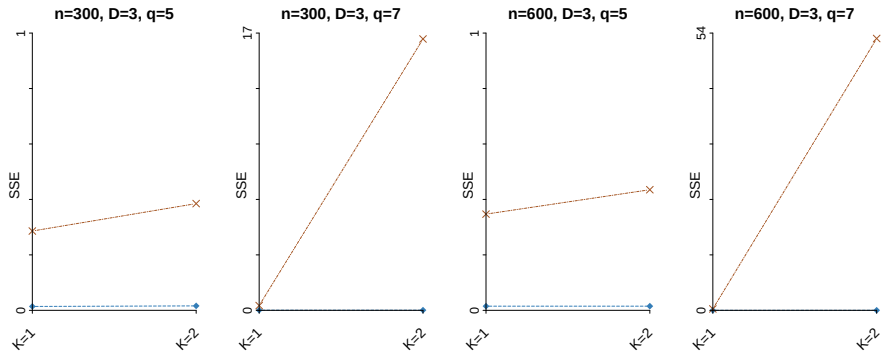

Figure 2.9: Parallel plot showing median SSE for correction for scenario II (no response style clusters).

Response-style-based clusters retrieval

The ARI results for the response-style-based clusters (response style ARI) are shown in Figures 2.12 and 2.13. As can be seen, the mean ARI's for CDS are always below those of CCRS. Furthermore, CCRS with $\lambda = 0.8$ outperforms the other methods in nearly all cases. Note that the response style ARI results for $\lambda = 0.8$ are generally better than those for CCRS tandem. An explanation for this could be that CCRS tandem only uses the boundary data $\bar{\boldsymbol{f}}_i$ to estimate the response functions, while simultaneous CCRS also exploits the underlying content related cluster structure in its estimation.

Scenario II: respondent-specific response styles

Content-based clusters retrieval

The ARI results concerning the content-based cluster structure are shown in Figures 2.14 and 2.15. As can be seen, there are no big differences with regard to scenario I, with the exception of the k-means results. The k-means results improved considerably compared to the results in scenario I. An explanation for this could be that the underlying content-based cluster structure in this scenario is no longer obscured by an additional (uncorrelated) response-style-based cluster structure. However, despite this improvement, CCRS still consistently outperforms k-means and appears to be useful to obtain content-based clusters even when no response-style-based clusters are present.

2.4.5 Conclusions of the simulation study

The results of the simulation study demonstrate that the proposed CCRS method outperforms CDS for correction in all cases, and CDS tandem, and k-means for clustering in all cases. Moreover, CCRS performed approximately equally well in both scenario I and scenario II.

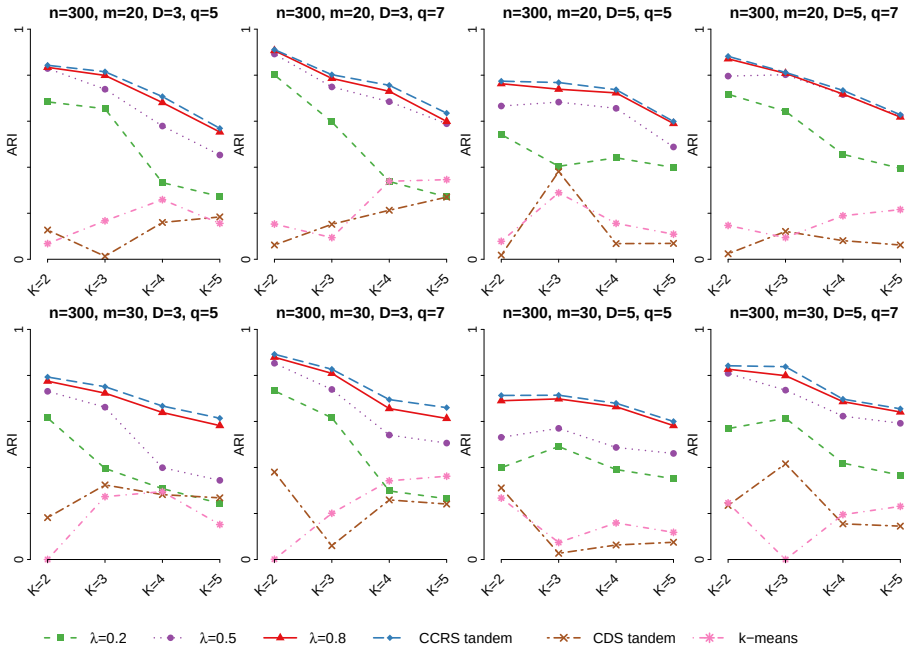

Figure 2.10: Parallel plot showing mean content ARI's for scenario I (presence of content and response-style-based clusters), $n = 300$ and different parameter settings and methods. $\lambda = 0.2, 0.5, 0.8$ indicate CCRS using each λ, where correction for response styles and clustering are simultaneously conducted.

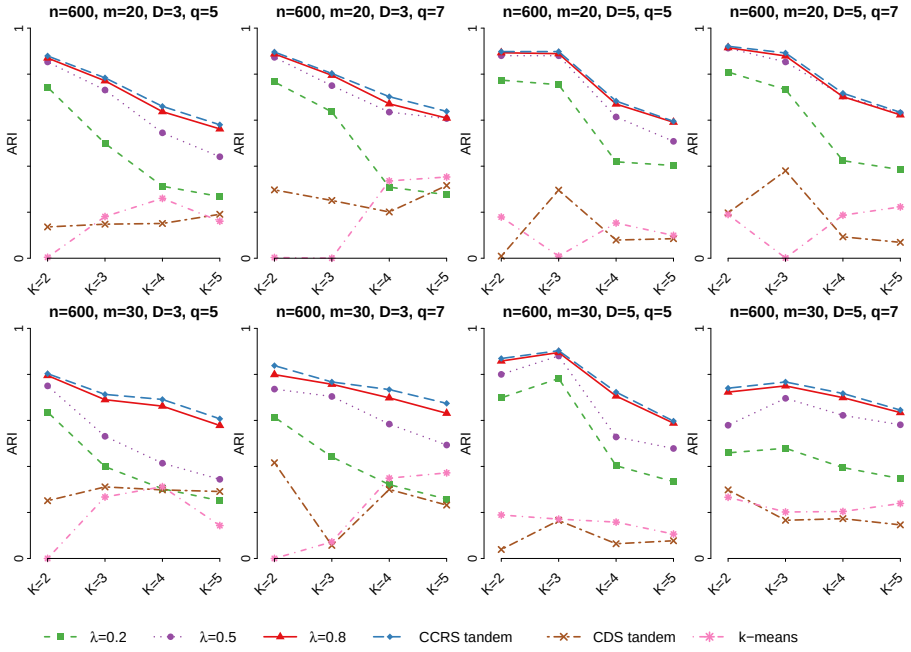

Figure 2.11: Parallel plot showing mean content ARI's for scenario I (presence of content and response style based clusters), $n = 600$ and different parameter settings and methods.

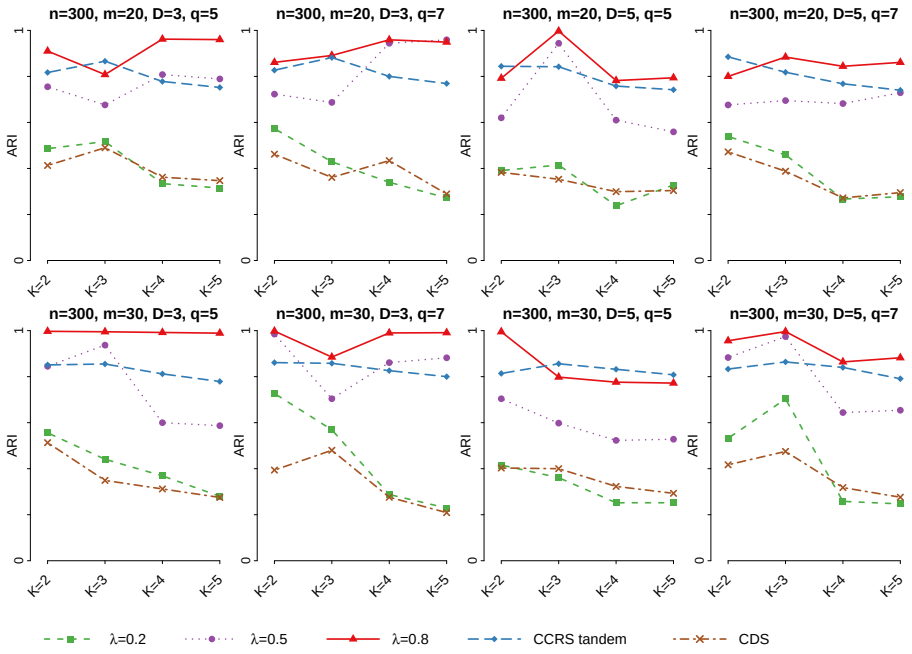

Figure 2.12: Parallel plot showing mean response style ARI's for scenario I (presence of content and response-style-based clusters), $n = 300$ and different parameter settings and methods.

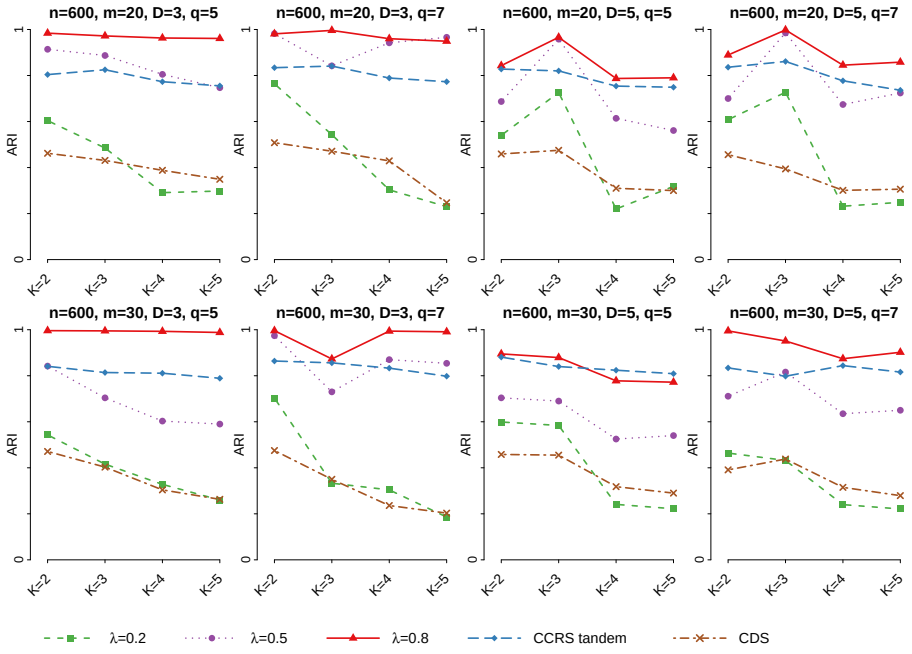

Figure 2.13: Parallel plot showing mean response style ARI's for scenario I (presence of content and response style based clusters), $n = 600$ and different parameter settings and methods.

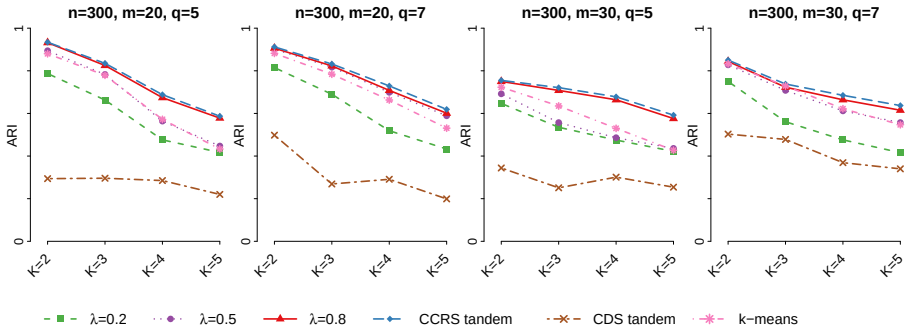

Figure 2.14: Parallel plot showing mean content ARI's for scenario II (no response style clusters) for $n = 300$.

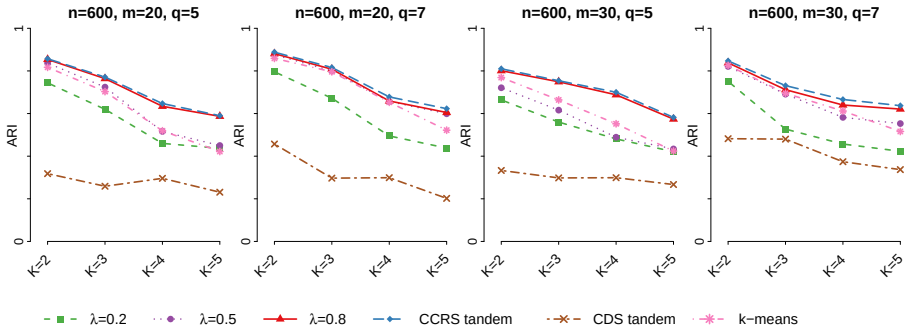

Figure 2.15: Parallel plot showing mean content ARI's for scenario II (no response style clusters) for $n = 600$.

For clustering evaluation, k-means clearly performed worse in scenario I. These results indicate that the proposed CCRS method appears to be robust to having both content and response-style-based cluster structures. Overall, CCRS performs better for greater q and smaller K. In addition, as the performance of CCRS does not appear to be strongly affected by an increase in response styles D, indicating that CCRS can account for more response styles.

The simulation study results showed that the content-based clustering results of CCRS improved when λ increased although differences between the cluster retrieval results for $\lambda = 0.8$ and CCRS tandem were very small. However, if a response-style-based clustering structure was present, this structure was better retrieved by selecting $\lambda = 0.8$. In addition, as is discussed in Section 2.3.3, the algorithm is more stable with high λ (e.g., $\lambda = 0.8$) than low λ, (e.g., $\lambda = 0.2$). Therefore, we suggest to use $\lambda \geq 0.8$, in order to obtain optimal results for both response style and content-based clustering.

2.5 Empirical example of CCRS

2.5.1 Data

Table 2.1: Value research selected items. Each statement is rated from 1 (strongly disagree) to 7 (strongly agree).

j	Statement	Value
1	It is more important for a wife to help her husband's career than to pursue her own career[1].	Patriarchy/Gender Role
2	The authority of father in a family should be respected under any circumstances.	Patriarchy/Gender Role
3	It is not desirable to oppose an idea which the majority of people accept, even if it is different from one's own.	Harmony
4	One should not express one's complaints about others in order to have good relationship with them.	Harmony
5	When hiring someone at a private company, even if an unacquainted person is more qualified, it would still be better to give the opportunity to relatives or friends.	In-Group Orientation
6	I would be honored when people who come from the same town play an important role in society.	In-Group Orientation
7	A subordinate should obey the superiors' instructions, even if the person cannot agree with them.	Hierarchy/Authority
8	If I have capable leaders, it is better to let them decide everything.	Hierarchy/Authority
9	A life full of risks and chances is more desirable than an ordinary and stable life.	Uncertainty Avoidance/Risk Taking
10	With extra money, I would invest in items for high returns even if they are risky.	Uncertainty Avoidance/Risk Taking

We illustrate the use of CCRS with an empirical application based on survey data collected in 2008 by the East Asian Social Survey (EASS). The survey data include 8745 respondents and 107 questions (except demographic variables). The data were downloaded from the ICPSR website (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34607). More information about the data is available in Chang, Iwai, Li, and Kim (2014). For our application, we selected 10 items from the survey in which respondents were asked to evaluate five values: Patriarchy/Gender Role, Harmony, In-Group Orientation, Hierarchy/Authority, and Uncertainty Avoidance/Risk Taking. For each of these values respondents were asked to assess two statements using a 7 point Likert scale ranging from 1: not important at all, to 7: very important. The 10 statements and corresponding values can be found in Table 2.1. we removed respondents having missing values in chosen 10 items, and as a result,

we have $n = 7634$ in data. In addition, the data contain respondents from four countries, 2838 Chinese, 1492 Japanese, 1448 Korean and 1856 Taiwanese.

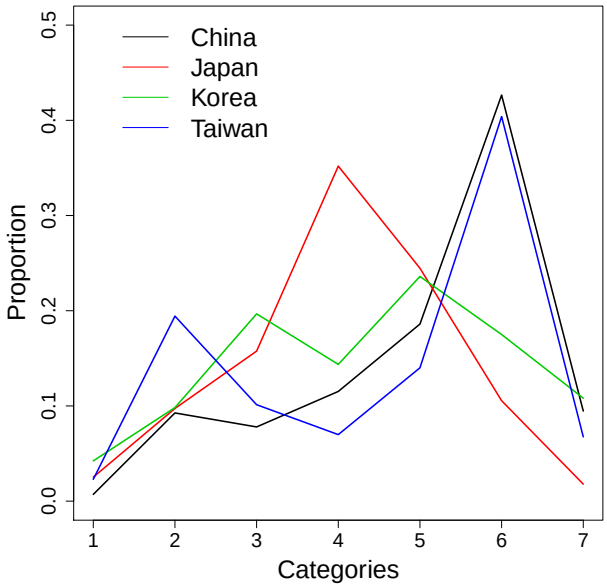

Figure 2.16: Proportion of rating categories 1-7 selected in different countries for all items.

In Figure 2.16, we see that there appears to be considerable difference in response tendencies among the four countries. For example, the Chinese and Taiwanese respondents selected the second highest category much more often than the Korean and Japanese respondents. Moreover, the Japanese respondents tended to select the midpoint more often than respondents from the other countries.

Below, we present the content and response-style-based clustering results obtained by the CCRS and CDS tandem methods, as well as the content clustering results obtained by k-means. For the response-style-based clustering in CCRS, k-means clustering was applied to the estimated $\boldsymbol{\beta}_i$ values ($i = 1, \ldots, n$), same as in the simulation study.

2.5.2 Setting

As these are empirical data, no known true clustering structure exist and all parameters need to be determined based on the data. Similar to the situation in cluster analysis, where selection of the number of clusters is a complex task, selection of such parameters in CDS and CCRS is difficult. In our application, we employed a pragmatic approach and based our selections on the KL index also used in simulation study. Furthermore, to ensure stability of the selected number of clusters, we based our choice on the results for 200 bootstrap samples. That is, from the complete sample we drew 200 bootstrap samples and, for each bootstrap sample, we selected the K value that maximized the KL index.

[1]For Japanese respondents, this statement was phrased differently, even though the same value was measured. Specifically, in the Japanese version the statement was: "A husband's job is to earn money; a wife's job is to look after the home and family."

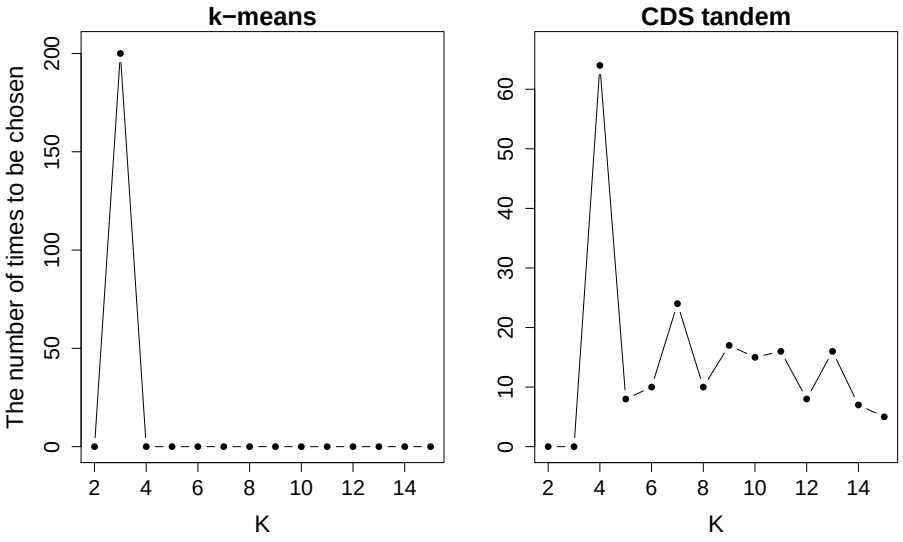

Figure 2.17: The number of times that different values of the parameter K was selected for content-based clustering methods using the KL index.

Next, the K value that was most often selected in these 200 samples was used in the final estimation. For the CCRS method, which requires two parameters (i.e., K and λ), the combination of parameters that was selected most frequently was used.

To reduce computation times, we used 20 different initial values for each CCRS run on bootstrap sample. In addition, for CDS tandem, K and D were selected sequentially. That is, first D is selected in a similar fashion as described above. Then, using the optimal D, K is determined in the same way.

Using the candidate values $K, D = 2, \ldots, 15$ and $\lambda = 0.7, 0.8, 0.9$ and CCRS tandem, we obtained $K = 7, \lambda = 0.8$ and $D = 6$ for CCRS, and $K = 4$ and $D = 3$ for CDS tandem, and $K = 3$ for k-means.

The total number of times each parameter was selected is shown in Figures 2.17, 2.18 and 2.19 for content and response-style-based clusters, respectively. Note that, for k-means, $K = 3$ was always selected. For the other methods, the value of K (or D) selected by the KL index varied among the bootstrap samples. However, a clear peak can be identified for most cases.

Once the parameters were set, the methods were applied using 500 different initial values in the same manner as used in the simulation study.

2.5.3 Clustering results

The content-based clustering results obtained by CCRS are shown in Figure 2.20. From the boxplots, we see how the clusters differ with respect to the assessment of the items. In some case, these differences are limited to only one item (e.g., clusters 1 and 5) but mostly difference concern at least two items corresponding to the same value (e.g., clusters 5 and 7).

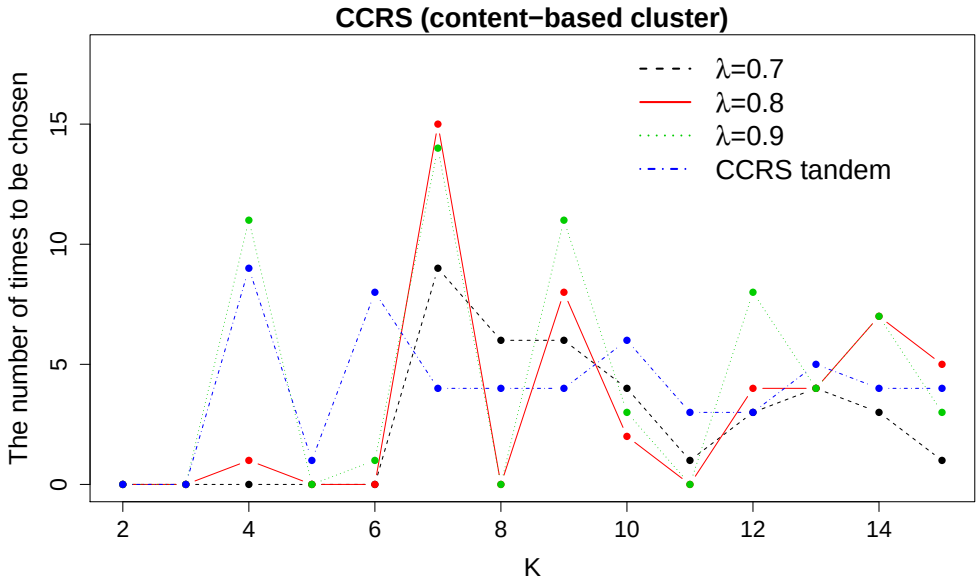

Figure 2.18: The number of times that different values of the parameter K was selected for content-based clustering methods using the KL index for different values of λ.

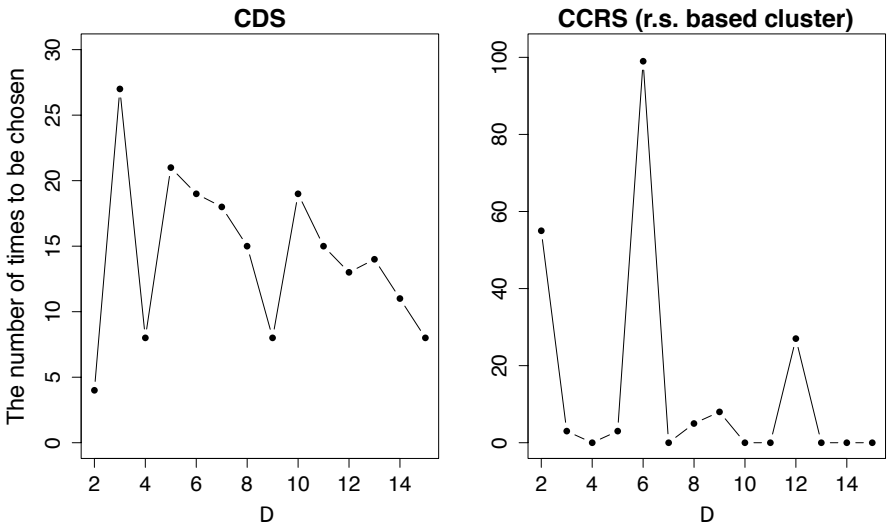

Figure 2.19: The number of times that different values of the parameter D was selected, using the KL index, in the response-style-based clustering.

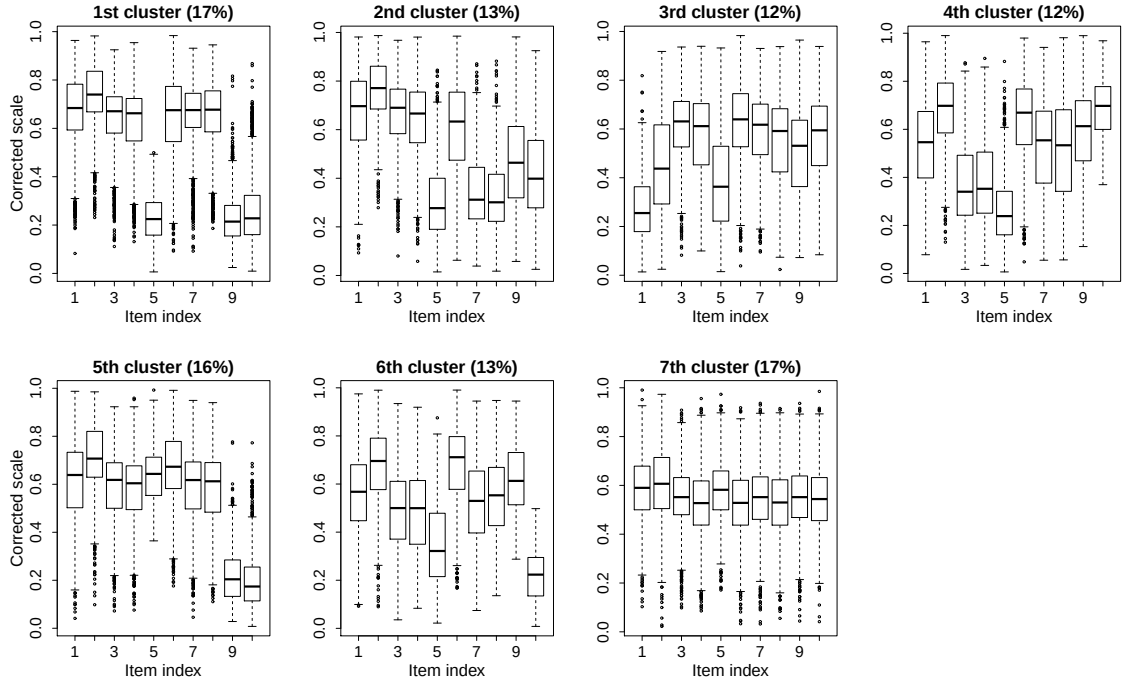

Figure 2.20: Boxplot of the 10 items (horizontal axis) for the content-based (cont.) clusters obtained with CCRS. The vertical axis indicates the scale of the corrected data using estimated response functions.

In most cases, items corresponding to the same values are similarly evaluated within a cluster. For the third value, i.e., "in-group orientation" this does not appear to be the case. Apparently, the evaluation of in-group orientation differs depending on the group considered. That is, relatives or people from the same town. However, the difference may also be due to the phrasing of the two items. In particular, in question 5 respondents assess whether they would favor relatives whereas question 6 merely ask respondents whether they appreciate success of others (from the same town in this case).

In CCRS, respondent-specific response functions are used. Clustering the resulting functions leads to a 6 cluster solution. The corresponding response functions are depicted in Figure 2.21 (left). Moreover, recall that the coefficients in the response functions estimated by CCRS can be used to visually capture the characteristics of response styles (c.f. Section 2.3.5). The results are shown in Figure 2.21 (right). The second and fourth response-style-based clusters correspond to a low β_1 and a high β_3 value. This indicates an "acquiescence" response style. Similarly, the third response-style-based cluster demonstrates low values for both β_1 and β_3 corresponding to a "midpoint" response style.

To see whether the response-style-based clusters are related to nationalities, we consider the distributions over the countries in Table 2.2. In the second and fourth clusters (acquiescence), most respondents are Chinese. On the other hand, the third response-style-based cluster (midpoint) comprises over 50% Japanese.

To see how the response style clusters and content-based clusters correspond, we consider

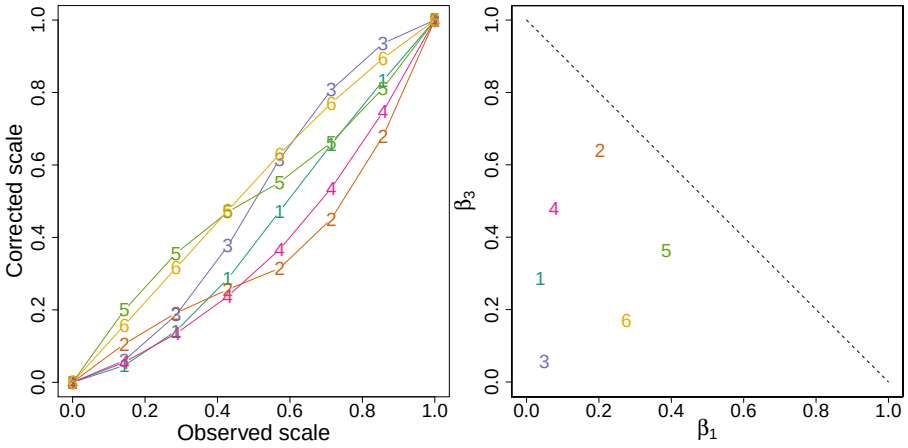

Figure 2.21: (Left) Estimated response functions of the response-style-based clusters obtained by CCRS. (Right) Low dimensional plot for $\boldsymbol{\beta}_d$ of the response-style-based clusters. Numbers indicate response-style-based clusters, and correspond to those used in the left plot.

a mosaic plot that visualizes the cross-tabulation of the two cluster solutions. Figure 2.22 shows that there does not appear to be significant overlap of respondents between the content and response-style-based clusters. In each content-based cluster respondents from all response-style-based clusters are present. That is, the content-based clusters and the response-style-based clusters do not coincide.

Table 2.2: (%'s) Distribution of respondents nationalities over the response-style-based clusters ($d = 1, \ldots, 6$) obtained by CCRS. The absolute frequencies of each response-based clusters are 1509, 692, 1781, 1822, 966 and 864, for the $d = 1, \ldots, 6$ th response-based clusters, respectively.

d	China	Japan	Korea	Taiwan
1	43.8	14.0	20.5	21.6
2	52.3	1.7	20.4	25.6
3	17.7	53.2	23.8	5.3
4	59.7	2.3	8.4	29.6
5	27.7	6.8	18.0	47.4
6	16.6	24.7	28.6	30.2

To assess whether results by CCRS are "better" than result obtained by the CDS tandem and k-means methods is cumbersome as we do not know whether there are true underlying cluster structures. Nevertheless, a comparison of results may be insightful and help to interpret the results.

The CDS tandem content-based clustering results and cluster-wise response functions are shown in Figure 2.23 and Figure 2.24, respectively. Looking at the association be-

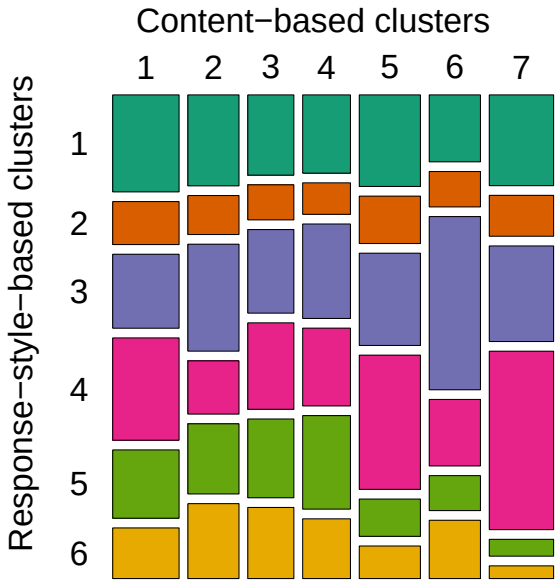

Figure 2.22: Mosaic plot of the CCRS content and response-style-based clusters. The index of response-style-based clusters corresponds to the number used in Figure 2.21.

tween the content-based and response-style-based clusters as shown in Figure 2.25, we see that there is significant overlap between respondents in the content and response-style-based clusters. This indicates that the cluster-wise correction of response styles results in content-based clusters that are similar to those response-style-based clusters. Consequently, when interpreting content-based clusters one may merely be considering response-style-based differences.

The k-means clustering results are shown in Figure 2.26. Here, the clusters also appear to correspond to some response tendencies. In the first and third content-based clusters, for example, we see that respondents predominantly select high and midpoint ratings, respectively. An interpretation relative to the item content appears difficult for this solution.

In summary, it appears that the k-means results may only reflect response tendencies. Moreover, when using CDS to correct for response style effect, the corrected data strongly reflects certain response-style-based clustering results. Consequently, the content-based cluster results obtained from the corrected data do not yield additional content-related insights. On the other hand, with the proposed CCRS method, we obtain content-based clusters that are dissimilar to the response-style-based clusters.

Finally, results of the empirical data application cannot be validated easily, and the fact that we find "dissimilar" clusters, does not provide evidence that CCRS should be preferred over CDS. This is because that we cannot observe "true preference". However, the results of this application, in combination with the results of the simulation study, do suggest that CCRS is able to better retrieve content-based cluster structures.

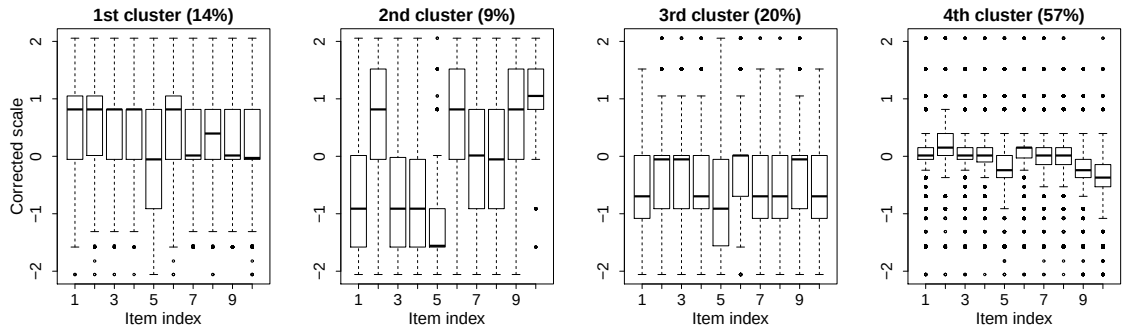

Figure 2.23: Boxplot of 10 items (horizontal axis) of the content-based clusters obtained by CDS tandem. The vertical axis indicates the scale of the corrected data using the estimated response functions.

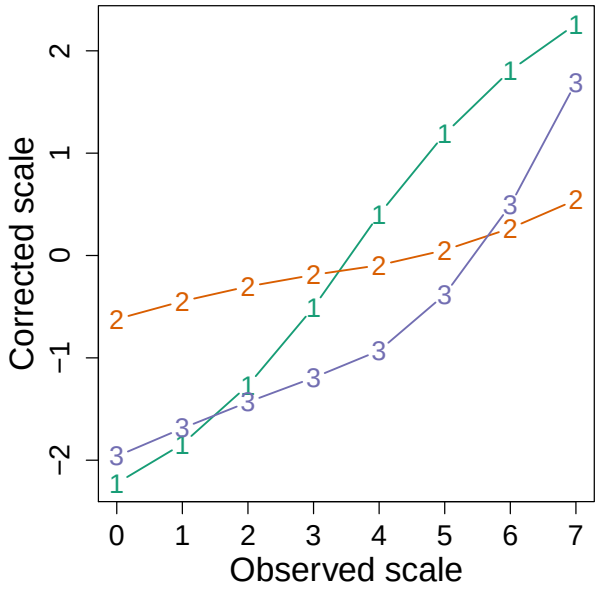

Figure 2.24: Estimated response functions of response-style-based cluster obtained by CDS.

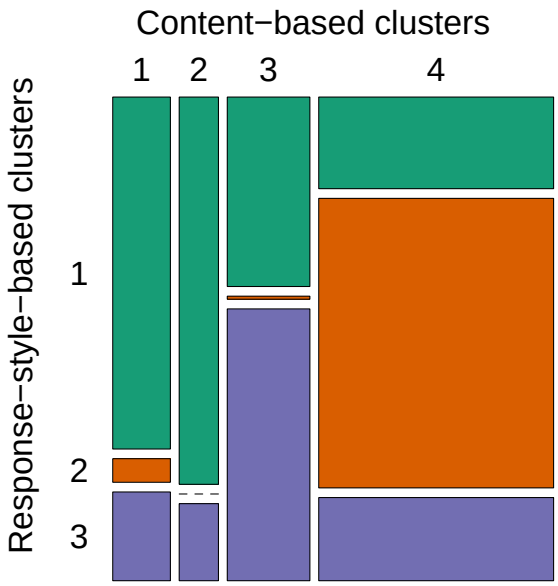

Figure 2.25: Mosaic plot of the content and response-style-based clusters by CDS tandem. The index of response-style-based clusters corresponds to the number in Figure 2.24.

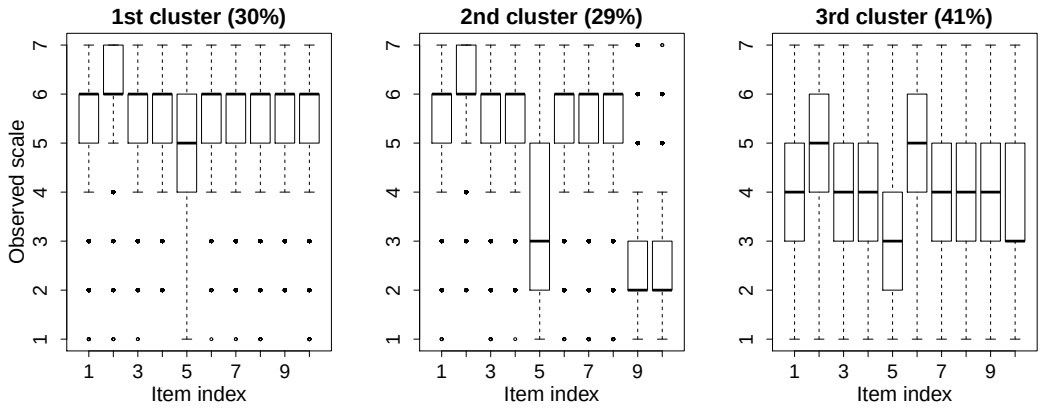

Figure 2.26: Boxplot of 10 items (horizontal axis) of the content-based clusters obtained by k-means. The vertical axis indicates the scale of the original preference data.

Chapter 3

Visualizing class specific heterogeneous tendencies in categorical data

3.1 Problem of interpretation of MCA biplot

Correspondence analysis (CA) and multiple correspondence analysis (MCA) are popular methods that support visual interpretations of the associations among categorical variables (e.g., Greenacre, 1984). In MCA, obtained quantifications of categories and individuals can be depicted in a biplot, which indicates not only the associations among categories and among individuals but also those between individuals and categories (e.g., Greenacre, 1993; J. C. Gower & Hand, 1996).

In an MCA biplot, if many individuals choose the same two categories, the quantifications for these categories and corresponding individuals tend to locate in close proximity. Therefore, an MCA biplot enables us to visually identify individuals with similar category choice tendencies. Moving beyond this benefit, adding pertinent external information about individuals can enhance interpretations of MCA biplots. By external information, we refer to information that might not be of use for the estimation of the coordinates, but that may be useful for interpreting the resulting biplot.

Several studies describe ways to incorporate external information about individuals into an MCA biplot (e.g., Yanai, 1986, 1988; Böckenholt & Böckenholt, 1990; Takane, Yanai, & Mayekawa, 1991; Van Buuren & de Leeuw, 1992; Böckenholt & Takane, 1994; Yanai & Maeda, 2002; Hwang, Yang, & Takane, 2005). Hwang and Takane (2002) also show that various objectives for incorporating the external information can be generalized into a linear constraint framework. Here, we focus specifically on external information that consists of a set of categorical variables, and we refer to subsets of these data that correspond to the categories of the external information as classes.

To visually explicate how individuals' tendencies differ depending on these classes, we could integrate external variables before applying MCA, but this approach transforms the information, such that the information is no longer external and instead becomes

part of the original analysis. As an alternative approach, we might seek to establish individual quantifications (i.e., points) visually, according to the classes. For example, if gender is an external variable, points corresponding to men can be colored black, and those corresponding to women are red. This approach incorporates external information corresponding to only one categorical variable at the time. Another option would be to obtain average quantifications for each class. By plotting these average points, as well as the category points of the original (non-external) variables, we can depict the relationship between the external information and the categories. We refer to this notion as the averaging approach.

Yet the averaging approach only reveals the average tendencies of many individuals within a class, obscuring their heterogeneous tendencies. When a relatively small group in a class has a strong tendency toward a particular category that the majority group in the class does not select, this preference would not be visible in a biplot that relies on an averaging approach. Despite representing a minority, such tendencies could be interesting to consider, especially to characterize tendencies by class.

Therefore, in this chapter, we propose a new approach to find class-specific clusters and depict them together with the categories of the (original) variables. The result is a visual depiction of the categories (i.e., category quantifications), together with points that represent clusters for the different classes of data. With this visualization, we can identify different heterogeneous tendencies within a class in a single MCA biplot, as well as perceive the relationships among classes that correspond to the categories of external variables.

The remainder of this paper is organized as follows. In Section 3.2, we introduce our proposed method and its relationship with existing approaches, including the linear row constraint framework. In addition, we compare a biplot obtained using the averaging approach and one obtained using our proposed method. The simulation study in Section 3.3 appraises the proposed method in various external information scenarios; the application of our method to empirical data in Section 3.4 confirms its appeal.

3.2 Multiple set cluster CA (MSCCA)

In this Section, we introduce our approach, which we call multiple-set cluster CA (MSCCA), as an extension of several existing methods, such as cluster CA (van de Velden, D'Enza, & Palumbo, 2017), CA, and the linear row constraint framework.

3.2.1 The MSCCA objective function

Suppose that we have n observations of m categorical variables, and in conjunction, that, for the same n observations, we have H additional categorical variables that contain external information. We refer to these H additional variables as supplementary variables. In this Section, notations to formulate the MSCCA objective function are defined as follows.

Let q_j $(j = 1, \ldots, m)$ be the number of categories for the jth variable, and let $Q =$

$\sum_{j=1}^{m} q_j$. We create dummy matrices $\boldsymbol{Z}_j$ for the m categorical variables using the categorical data, so the rows of $\boldsymbol{Z}_j$ are $(q_j \times 1)$ vectors $\boldsymbol{z}_{ji} = (z_{ji\ell})$ $(i = 1, \ldots, n\,;\, \ell = 1, \ldots, q_j)$, where $z_{ji\ell} = 1$ if individual i chooses the ℓth category in the jth variable, and the other elements are 0. Similarly, we create dummy matrices for the H supplementary variables, with $\boldsymbol{V}_h = (v_{his})$ $(h = 1, \ldots, H;\, s = 1, \ldots, r_h)$, where r_h is the number of categories for the hth supplementary variable.

In addition, let K_{hs} be the number of clusters for the sth category (class) of the hth supplementary variable $(h = 1, \ldots, H\,;\, s = 1, \ldots, r_h)$, with $K_h = \sum_{s=1}^{r_h} K_{hs}$. Let $\boldsymbol{B}_j$ be the $q_j \times p$ quantification matrix for the categories of the jth variable, where p denotes the number of dimensions, and let $\boldsymbol{U}_h$ and $\boldsymbol{G}_h$ be $n \times K_h$ cluster indicator matrices and $K_h \times p$ quantification matrices for cluster centers in the hth supplementary variable, respectively. The objective function of MSCCA then can be defined as

$$\min_{\boldsymbol{U}_h, \boldsymbol{G}_h, \boldsymbol{B}_j} \psi(\boldsymbol{U}_h, \boldsymbol{G}_h, \boldsymbol{B}_j \mid \boldsymbol{Z}_j, \boldsymbol{V}_h) = \frac{1}{nHm} \sum_{j=1}^{m} \sum_{h=1}^{H} \|\boldsymbol{U}_h \boldsymbol{G}_h - \boldsymbol{Z}_j \boldsymbol{B}_j\|^2 \tag{3.2.1}$$

$$\text{s.t.} \quad \frac{1}{nm} \sum_{j=1}^{m} \boldsymbol{B}_j' \boldsymbol{Z}_j' \boldsymbol{Z}_j \boldsymbol{B}_j = \boldsymbol{I}_p, \quad \boldsymbol{J}_n \boldsymbol{U}_h \boldsymbol{G}_h = \boldsymbol{U}_h \boldsymbol{G}_h$$

$$\text{where} \quad \underset{(n \times K_h)}{\boldsymbol{U}_h} = \begin{pmatrix} \boldsymbol{u}_{h11}' & \cdots & \boldsymbol{u}_{h1r_h}' \\ \vdots & \ddots & \vdots \\ \boldsymbol{u}_{hn1}' & \cdots & \boldsymbol{u}_{hnr_h}' \end{pmatrix}$$

$$\boldsymbol{u}_{his} = (u_{his1}, \ldots, u_{hisK_{hs}})'$$

$$\text{s.t.} \quad \begin{cases} u_{hisk} \in \{0, 1\},\ (k = 1, \ldots, K_{hs}),\ \sum_{k=1}^{K_{hs}} u_{hisk} = 1 & (v_{his} = 1) \\ u_{hisk} = 0,\ (k = 1, \ldots, K_{hs}) & (v_{his} = 0) \end{cases} \tag{3.2.2}$$

$$(i = 1, \ldots, n\,;\, s = 1, \ldots, r_h\,;\, h = 1, \ldots, H).$$

Here, $\boldsymbol{J}_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ is the centering matrix, $\boldsymbol{I}_n$ is an $n \times n$ identity matrix, and $\mathbf{1}_n$ is an $n \times 1$ vector of ones. When we estimate parameters, the number of dimension p, the number of clusters for each class, K_{hs} $(h = 1, \ldots, H\,;\, s = 1, \ldots, r_h)$, must be pre-specified.

The constraint on $\boldsymbol{U}_h$ in Equation (3.2.2) defines a two-level hierarchical cluster structure. Specifically, for each supplementary variable h, individuals first are divided into r_h *known* classes, corresponding to the categories of the variable as indicated by $\boldsymbol{u}_{hi1}, \ldots, \boldsymbol{u}_{hir_h}$. Then within each class s $(s = 1, \ldots, r_h)$, individuals are assigned to K_{hs} unknown clusters based on variables, as indicated by $u_{his1}, \ldots, u_{hisK_{hs}}$.

We can illustrate the construction of $\boldsymbol{U}_h$ with a small example. Suppose that we have five observations and that one supplementary variable, (e.g., h), corresponds to gender. In addition, assume we want to find two clusters for the males and one cluster for females, so that $K_{h1} = 2$ and $K_{h2} = 1$. Let observations $i = 1, 3, 5$ be males where $i = 1, 3$ are in the first male cluster and $i = 5$ is in the second one, individuals $i = 2, 4$ are females.

When we consider the cluster indicator vector for individual $i = 1$, $\boldsymbol{u}_{h1}$, because we partition the data by gender, the vector is split as $\boldsymbol{u}_{h1}' = (\boldsymbol{u}_{h11}', \boldsymbol{u}_{h12}')$, where $\boldsymbol{u}_{h11}$ and $\boldsymbol{u}_{h12}$ denote the cluster indicator vectors of individual i in the male and female classes,

respectively. Performing this partitioning for all individuals, we obtain

$$\boldsymbol{U}_h = \begin{pmatrix} \boldsymbol{u}'_{h11} & \boldsymbol{u}'_{h12} \\ \boldsymbol{u}'_{h21} & \boldsymbol{u}'_{h22} \\ \boldsymbol{u}'_{h31} & \boldsymbol{u}'_{h32} \\ \boldsymbol{u}'_{h41} & \boldsymbol{u}'_{h42} \\ \boldsymbol{u}'_{h51} & \boldsymbol{u}'_{h52} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Then, to relate our method to other methods, we can rewrite Equation (3.2.1) as

$$\min_{\boldsymbol{U},\boldsymbol{G},\boldsymbol{B}} \psi(\boldsymbol{U},\boldsymbol{G},\boldsymbol{B} \mid \boldsymbol{Z},\boldsymbol{V}) = \frac{1}{nHm} \sum_{j=1}^{m} \|\boldsymbol{U}\boldsymbol{G} - \boldsymbol{Z}_j^H \boldsymbol{B}_j\|^2 \tag{3.2.3}$$

$$\text{s.t.} \quad \frac{1}{nHm} \sum_{j=1}^{m} \boldsymbol{B}'_j \boldsymbol{Z}_j^{H\prime} \boldsymbol{Z}_j^H \boldsymbol{B}_j = \boldsymbol{I}_p, \quad \boldsymbol{J}_{nH}\boldsymbol{U}\boldsymbol{G} = \boldsymbol{U}\boldsymbol{G}$$

$$\text{where,} \underset{(nH\times q_j)}{\boldsymbol{Z}_j^H} = \begin{pmatrix} \boldsymbol{Z}_j \\ \vdots \\ \boldsymbol{Z}_j \end{pmatrix}, \underset{(K\times p)}{\boldsymbol{G}} = \begin{pmatrix} \boldsymbol{G}_1 \\ \vdots \\ \boldsymbol{G}_H \end{pmatrix}, \tag{3.2.4}$$

$$\boldsymbol{U} = \text{b-diag}(\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_H) = \begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{U}_H \end{pmatrix},$$

and $K = \sum_{h=1}^{H} K_h = \sum_{h=1}^{H} \sum_{s=1}^{r_h} K_{hs}$. If we set $H = 1$ and define $\boldsymbol{U}_H$ as a cluster indicator matrix for K_H clusters without the hierarchical clustering structure—that is, $\boldsymbol{U}_H = (u_{ik})$, $(i = 1, \ldots, n\,;\, k = 1, \ldots, K_H)$ where $\sum_{k=1}^{K_H} u_{ik} = 1$ and $u_{ik} \in \{0, 1\}$—then Equation (3.2.3) is equivalent to cluster CA (van de Velden et al., 2017), which is equivalent to GROUPALS (Van Buuren & Heiser, 1989) when applied to categorical variables.

Thus, MSCCA represents an extension of cluster CA that is able to specify the cluster allocation for each class simultaneously in a common low-dimensional space, in which the quantifications for categories $\boldsymbol{B}_j$ $(j = 1, \ldots, m)$ are optimally estimated for all clusters. The advantage of MSCCA is, by plotting the cluster results of each class in a common low-dimensional space, we can not only identify different heterogeneous tendencies within a class in a single MCA biplot, but also perceive the relationships among classes.

3.2.2 Algorithm of MSCCA

Algorithm to estimate MSCCA parameters

To estimate the parameters $\boldsymbol{U}, \boldsymbol{G}$, and $\boldsymbol{B}_j$ $(j = 1, \ldots, m)$, we use an alternating least squares algorithm. The updating formulas come from a direct extension of cluster CA (van de Velden et al., 2017).

Step 1: Initialization. Determine $K_{hs}\,(h = 1, \ldots, H\,; s = 1, \ldots, r_h)$ and p. Set the number of iterations to $t = 0$, and set a convergence criterion ε. Then, randomly generate initial clusters for each class.

Step 2: Update $\boldsymbol{B}_j$. Let $\boldsymbol{B} = (\boldsymbol{B}_1', \ldots, \boldsymbol{B}_m')'$ and $\boldsymbol{Z}^H = (\boldsymbol{Z}_1^H, \ldots, \boldsymbol{Z}_m^H)$. Then find $\boldsymbol{B}^{(w+1)}$ as

$$\boldsymbol{B}^{(w+1)} = \sqrt{nm}\boldsymbol{D}^{-1/2}\boldsymbol{B}^*$$

$$\text{where} \quad \frac{1}{m}\boldsymbol{D}^{-1/2}\boldsymbol{Z}^{H\prime}\boldsymbol{J}_{nH}\boldsymbol{U}^{(w)}(\boldsymbol{U}^{(w)\prime}\boldsymbol{U}^{(w)})^{-1}\boldsymbol{U}^{(w)\prime}\boldsymbol{J}_{nH}\boldsymbol{Z}\boldsymbol{D}^{-1/2} = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{*\prime}$$

$$\boldsymbol{D} = \widetilde{\boldsymbol{Z}}'\widetilde{\boldsymbol{Z}}, \quad \widetilde{\boldsymbol{Z}} = \text{b-diag}(\boldsymbol{Z}_1^H, \ldots, \boldsymbol{Z}_m^H)$$

Step 3: Update $\boldsymbol{G}$. Obtain $\boldsymbol{G}^{(w+1)}$ as follows:

$$\boldsymbol{G}^{(w+1)} = \frac{1}{m}(\boldsymbol{U}^{(w)\prime}\boldsymbol{U}^{(w)})^{-1}\boldsymbol{U}^{(w)\prime}\boldsymbol{J}_{nH}\boldsymbol{Z}\boldsymbol{B}^{(w+1)}$$

Step 4: Update $\boldsymbol{U}$. To obtain $\boldsymbol{U}_h^{(w+1)}$, the update proceeds by row. Specifically, each element in the ith row of $\boldsymbol{U}_h$, or $\boldsymbol{u}_{his} = (u_{hisk})$ $(k = 1, \ldots, K_{hs})$, gets updated as follows: If $v_{his} = 1$,

$$u_{hisk}^{(w+1)} = \begin{cases} 1 & (k = \underset{\ell \in \{1,\ldots,K_{hs}\}}{\arg\min} \|\boldsymbol{f}_i - \boldsymbol{g}_{hs\ell}^{(w+1)}\|^2) \\ 0 & (\text{others}) \end{cases}$$

and otherwise, $u_{hisk}^{(w+1)} = 0$. Here, $\boldsymbol{f}_i$ is the ith row of $\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}^{(w+1)}$, and $\boldsymbol{g}_{hsk}^{(w+1)}$ is the cluster center of the kth cluster in the sth category in the hth supplementary variable.

Step 5: Convergence test Compute $\psi^{(w)}$, the value of the objective function from Equation (3.2.1), using updated parameters. For $t > 1$, if $\psi^{(w)} - \psi^{(w-1)} < \varepsilon$, terminate; otherwise, let $t \leftarrow t + 1$ and return to Step 2.

Problem of local minimum in MSCCA

Similar to CCRS algorithm, MSCCA applies k-means type algorithm, and thus has a local minimum problem. Figure 3.1 shows the value of optimized MSCCA objective function over the number of algorithm runs, similar to Figure 2.5. This figure suggests that with more runs, the more stable results. However, after some T (in this example, T should be around 25), the algorithm gets relatively stable. Therefore, in practice, it is recommended to make a plot like Figure 3.1, and specify T around which the algorithm gets relatively stable.

3.2.3 Biplot by MSCCA

Here, we show how MSCCA can be used to construct a biplot. In van de Velden et al. (2017), cluster CA is formulated as a maximization problem.

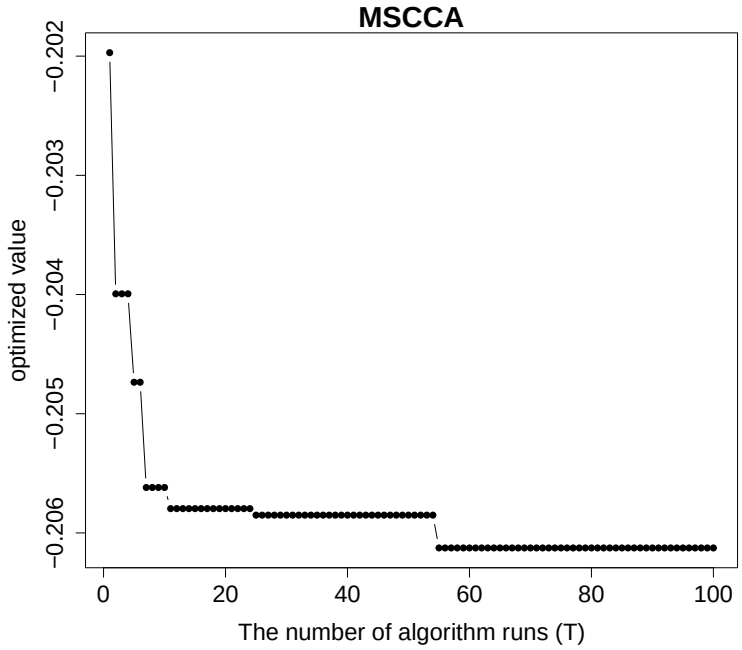

Figure 3.1: The graph of optimized value of MSCCA objective function over the number of algorithm runs T with different initial values, similar to Figure 2.5. The artificial data used in this example is $n = 300$, $m = 10$, $H = 3$, $r_h = 2$, $(h = 1, \ldots, H)$, $K_{hs} = 2$, $(h = 1, \ldots, H\,;\, s = 1. \ldots, r_h)$, and $q = 5$. How to generate the artificial data is explained in Section 3.3.1.

Proposition 3.2.1. *the MSCCA in Equation (3.2.3) can be rewritten as the following maximization problem:*

$$\max_{\boldsymbol{U},\boldsymbol{B}} \psi(\boldsymbol{U},\boldsymbol{B} \mid \boldsymbol{Z}^H) = \operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}^{H\prime}\boldsymbol{J}_{nH}\boldsymbol{U}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_{nH}\boldsymbol{Z}^H\boldsymbol{B} \tag{3.2.5}$$

$$\text{s.t.} \quad \frac{1}{nHm}\sum_{j=1}^{m} \boldsymbol{B}_j'\boldsymbol{Z}_j^{H\prime}\boldsymbol{Z}_j^H\boldsymbol{B}_j = \boldsymbol{I}_p$$

Proof. To simplify the notation, in this proof we only consider $H = 1$ case without loss of generality, and denote $\boldsymbol{Z}^H$ and $\boldsymbol{Z}_j^H$ by $\boldsymbol{Z}$ and $\boldsymbol{Z}_j$. At first the equivalence is shown when $\boldsymbol{U}$ is fixed. Considering the constraints, Equation (3.2.3) can be rewritten as

$$\begin{aligned}
\psi(\boldsymbol{U},\boldsymbol{G},\boldsymbol{B}) &= \frac{1}{nm}\sum_{j=1}^{m} \|\boldsymbol{U}\boldsymbol{G} - \boldsymbol{Z}_j\boldsymbol{B}_j\|^2 \\
&= \frac{1}{nm}\left(m \operatorname{tr} \boldsymbol{G}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{G} - 2\operatorname{tr}\sum_{j=1}^{m} \boldsymbol{B}_j'\boldsymbol{Z}_j'\boldsymbol{U}\boldsymbol{G} + \operatorname{tr}\sum_{j=1}^{m} \boldsymbol{B}_j'\boldsymbol{Z}_j'\boldsymbol{Z}_j\boldsymbol{B}_j \right) \\
&= \frac{1}{n}\operatorname{tr} \boldsymbol{G}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{G} - \frac{2}{nm}\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{U}\boldsymbol{G} + p,
\end{aligned}$$

because $\frac{1}{nm}\sum_{j=1}^{m} \boldsymbol{B}_j'\boldsymbol{Z}_j'\boldsymbol{Z}_j\boldsymbol{B}_j = \boldsymbol{I}_p$. Using $\boldsymbol{J}_n\boldsymbol{U}\boldsymbol{G} = \boldsymbol{U}\boldsymbol{G}$ and omitting the constant, this minimization will be

$$\frac{1}{n}\operatorname{tr} \boldsymbol{G}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{G} - \frac{2}{nm}\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}\boldsymbol{G} \tag{3.2.6}$$

Solving this for $\boldsymbol{G}$, we obtain

$$\boldsymbol{G} = \frac{1}{m}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}$$

Inserting this in Equation (3.2.6), it will be

$$\begin{aligned}
&\frac{1}{nm^2}\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B} - \frac{2}{nm^2}\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B} \\
&= -\frac{1}{nm^2}\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}
\end{aligned}$$

Minimizing this is equivalent to maximizing Equation (3.2.5). Next, the equivalence is shown when $\boldsymbol{B}$ is fixed and $\boldsymbol{U}$ is not. At first, a k-means type optimization problem

$$\min_{\boldsymbol{U},\boldsymbol{G}} \|\boldsymbol{U}\boldsymbol{G} - \boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}\|^2$$

is equivalent to the optimization problem in Equation (3.2.5), since this can be rewritten as

$$\begin{aligned}
\|\boldsymbol{U}\boldsymbol{G} - \boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}\|^2 &= \operatorname{tr} \boldsymbol{G}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{G} - 2\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}\boldsymbol{G} + \operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B} \\
&= -\operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B} + \operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}
\end{aligned} \tag{3.2.7}$$

Here, we use $\boldsymbol{G} = m^{-1}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}$. Omitting a constant term, minimizing Equation (3.2.7) is equivalent to maximizing Equation (3.2.5). On the other hand, with $\boldsymbol{B}_j$ ($j =$

$1, \ldots, m$) fixed, Equation (3.2.3) can be written as

$$\begin{aligned}\sum_{j=1}^{m} \|\boldsymbol{UG} - \boldsymbol{Z}_j\boldsymbol{B}_j\|^2 &= \|\boldsymbol{UG} - \boldsymbol{ZB}\|^2 \\ &= \operatorname{tr}\boldsymbol{G}'\boldsymbol{U}'\boldsymbol{UG} - 2\operatorname{tr}\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{UG} + \operatorname{tr}\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{ZB} \\ &= -\operatorname{tr}\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}'(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{ZB} + \operatorname{tr}\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{ZB}\end{aligned}$$

This is the same as Equation (3.2.7). Thus, we obtain the proposition. □

When we leave $\boldsymbol{U}$ fixed, maximizing Equation (3.2.5) is equivalent to minimizing

$$\min_{\boldsymbol{G},\boldsymbol{B}} \psi^{CA}(\boldsymbol{G}, \boldsymbol{B} \mid \boldsymbol{Z}^H, \boldsymbol{V}, \boldsymbol{U}) = \|\widetilde{\boldsymbol{P}} - \boldsymbol{D}_r^{1/2}\boldsymbol{G}\boldsymbol{B}'\boldsymbol{D}_c^{1/2}\|^2 \tag{3.2.8}$$

$$\text{s.t.} \quad \frac{1}{nm}\boldsymbol{B}'\boldsymbol{D}_c\boldsymbol{B} = \boldsymbol{I}_p$$

$$\text{where} \quad \widetilde{\boldsymbol{P}} = \boldsymbol{D}_r^{-1/2}(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1/2} \tag{3.2.9}$$

$$\boldsymbol{P} = (nm)^{-1}\boldsymbol{U}'\boldsymbol{Z}^H, \ \boldsymbol{r} = \boldsymbol{P}\boldsymbol{1}_Q, \ \boldsymbol{c} = \boldsymbol{P}'\boldsymbol{1}_K, \quad \boldsymbol{D}_r = \operatorname{diag}(\boldsymbol{r}), \quad \boldsymbol{D}_c = \operatorname{diag}(\boldsymbol{c})$$

The proof of the equivalence is available from van de Velden et al. (2017). Here, $\boldsymbol{P}$ indicates a $K \times Q$ scaled contingency table of clusters for each class (row) and category (column), and each element in $\boldsymbol{r}\boldsymbol{c}'$, $r_k c_\ell$ ($k = 1, \ldots, K$; $\ell = 1, \ldots, Q$), indicates the scaled expected frequency with an assumption of independence between the kth cluster and the ℓth category. Thus, the matrix $\widetilde{\boldsymbol{P}}$ represents the standardized deviations from the the assumption of independence between cluster membership and the categorical variables.

From Equation (3.2.8), it follows that the inner product of $\boldsymbol{D}_r^{1/2}\boldsymbol{G}$ and $\boldsymbol{D}_c^{1/2}\boldsymbol{B}$ approximates the matrix of standardized deviations from independence, $\widetilde{\boldsymbol{P}}$. That is, in MSCCA, we can use $\boldsymbol{G}$ and $\boldsymbol{B}$ to construct a biplot in which a greater the inner product of the kth row vector of $\boldsymbol{G}$ and the ℓth row vector in $\boldsymbol{B}$ generally indicates a stronger association between the kth cluster and the ℓth category.

Note that in the resulting biplot, the points of the row and column are not necessarily similarly spread (e.g., J. Gower, Groenen, & van de Velden, 2010). In this case, the points can be scaled using a constant, such that the average squared deviation from the origin of the row and column points is the same. See van de Velden et al. (2017) for detail.

3.2.4 Relationship with linear row constraint approach

Hwang and Takane (2002) show that several approaches for incorporating external information about individuals into an MCA biplot can be generalized, as a linear row constraint framework.

To add linear row constraints in MCA, we formulate the following objective function

$$\min_{\boldsymbol{\Gamma},\boldsymbol{B}} \psi^{const}(\boldsymbol{\Gamma}, \boldsymbol{B} \mid \boldsymbol{Z}_j, \boldsymbol{V}_h) = \frac{1}{nm}\sum_{j=1}^{m} \|\boldsymbol{C\Gamma} - \boldsymbol{Z}_j\boldsymbol{B}_j\|^2 \tag{3.2.10}$$

$$\text{s.t.} \quad \frac{1}{nm}\sum_{j=1}^{m} \boldsymbol{B}_j'\boldsymbol{Z}_j'\boldsymbol{Z}_j\boldsymbol{B}_j = \boldsymbol{I}_p, \quad \boldsymbol{J}_n\boldsymbol{C\Gamma} = \boldsymbol{C\Gamma},$$

where $\boldsymbol{C}$ is the $n \times n$ matrix that contains linear row constraints for the quantifications. If $\boldsymbol{C} = \boldsymbol{I}$, the problem reduces to the homogeneity formulation of MCA.

The choice of $\boldsymbol{C}$ depends on the objective that underlies the incorporation of the external information. For example, if we were to use $\boldsymbol{C} = \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'$, where $\boldsymbol{V} = \text{b-diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_H)$, and we inserted $\boldsymbol{Z}_j^H$ for $\boldsymbol{Z}_j$, then Equation (3.2.10) would produce the averaging approach we described previously, because the class (category) would be represented by the average quantification of individuals corresponding to that class. Alternatively, if we aimed to "remove" the effect of external information from a biplot, then we might use $\boldsymbol{C} = \boldsymbol{I} - \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'$ (e.g., Takane & Shibayama, 1991; Takane & Hwang, 2002; Hwang & Takane, 2002), which is equivalent to deducting the class conditional means from the data. For example, if as supplementary variable we have gender, the mean of all males is deducted from all male observations.

Although MSCCA follows a different approach from these two examples to incorporate external information, we can reformulate this method to fit into the linear row constraint framework. In particular, for a fixed $\boldsymbol{U}$, the MSCCA objective function in Equation (3.2.3) can be rewritten as a minimization problem:

Proposition 3.2.2. *Minimizing Equation* (3.2.3) *with respect to* $\boldsymbol{B}$ *is equivalent to minimizing*

$$\min_{\boldsymbol{\Gamma},\boldsymbol{B}} \psi^{MSCCA}(\boldsymbol{\Gamma}, \boldsymbol{B} \mid \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{V}) = \frac{1}{nHm} \sum_{j=1}^{m} \|\boldsymbol{C}\boldsymbol{\Gamma} - \boldsymbol{Z}_j^H \boldsymbol{B}_j\|^2 \tag{3.2.11}$$

$$\text{s.t.} \quad \frac{1}{nm} \sum_{j=1}^{m} \boldsymbol{B}_j' \boldsymbol{Z}_j^{H\prime} \boldsymbol{Z}_j^H \boldsymbol{B}_j = \boldsymbol{I}_p, \quad \boldsymbol{J}_{nH}\boldsymbol{C}\boldsymbol{\Gamma} = \boldsymbol{C}\boldsymbol{\Gamma}$$

$$\text{where} \quad \boldsymbol{C} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'$$

where, $\boldsymbol{U}$ *still features the hierarchical cluster structure constraint imposed by Equation* (3.2.2)*.*

Proof. Similar to Proposition 3.2.1, in this proof we only consider $H = 1$ case without loss of generality in order to simplify the notation.

Using constraints, Equation (3.2.11) can be rewritten as

$$\psi^{MSCCA}(\boldsymbol{\Gamma}, \boldsymbol{B}) = \frac{1}{n} \operatorname{tr} \boldsymbol{\Gamma}'\boldsymbol{C}\boldsymbol{\Gamma} - \frac{2}{nm} \operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{C}\boldsymbol{\Gamma} \tag{3.2.12}$$

Solving this for $\boldsymbol{\Gamma}$, we obtain

$$\boldsymbol{\Gamma} = \frac{1}{m}\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}$$

Inserting this into Equation (3.2.12), we obtain a minimization problem of

$$-\frac{1}{nm^2} \operatorname{tr} \boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{J}_n\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{J}_n\boldsymbol{Z}\boldsymbol{B}. \tag{3.2.13}$$

On the other hand, using the proof in Proposition 3.2.1, the optimization problem in Equation (3.2.3) can also be rewritten as (3.2.13). Thus we obtain the proposition. □

Table 3.1: Categories of variables of artificial data for the simple illustration

Variable type	Variable name	Category
Variables to estimate quantifications	Meal	Western, Asian
	Drink	Fruits juice, Tea, Alcohol
Supplementary variables	Nationality	American, Japense
	Gender	Male, Female

From this formulation, it immediately follows that MSCCA represents a special case of Equation (3.2.10), with $\boldsymbol{C} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'$.

Thus, even though MSCCA has a different objective to incorporate external information from the existing approaches, this can be still considered as an extension of the linear row constraint framework.

3.2.5 Numerical illustration of an MSCCA biplot

Here, we present a small example, using artificial data, to illustrate how MSCCA works. With this example, we zoom in specifically on the differences between MSCCA and the averaging approach for the visualization of heterogeneous tendencies.

To start, we artificially create categorical data of 200 individuals represented in Figure 3.2 that have two categorical variables (meal and drink preference), and two supplementary variables (nationality and gender). Table 3.1 contains the variables and corresponding categories. With this analysis, we seek to determine if different tendencies, with respect to the meal and drink preferences, emerge for groups of individuals, depending on their nationality and gender.

We created the data to establish three true clusters in the full data set. All individuals in the first cluster choose "Western meal" for the meal variable, and "fruit juice" for the drink variable (W&J), all in the second cluster choose "Asian meal" and "Tea" (A&T), and in the third cluster, all individuals choose "Western meal" and "alcohol" (W&A). The frequency distribution of the generated artificial data over each cluster in each class is shown in Figure 3.2, revealing there are two clusters for Americans, Japanese and females and three clusters for males.

The biplot that results from an averaging approach, in Figure 3.3 (left), clearly reveals overall tendencies of many Americans and Japanese consumers, strongly associated with W&J and A&T, respectively. However, the much smaller number of individuals who choose "alcohol" makes it impossible to specify who (i.e., which nationality or gender) makes this choice.

In contrast, by obtaining clusters for each class, the MSCCA biplot makes the tendencies of this relatively small number of individuals visible. When we use the correct number of clusters for each class, the MSCCA biplot result in Figure 3.3 (right) clearly reveals that a small number of male Americans choose "alcohol". In addition, this biplot still depicts the tendencies of the larger groups, as obtained in the averaging approach. That is, MSCCA reveals the tendencies of small groups, without losing the information about

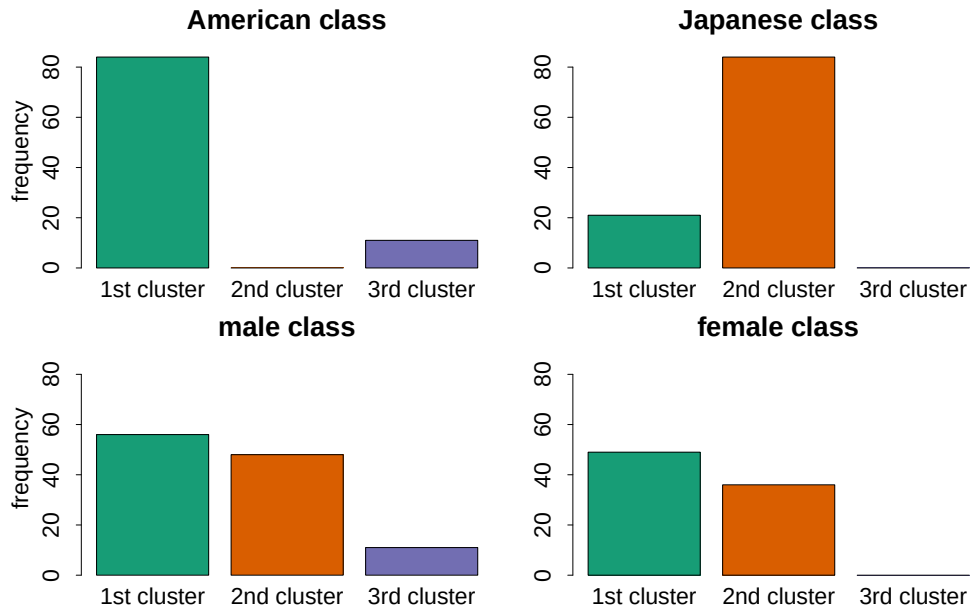

Figure 3.2: Frequency distributions for each combination of supplementary categories for each true cluster

the tendencies of larger groups.

To understand why MSCCA can depict heterogeneous tendencies more clearly than the averaging approach, we compare the methods that the two approaches use to calculate associations between classes and categories. That is, both MSCCA and averaging reflect a CA framework. Averaging is equivalent to CA for the $\sum_{h=1}^{H} r_h \times Q$ contingency table (row is class, column is category); MSCCA is equivalent to CA for the $\sum_{h=1}^{H}\sum_{s=1}^{r_h} K_{hs} \times Q$ contingency table (row is clusters in each class, column is category), for a given cluster allocation. Figure 3.4 shows heat maps of the relative deviations, $\widetilde{\boldsymbol{P}}^{ave}$ and $\widetilde{\boldsymbol{P}}^{MSCCA}$, for each method calculated based on their respective contingency tables. Thus, using this framework, we can say that the difference between the two methods is whether the rows of the contingency table are split by clusters in each class.

This factor then distinguishes between averaging and MSCCA in the calculation of the expected frequency, $\boldsymbol{rc}'$. Specifically, in the averaging approach, the expected frequency in the (3,1) element in $\widetilde{\boldsymbol{P}}^{ave}$ is calculated using the number of individuals who are American and choose "alcohol", whereas that for the (2,5) element in $\widetilde{\boldsymbol{P}}^{MSCCA}$ results from calculating the number of individuals who are in the second cluster in the American class and choose "alcohol". That is, in MSCCA, the number of individuals used to calculate expected frequency is less for each row in the contingency table than the number for the averaging approach.

Note that the relative deviation indicates the size of the observed frequency (i.e., the number of individuals choosing a particular category), compared with the expected frequency (i.e., the expected number of individuals choosing the category under an assumption of independence). Therefore, the relative deviation tends to increase when the ex-

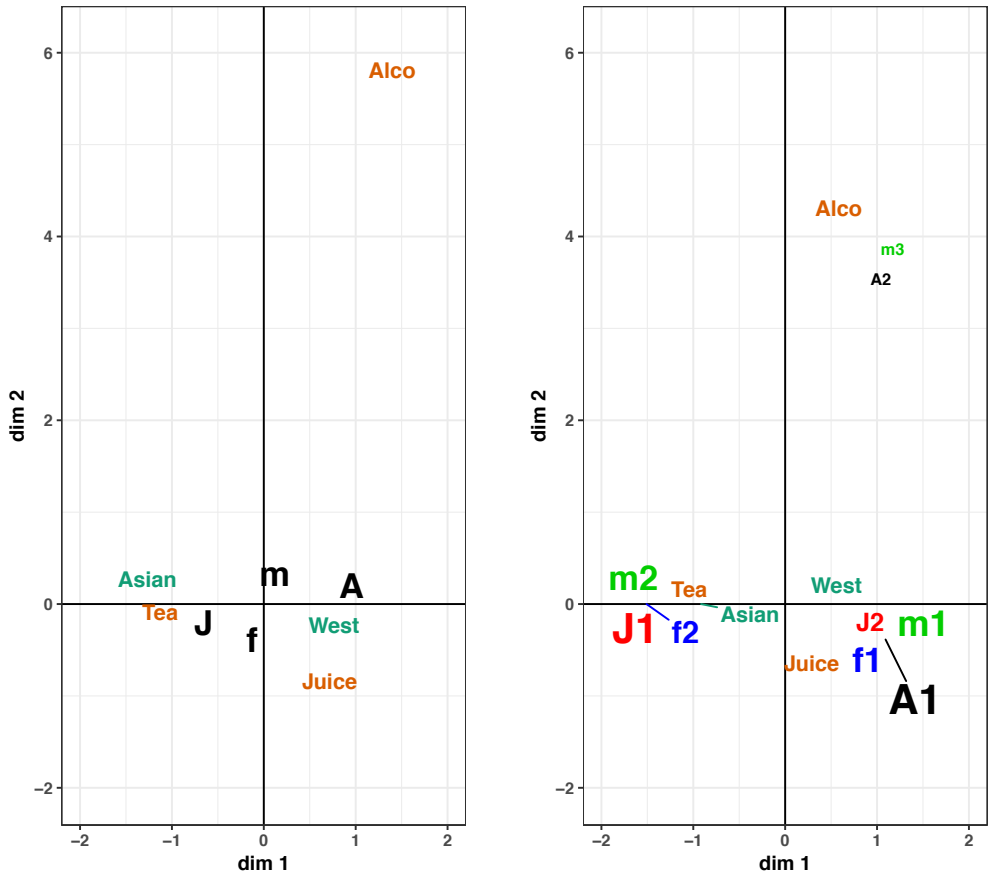

Figure 3.3: Results (Left) Averaging approach. Average points are labelled "A" (American), "J" (Japanese), "m" (male), and "f" (female), and the label sizes correspond to class sizes. Other character labels indicate category points. (Right) MSCCA. The points labeled "A," "J," "m," or "f" followed by a number, correspond to the cluster points for each class.

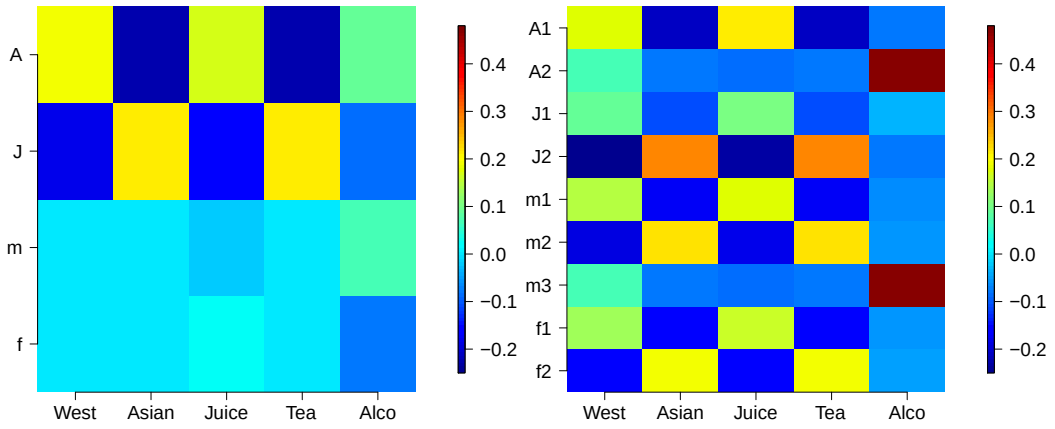

Figure 3.4: Heatmaps of $\widetilde{\boldsymbol{P}}^{ave}$ (left) and $\widetilde{\boldsymbol{P}}^{MSCCA}$ (right), respectively. These matrices are calculated as in Equation (3.2.9), using $\boldsymbol{P}^{ave} = (nHm)^{-1}\boldsymbol{V}'\boldsymbol{Z}^H$ and $\boldsymbol{P}^{MSCCA} = (nHm)^{-1}\boldsymbol{U}'\boldsymbol{Z}^H$, respectively.

pected frequency is calculated using the limited number of individuals who select the same categories.

Thus in MSCCA, clustering individuals for each class reveals the heterogeneous tendencies within each class clearly, regardless of the size of the groups that exhibit similar tendencies.

3.3 Simulation study of MSCCA

We conducted a simulation study to evaluate the performance of MSCCA in different scenarios. By using simulations, we can determine the effects of the supplementary variables on the accuracy of the clustering and biplots achieved through MSCCA.

3.3.1 Data generation

The data generation process consists of two steps: generating an $n \times m$ data matrix, and generating $n \times H$ matrix of supplementary variables. First, we start by dividing the m variables into two groups: active variables that relate to the clustering structure, and noise variables that are unrelated to the cluster structure. Furthermore, we determine the cluster allocation with a multinomial distribution. To generate data for the active variables, we assign one category for each variable a high probability of 0.8. Then the (low) probabilities assigned to the remaining categories are determined according to $\bar{\boldsymbol{p}} = (\bar{p}_\ell)$ $(\ell = 1, \ldots, q-1)$, where $\bar{\boldsymbol{p}} = ((1-0.8) \times (p_1, \ldots, p_{q-1}) / \sum_{\ell=1}^{q-1} p_\ell)$ and $p_\ell \sim U(0, (1-0.8))$. The high probability categories are cluster specific. Then to generate noise variables, we use a multinomial distribution in which the proportion for each category is $1/q$. In our simulation study, we set the ratio of active to noise variables to $1:1$.

Second, to generate the data matrix corresponding to the H supplementary variables, we consider two scenarios: balanced and unbalanced distributions over the categories. In

the balanced scenarios, the multinomial probabilities for all categories are equal. In the unbalanced scenario, the probabilities are $1/S, \ldots, r_h/S$, where r_h denotes the number of categories for the supplementary variable, and $S = \sum_{s=1}^{r_h} s$.

3.3.2 Simulation study design

To assess the performance of the methods in different settings, we fix the number of observations $n = 300$ and the number of variables $m = 10$. Then, we consider a full factorial design with the number of categories for each variable $q = 5, 7$; the number of clusters $K = 2, 3$; the number of supplementary variables $H = 1, 3$; and the number of categories for the supplementary variables $r_h = 3, 5$. Finally, for the supplementary variables we note the balanced and unbalanced scenarios. For each combination of parameters in the simulation, we randomly generate 100 different $n \times m$ data matrices and $n \times H$ supplementary variable matrices. For each data set, we apply MSCCA using 100 random initial values.

3.3.3 Evaluation

We evaluate the performance of the proposed methods by checking the accuracy of both the clustering and the biplots. To measure clustering accuracy, we turn to the ARI, same as in Section 2.4. We calculate the ARI for the class-specific clustering results separately.

For biplot accuracy, we use a goodness-of-fit (GF) index (e.g., Gabriel, 2002), which is equivalent to the so-called congruence coefficient (e.g., Lorenzo-Seva & Ten Berge, 2006). The GF between configurations $\boldsymbol{Y}$ and $\boldsymbol{H}$ is defined as

$$\mathrm{GF}(\boldsymbol{Y}, \boldsymbol{H}) = \frac{\mathrm{tr}^2(\boldsymbol{Y}'\boldsymbol{H})}{\mathrm{tr}(\boldsymbol{Y}'\boldsymbol{Y})\,\mathrm{tr}(\boldsymbol{H}'\boldsymbol{H})} = \cos^2(\boldsymbol{Y}, \boldsymbol{H}).$$

Therefore, we calculate the GF between $\boldsymbol{Y}$ and $\boldsymbol{H}$, where $\boldsymbol{H} = \boldsymbol{G}\boldsymbol{B}'$ (with $\boldsymbol{G}$ and $\boldsymbol{B}$ as the MSCCA solutions) and $\boldsymbol{Y} = \tilde{\boldsymbol{P}}^{true} = \boldsymbol{D}_r(\boldsymbol{P}^{true} - \boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c$, such that $\boldsymbol{P}^{true} = \boldsymbol{U}'\boldsymbol{Z}$ and $\boldsymbol{U}$ is the true cluster allocation. Note that by definition, GF$\in [0, 1]$. In our calculation of the GF index, we assume that the true cluster allocation is known. Therefore, cluster accuracy does not affect the GF index.

3.3.4 Result

The results for the GF index in Figure 3.5 indicate that it tends to decrease as the number of categories q increases. The number of supplementary variables H does not substantially affect the GF. Rather, the GF tends to be somewhat better when there are fewer categories r_h in the supplementary variables and when the distribution over the categories is balanced.

The cluster retrieval results in Figure 3.6 show that overall, ARI decreases when the number of clusters K increases and when the number of categories q decreases. In contrast, the number of supplementary variables H and whether the distributions over the categories are unbalanced do not affect the median ARI substantially. However, for more supplementary variables with balanced distributions, we note more outlying results. In

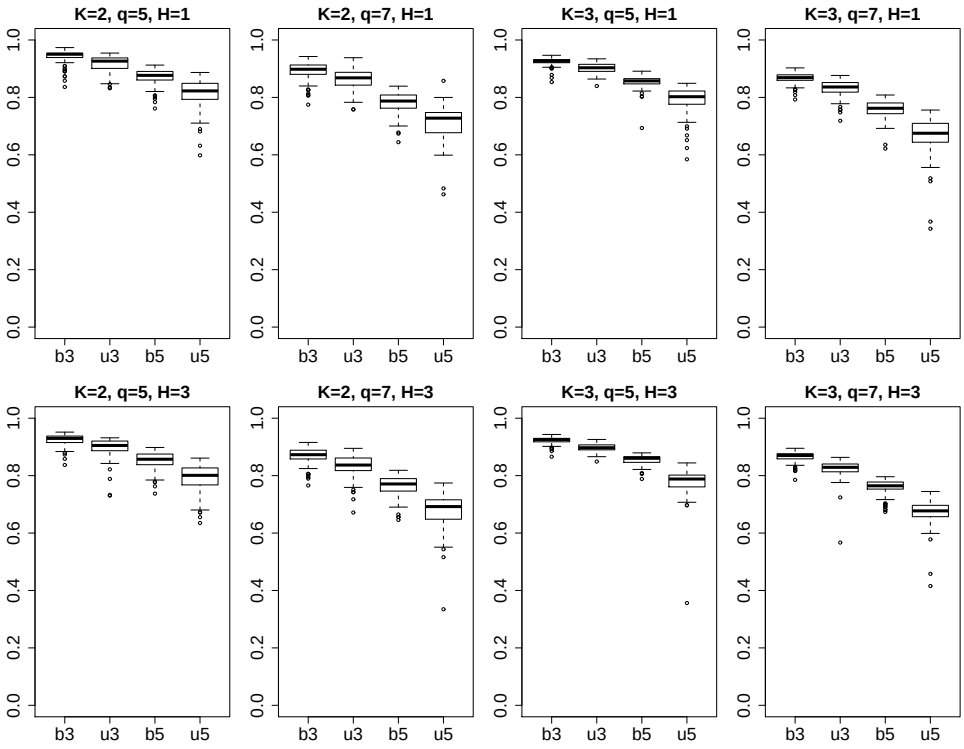

Figure 3.5: Boxplot of GF for each case. On the horizontal axis, b3 indicates balanced categories in supplementary variables, and $r_h = 3$ for all $h = 1, \ldots, H$; while b5 has balanced categories with $r_h = 5$ for all h. Similarly, u3 and u5 indicate unbalanced categories with $r_h = 3, 5$, respectively.

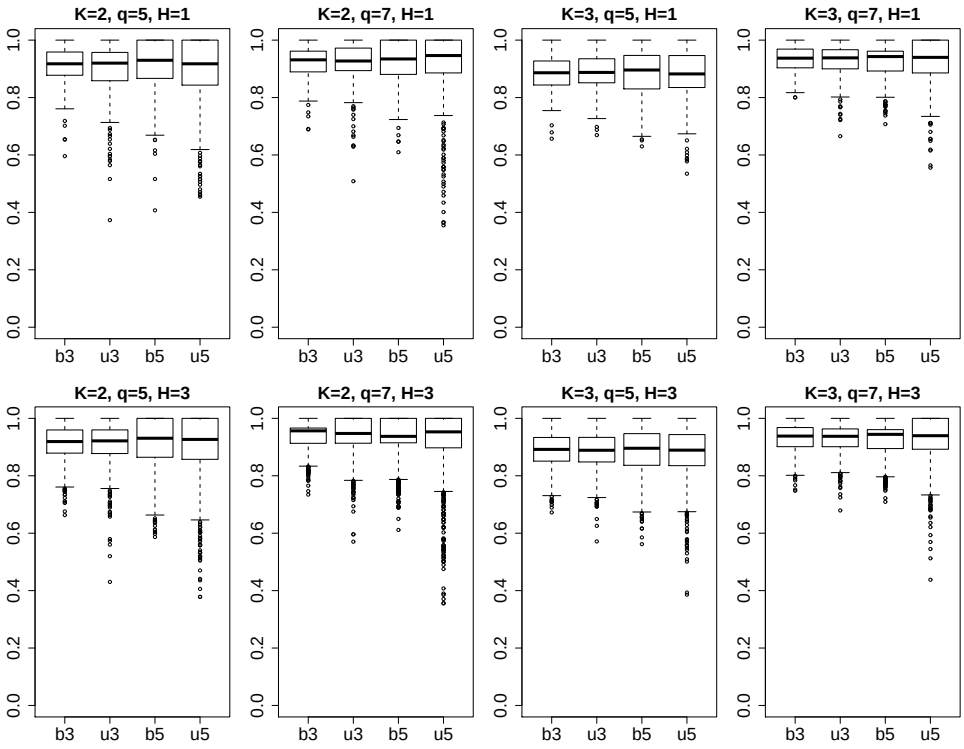

Figure 3.6: Boxplot of ARI for each case.

addition, the number of categories for the supplementary variables appears to affect variance in the ARI results, such that the ARI for $r_h = 5$ has greater variance than that for $r_h = 3$.

3.3.5 Conclusions from the simulation study

The simulation study shows that the number of supplementary variables does not affect the accuracy of the biplot or the clustering. We can increase the number of supplementary variables H without harming the accuracy of the results. However, increasing the number of supplementary variables H leads to more points in the biplot, resulting in a more complicated visualization. We thus assert that H can be increased as long as the biplot remains interpretable.

In addition, though the clustering results are hardly affected by the nature of the supplementary variables (i.e., number of categories r_h, and whether the distribution over the categories is balanced), the simulation study indicates that biplot accuracy is affected. In particular, using supplementary variables with more categories and unbalanced distributions over categories leads to a decrease in biplot accuracy. In conclusion, when there are several candidates for supplementary variables, it is better to select balanced supplementary variables with fewer categories.

Table 3.2: Categories for each variable and their corresponding labels in biplots and descriptions.

Variable type	Variable name	Label	Description
Non-supplementary variables	Light conditions	Dark0	Daylight
		Dark1	Darkness: street lights present and lit
		Dark2	Darkness: street lights present but unlit
		Dark3	Darkness: no street lighting
	Weather conditions	Fine	Fine without high winds
		Rain	Raining without high winds
		Snow	Snowing without high winds
		Fine_w	Fine with high winds
		Rain_w	Raining with high winds
		Snow_w	Snowing with high winds
		Fog	Fog or mist — if hazard
		Other	Other
	Road surface conditions	Dry	Dry
		Wet	Wet / Damp
		Snow	Snow
		Frost	Frost / Ice
		Flood	Flood (surface water over 3cm deep)
	Speed limit	∼30	Speed limit is up to 30km/h
		∼70	Speed limit is up to 70km/h
Supplementary variables	Casualty class	Driver	Casualty is one driver
		Ped	Casualty is one pedestrian
	Area	Urban	Occurring in urban area
		Rural	Occurring in rural area

3.4 Empirical example of MSCCA

In this Section, we illustrate the proposed method using data that reflect road accidents in the United Kingdom. With these data, we seek to determine how the circumstances in which a car accident occurs depends on the type of accident. We compare the results using MSCCA, the averaging approach, and cluster CA, to establish how each method would visualize the relationships.

3.4.1 Data and setting

The data were obtained from the U.K. Department for Transport's road safety statistics (https://data.gov.uk/dataset/road-accidents-safety-data). In these data, observations are accidents, and the (categorical) variables refer to information about those accidents. For this illustration, we selected accidents that occurred in January 2016, that involved one casualty (either a driver or a pedestrian), and in which at most two parties were involved. The resulting data set contains $n = 3,026$ observations.

Regarding the circumstances of the accident, we consider four (i.e., $m = 4$) variables: lighting conditions, weather conditions, road surface conditions, and speed limit. For the types of accident, we select two ($H = 2$) supplementary variables: casualty class and area.

Table 3.2 summarizes the variables and their categories.

As is true of any cluster analysis method, determining the number of clusters is not trivial. In MSCCA, the number of clusters must be prespecified for each class K_{hs} ($h = 1, \ldots, H; s = 1, \ldots, r_h$). For this study, we use the KL index to determine the number of clusters for each class, with separate cluster CA analyses. Specifically, we apply cluster CA to class-specific data (i.e., data corresponding to one category of the supplementary variables) to determine the number of clusters K_{hs} that corresponds to the optimal KL index. This procedure results in four clusters for the driver class, five clusters for the pedestrian class, four in the urban class and four clusters for the rural class (i.e., $K_{11} = 4$, $K_{12} = 5$, $K_{21} = 4$ and $K_{22} = 4$). Henceforward, we refer to a cluster from the driver class as driver cluster, clusters from the pedestrian class as pedestrian clusters, and so on.

In a comparative analysis, we also consider the averaging approach and cluster CA with complete data (i.e., including the supplementary variables in the analysis to determine clusters and quantifications). To select the number of clusters for the complete cluster CA analysis, we employed the KL index and obtained $K = 7$ clusters.

3.4.2 Result

MSCCA result

In the biplot for the MSCCA solution (Figure 3.7), we see that the largest pedestrian clusters, as well as the largest urban and rural clusters (P1, U1, and Ru1, respectively) are related to categories such as "Fine," "Fine_w," and "Dry." That is, many accidents in urban and rural areas result in pedestrian casualties and have a strong association with what is generally be considered good driving conditions (e.g., fine weather, dry roads).

The driver cluster (D1) instead is related to categories such as "Dark3," "Snow_w (weather condition)," and "Snow (road surface)." Therefore, accidents that result in driver casualties tend to have a strong association with bad driving conditions, such as a dark night or slippery road. Another driver cluster, close to the good conditions, is the smallest one, indicating that accidents resulting in a driver casualty are less likely under good driving conditions.

In the rural class, we also recognize that though the largest rural cluster is proximal to categories that correspond to good conditions, the second largest rural cluster is close to bad conditions. Therefore, accidents in rural areas occur in both good and bad driving conditions.

The fourth-largest cluster for rural data and the third-largest clusters for the three other classes indicate similar associations with categories such as "Rain," "Rain w," and (to some extent) "Wet." This indicates that for all classes of supplementary variables, some clusters of accidents occur in rainy weather.

By inspecting the MSCCA biplot and relating the class-specific cluster points to the category quantifications, we can visually perceive how accidents, split into different classes, relate differently to weather and road conditions. For example, for pedestrians, the risk of casualties exists even in favorable conditions, but accidents involving drivers are more strongly related to bad conditions.

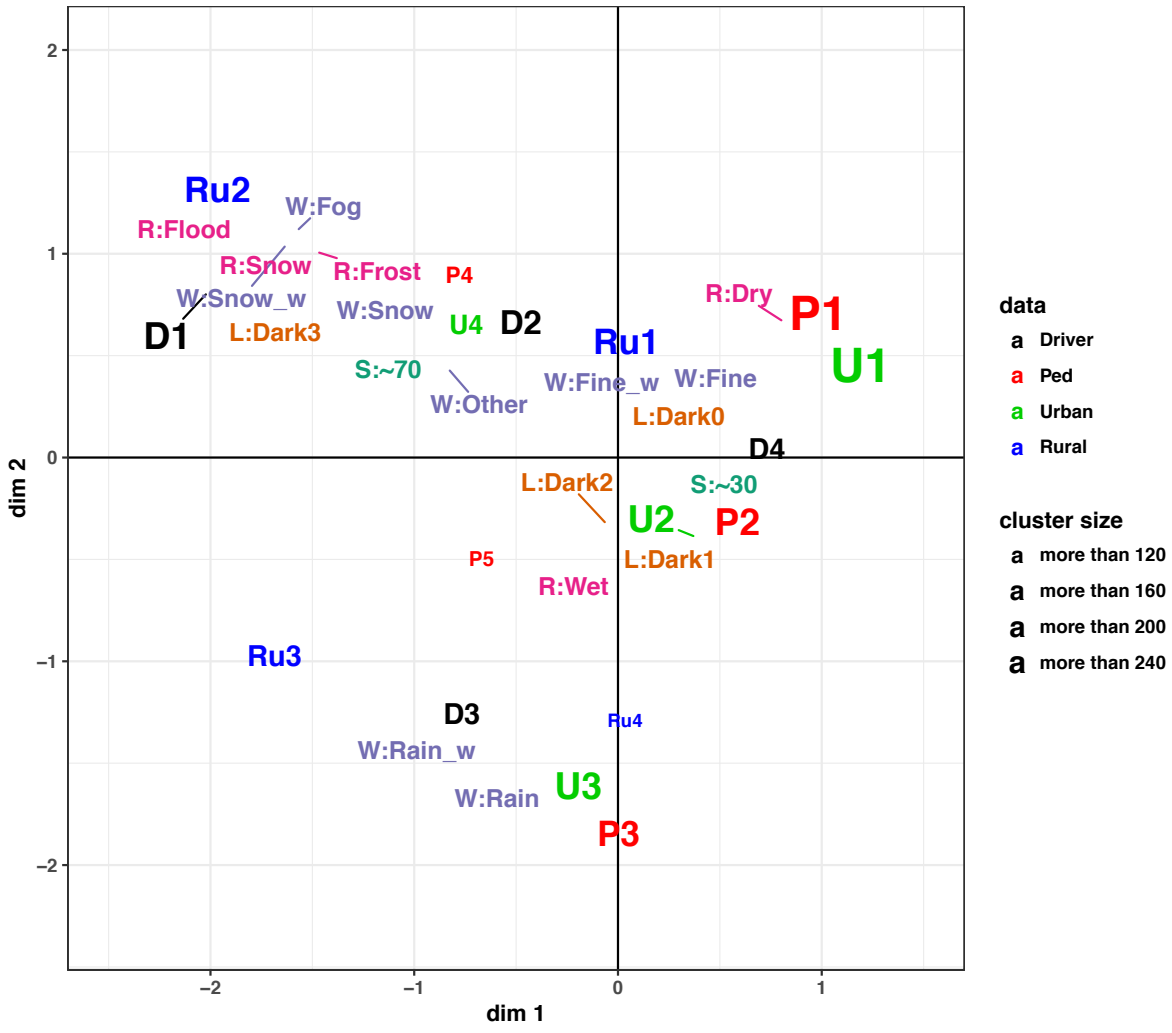

Figure 3.7: Results using MSCCA. The numbered labels indicate cluster points with "D" indicating driver clusters, "P" corresponding to the pedestrian class, "U" to the urban class, and "Ru" to rural class clusters. The numbers reflect the size of each cluster within its class (e.g., "D1" indicates the largest size cluster in the driver class), as also indicated by the label sizes. Character labels also indicate light conditions "L", weather conditions "W", road surface conditions "R", and speed limits "S".

Note that since there are more sunny or cloudy days than rainy or other whether which is unfavorable to drivers, we could not simply state that the good weather is more associated with the car accidents. However, this result at least indicates the possible association compared with other categories used in the data analysis.

Averaging approach result

The results using the averaging approach are in Figure 3.8. We can still interpret the information regarding classes with respect to categories, but the averaging of the results limits the available information. Specifically, we see that "Driver" and "Rural" relate to categories indicating bad driving conditions (e.g., "∼70", "Show"), while "Pedestrian" and "Urban" are related to categories corresponding to good driving conditions (e.g., "∼30", "Fine", "Dark0"). However, it is difficult to interpret the relationship between classes and categories that are not close to the class quantifications. Averaging limits us to interpreting tendencies that many accidents in each class have in common. Differentiation with respect to smaller, relatively homogeneous subgroups is no longer possible.

Cluster CA result

Figure 3.9 shows the results using the cluster CA approach. In contrast with the averaging approach, we can now distinguish different clusters corresponding to several accident tendencies. For example, we find a cluster associated with rain-related categories, whereas this relationship was not clear in the averaging results. Yet the cluster CA approach still limits interpretations with respect to classes. For example, we can see that "Pedestrian" and "Urban" are related to good driving conditions, but we cannot interpret the relationship between the "Pedestrian" and "Urban" class in conditions such as rainy or bad driving conditions (e.g, "∼70" and "Dark3"). In contrast, with MSCCA, we can better interpret these relationships (e.g., we can see that the "Pedestrian" class has a weaker association with bad driving conditions than with good ones or with rainy conditions, because the smallest pedestrian cluster is closest to bad driving conditions.)

3.4.3 Conclusions of empirical data analysis

In this Section, we have compared three visualization results to appraise differences in how the biplots incorporate external information. All three methods can identify situations in which many accidents occur in each class. However, only by using MSCCA were we able to differentiate across conditions in which many or few accidents occurred. Specifically, this method reveals that relatively many accidents in the "Pedestrian" and "Urban" classes occur when conditions are good, but fewer occur when conditions are bad. Conversely, for the "Driver" class, accidents predominantly occur under bad conditions, with only a few appearing when conditions are good. For accidents corresponding to the "Rural" class, we find that they occur in both good and bad conditions. Finally, for all classes, we uncover relatively small clusters of accidents that relate strongly to rainy conditions.

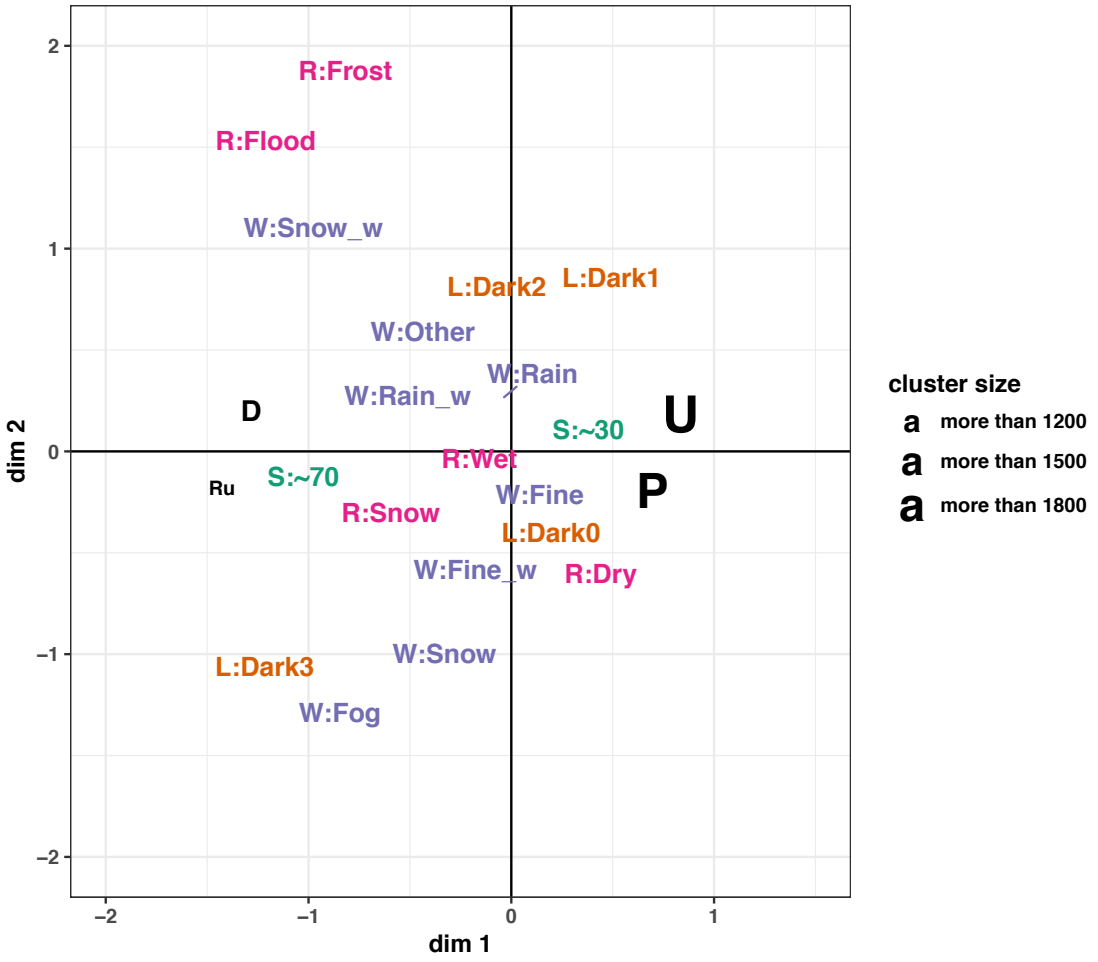

Figure 3.8: Results using the averaging approach. The character labels "D", "P", "U" and "Ru" indicate classes defined by supplementary variables, label sizes correspond to class sizes. Other character labels indicate category points, same as Figure 3.7.

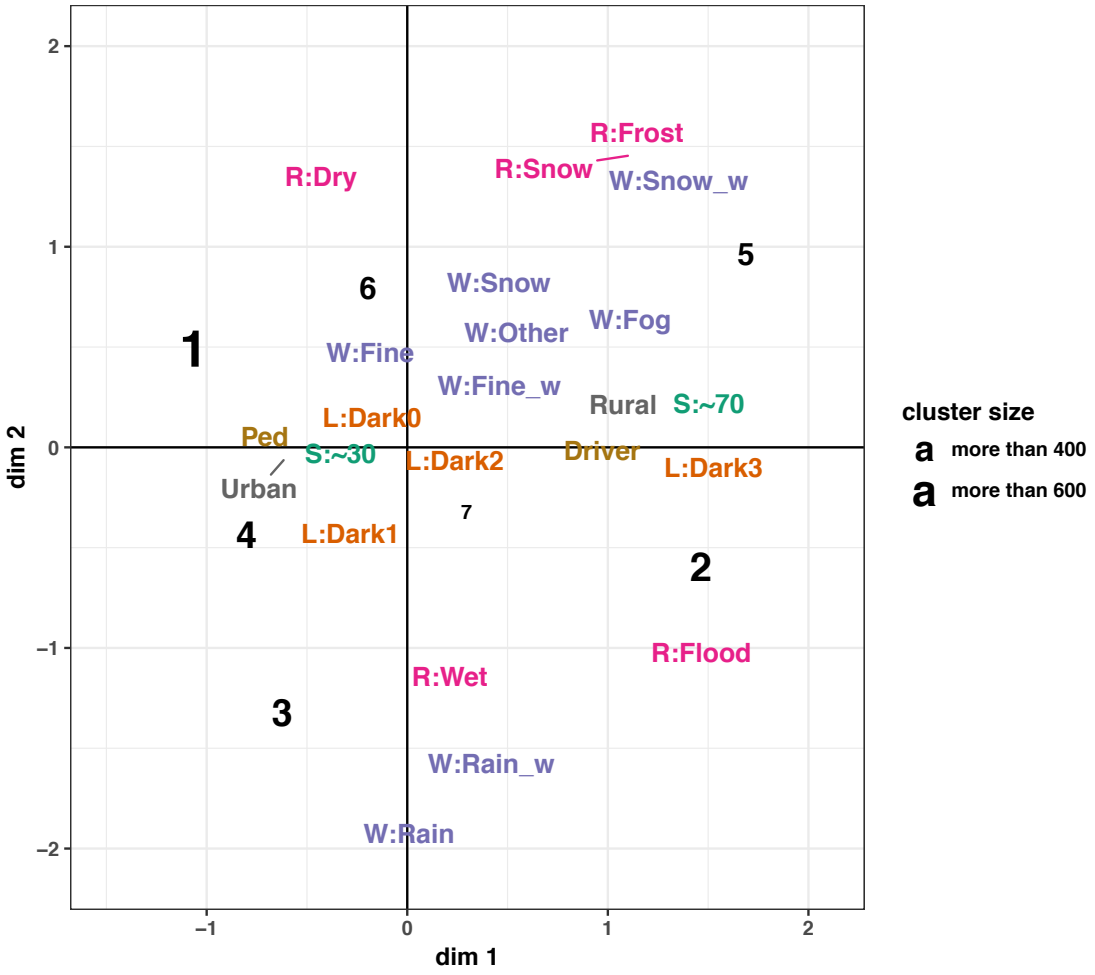

Figure 3.9: Results using cluster CA. Numbered labels indicate cluster points and the numbers are ordered according to the sizes of the clusters. The other labels are the same as in Figures 3.7 and 3.8.

Chapter 4

Conclusion

In this paper, two problems in interpretation of large-scale categorical data are considered. When data is large scale, it is useful to apply methods to capture the latent structure of data for simple interpretation. As such a method, a cluster analysis and a visualization method are often used. However, in some situation, it is still difficult to interpret with these methods. To overcome issues, new methods are proposed, in Chapter 2 and 3, respectively.

In Chapter 2, the problem of response style bias when clustering ordinal categorical data was discussed. In this paper, as the response styles we focus on acquiescence, disacquiescence, midpoint, or extreme response styles, similar to Schoonees et al. (2015), and assume that that there are no respondents having response-style-like preference. There have been many methods to detect response style bias in both probabilistic and non-probabilistic models, but not so many to correct for response styles. CDS, non-probabilistic model, proposed by Schoonees et al. (2015), can correct for response styles, and they investigated the accuracy of correction using the simulation study. In this paper, to conduct correction and clustering on the corrected data more efficiently than CDS, which requires a dimensional reduction, we proposed a new method, called CCRS, for simultaneously correcting for response style bias and performing content-based clustering.

By generalizing the concept of a response function as introduced by van de Velden (2008) and Schoonees et al. (2015), respondent-specific response functions were estimated without first applying a dimension reduction technique. In CCRS, we obtain clusters which are not affected by response style bias. Note that our new correction method explained in Section 2.3.1, which is a part of CCRS, can also be used to correct for response style bias in combination with other methods and applications.

In a simulation study, we demonstrated that our proposed CCRS method outperforms existing methods such as a CDS tandem (CDS and k-means) as well as k-means in most cases. In particular, when both content and response-style-based clustering structures exist, CCRS performs better concerning the retrieval of the content-based clustering structure. Overall, fewer clusters and more rating categories, i.e. a larger rating scale, yields better CCRS results for both content and response-style-based clusters. In addition, we showed that the performance of CCRS is not strongly affected by an increase in the number

of response styles D and a decrease in the sample size n.

Using an empirical dataset, we illustrated that CCRS yields different content and response-style-based clusters whereas both CDS tandem and k-means lead to content-based clusters that are hard to distinguish from response-style-based clusters. Obviously, the results of the empirical data are, as is often the case in cluster analysis studies, difficult to validate. Nonetheless, these results do illustrate that the potential challenge associated with existing methods (i.e., identifying clusters that are merely related to response tendencies) can be mitigated with the proposed approach. Moreover, we implemented an R package, ccrs, which can be download from https://cran.r-project.org/web/packages/ccrs/.

There are many opportunities for future work based on CCRS. At first, in this paper, only content-based clusters that differ from response-style-based clusters were extracted in this paper; however, if there exists a response-style-like content-based cluster (such as a content-based cluster of mostly midpoint values), additional tools, such as Anchoring vignette (King, Murray, Salomon, & Tandon, 2004), may be required to distinguish them.

In addition, for the new framework described in Section 2.2, only the relationship with CDS was considered in this paper. However, we should consider its relationship with probability-based model, such as IRT methods as well as the method proposed by van Rosmalen et al. (2010). Such an evaluation could possibly result in a very general framework for correction that includes various existing correction methods and that would facilitate a comparison of the correction accuracies of different correction methods.

In Chapter 3, we have proposed a new approach to incorporate and interpret external information in a biplot for categorical data. Specifically, we introduce a multiple-set extension to cluster CA, MSCCA, that can visually establish the relationship between external information and categories. In MSCCA, unlike the averaging approach, the class-specific clusters obtained make it possible to identify heterogeneous tendencies within each class. In addition, by simultaneously biplotting clusters in different classes in a common low dimension space, the relationships among classes can be perceived in a single MCA biplot. Moreover, we show how MSCCA relates to the existing linear row constraint framework, discussed in Hwang and Takane (2002). Note that MSCCA is especially useful when there are many individuals in each class which is of interest.

To investigate the performance of this proposed method, we consider different conditions, according to a simulation study. The results show that increasing the number of supplementary variables H has little effect on cluster or biplot accuracies. However, the results are better if the supplementary variables feature few categories and a balanced distribution over categories.

Then with an empirical analysis of road accident data, we show that that the averaging and cluster CA approaches can uncover only tendencies corresponding to the majority of accidents in each class. The MSCCA biplot instead makes it possible to interpret heterogeneous tendencies within each class, regardless of cluster sizes.

Finally, MSCCA introduced in Section 3.2 can be applied to different settings. In particular, it could be adopted in a three-way setting to depict the relationship among multiple two-way data sets. For example, if we have $n \times m$ categorical data sets corresponding to

T different time points, we could use MSCCA to reveal the relationships among clusters at different times.

Since the increasing number of large-scale categorical data has been obtained recently, it is important to capture the latent structure of data for simple interpretation. Therefore, we can expect that these two proposed methods help data analyser interpret the result of data analysis which has been considered complicated.

Aknowledgement

I would like to express my sincere gratitude to Prof. Dr. Hiroshi Yadohisa for generously providing an opportunity to start studying statistics, and keep on researching in his laboratory. Also, I am indebted to Professor Dr. K. Kawasaki, Professor Dr. M. Jin, Professor Dr. Y. Zheng, and Accosiate Professor Dr. K. Okada for their helpful comments, which greatly improved this paper.

In addition, I also would like to show my deep gratitude to Dr. Michel, van de Velden, for providing me an opportunity to do research with him and accepting me in his university, Erasmus University Rotterdam, as a visiting researcher. Also I am very grateful to my laboratory members for accepting me in the laboratory.

Finally, I would like to show my deep gratitude to my family for their help.

References

Agresti, A. (2013). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons.

Arabie, P. (1994). Cluster analysis in marketing research. *Advanced methods of marketing research*, 160–189.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*, 1235–1245.

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

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678.

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*, 69–83.

Böckenholt, U., & Böckenholt, I. (1990). Canonical analysis of contingency tables with linear constraints. *Psychometrika*, *55*(4), 633–639.

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*, 159–181.

Böckenholt, U., & Takane, Y. (1994). Linear constraints in correspondence analysis. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence analysis in social sciences* (pp. 112–127). London: Academic Press.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352.

Chang, Y. H., Iwai, N., Li, L., & Kim, S. W. (2014). *East Asian Social Survey (EASS), cross-national survey data sets: Culture and Globalization in East Asia, 2008*. Ann Arbor, MI: EASSDA, Inter-university Consortium for Political and Social Research (ICPSR).

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187–212.

Finch, W. H., & French, B. F. (2012). Parameter estimation with mixture item response theory models: A monte carlo comparison of maximum likelihood and bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 14.

Gabriel, K. R. (2002). Goodness of fit of biplots and correspondence analysis. *Biometrika*, *89*(2), 423–436.

Gower, J., Groenen, P., & van de Velden, M. (2010). Area biplots. *Journal of Computational and Graphical Statistics*, *19*(1), 46–61.

Gower, J. C., & Hand, D. J. (1996). *Biplots.* London: Chapman & Hall.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics*, *20*(2), 251–269.

Haskell, K. H., & Hanson, R. J. (1981). An algorithm for linear least squares problems with equality and nonnegativity constraints. *Mathematical Programming*, *21*, 98–118.

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

Hwang, H., Dillon, W. R., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, *71*(1), 161–171.

Hwang, H., & Takane, Y. (2002). Generalized constrained multiple correspondence analysis. *Psychometrika*, *67*(2), 211–224.

Hwang, H., Yang, B., & Takane, Y. (2005). A simultaneous approach to constrained multiple correspondence analysis and cluster analysis for market segmentation. *Asia Pacific Advances in Consumer Research*, *6*, 197-199.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, *98*(1), 191–207.

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, *44*, 23–34.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57–64.

MacQueen. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*, 1539–1550.

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, *24*(1), 27–34. doi: https://doi.org/10.1002/job.2112

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*, 271–298.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology:*

European Journal of Research Methods for the Behavioral and Social Sciences, *8*(4), 159–170. doi: https://doi.org/10.1027/1614-2241/a000048

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

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*, 425–441.

Ramsay, J. O., & Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, *84*(408), 906–915.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75–92.

Schoonees, P. C. (2016). *cds: Constrained dual scaling for detecting response styles.* Retrieved from `https://cran.r-project.org/web/packages/cds/` (R package version 1.0.3)

Schoonees, P. C., van de Velden, M., & Groenen, P. J. (2015). Constrained dual scaling for detecting response styles in categorical data. *Psychometrika*, *80*, 968–994.

Stukovsý, R., Palat, M., & Sedlakova, A. (1982). Scoring position styles in the elderly. *Studia Psychologica*, *24*, 145.

Takane, Y., & Hwang, H. (2002). Generalized constrained canonical correlation analysis. *Multivariate Behavioral Research*, *37*(2), 163–195.

Takane, Y., & Shibayama, T. (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika*, *56*(1), 97–120.

Takane, Y., Yanai, H., & Mayekawa, S. (1991). Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika*, *56*(4), 667–684.

Van Buuren, S., & de Leeuw, J. (1992). Equality constraints in multiple correspondence analysis. *Multivariate behavioral research*, *27*(4), 567–583.

Van Buuren, S., & Heiser, W. J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, *54*(4), 699–706.

van de Velden, M. (2000). Dual scaling and correspondence analysis of rank order data. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in Multivariate Statistical Analysis* (pp. 87–99). Dordrecht, Netherland: Kluwer Academic Publisher.

van de Velden, M. (2008). Detecting response styles by using dual scaling of successive categories. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New Trends in Psychometrics* (pp. 517–524). Tokyo, Japan: Universal Academy Press.

van de Velden, M., D'Enza, A. I., & Palumbo, F. (2017). Cluster correspondence analysis.

Psychometrika, *82*(1), 158–185.

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

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch models* (pp. 99–115). New York, NY: Springer.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*, 96–110.

Yanai, H. (1986). Some generalizations of correspondence analysis in terms of projectors. In E. Diday, Y. Escoufier, L. Lebaxt, J. E. Pages, Y. Schektman, & R. Thomassone (Eds.), *Data analysis and informatics IV* (pp. 193–207). Amsterdam: North Holland.

Yanai, H. (1988). Partial correspondence analysis and its properties. In C. Hayashi, M. Jambu, E. Diday, & N. Ohsumi (Eds.), *Recent developments in clustering and data analysis* (pp. 259–266). Boston: Academic Press.

Yanai, H., & Maeda, T. (2002). Partial multiple correspondence analysis. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefuji (Eds.), *Measurement and Multivariate Analysis* (pp. 57–68). Tokyo: Springer.