Visualizing Class Specific Heterogeneous Tendencies in Categorical Data

Mariko Takagishi and Michel van de Velden

Graduate School of Engineering Science, Osaka University, Japan; Department of Econometrics, Erasmus University Rotterdam, Netherlands

ABSTRACT

In multiple correspondence analysis, both individuals (observations) and categories can be represented in a biplot that jointly depicts the relationships across categories and individuals, as well as the associations between them. Additional information about the individuals can enhance interpretation capacities, such as by including class information for which the interdependencies are not of immediate concern, but that facilitate the interpretation of the plot with respect to relationships between individuals and categories. This article proposes a new method which we call multiple-class cluster correspondence analysis that identifies clusters specific to classes. The proposed method can construct a biplot that depicts heterogeneous tendencies of individual members, as well as their relationships with the original categorical variables. A simulation study to investigate the performance of the proposed method and an application to data regarding road accidents in the United Kingdom confirms the viability of this approach. Supplementary materials for this article are available online.

1. Introduction

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; Gower and Hand 1996).

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 and Böckenholt 1990; Takane, Yanai, and Mayekawa 1991; Van Buuren and de Leeuw 1992; Böckenholt and Takane 1994; Yanai and Maeda 2002; Hwang, Yang, and Takane 2005). Hwang and Takane (2002) also show that various objectives for incorporating the external information can be generalized into a linear constraint framework, which they call generalized constrained MCA (GCMCA). 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 visualize how individuals’ tendencies differ depending on these classes, we could integrate external variables before applying MCA, but this approach transforms the information, such that it 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, points corresponding to men and women could be colored differently. 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 nonexternal variables, which, from here on, we shall refer to as “analysis” variables, we can depict the relationship between the external information and the categories. We refer to this as the averaging approach and we show in Section 2.4, that this approach can be regarded as a special case that fits into the linear constraint framework proposed by Hwang and Takane (2002).

The averaging approach only reveals average tendencies of many individuals within a class, obscuring their heterogeneous tendencies. When, for example, 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 only a minority, such tendencies could still be interesting to consider, especially to characterize tendencies by class.

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 article is organized as follows. In Section 2, we introduce our proposed method and its relationship with existing approaches, including the linear row constraint framework. Then in Section 3, we compare a biplot obtained using the averaging approach and one obtained using our proposed method. A simulation study in Section 4 appraises the proposed method. A simulation study in Section 4 appraises the proposed method. Finally, we illustrate our method by applying it to empirical data in Section 5.

2. Multiple Class Cluster Correspondence Analysis

In this section, we introduce our approach, which we call multiple-class cluster correspondence analysis (MCCCA from here on). We discuss how our new method is related to several existing methods, such as the averaging approach and cluster correspondence analysis (cluster CA, Van Buuren and Heiser 1989; van de Velden, D’Enza, and Palumbo 2017), which simultaneously applies dimension reduction and cluster analysis. In fact, we show that, although cluster CA is not a method to incorporate external information, mathematically, MCCCA can be seen as an extension of cluster CA.

2.1. The MCCCA Objective Function

Suppose that we have $N$ observations on $J$ 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 external variables, and categories in external variables as external categories.

Table 1 summarizes some of the notation required to formulate the MCCCA objective function. Note that the rows of the category indicator matrices $Z_j$ are $(q_j \times 1)$ vectors $z_{ij}^{(j)} = (z_{ij}^{(j)}; i = 1, \ldots, N; \ell = 1, \ldots, q_j)$, where $z_{ij}^{(j)} = 1$ if individual $i$ chooses the $\ell$th category in the $j$th variable, and the other elements are 0. Category indicator matrices for external variables $V_h$ are similarly defined.

MCCCA amounts to the following: Given $(N \times J)$ categorical data, individuals are divided into $C$ known classes. Here, a class may be defined as either a specific category of an external variable or as a combination of categories of some or all of the $H$ external categories. For example, an individual choosing the first category for all $H$ external variables belongs to the $c_1 = 1, c_2 = 1$, and $c_3 = 1$ class. Within each class, individuals are assigned to $K_{c_1=\cdots=c_H}$ unknown class-specific clusters ($c_h = 1, \ldots, C_h; h = 1, \ldots, H$) and a low-dimensional configuration of class-specific clusters and categories of $J$ active variables is obtained.

To achieve this, the objective function of MCCCA is defined as

$$\min_{U,G,B_j} \phi(U,G,B_j | Z_j, V_h) = \frac{1}{N^J} \sum_{j=1}^J \|UG - Z_j B_j\|^2$$

$$\quad \text{s.t.} \quad \frac{1}{N^J} \sum_{j=1}^J B_j' Z_j' Z_j = I_p, \quad M_N U G = U G$$

where $U = (U_{11}, \ldots, U_{C_1 C_2 \cdots C_H})$.

Here, $M_N = I_N - N^{-1} I_N$ is the centering matrix, $I_N$ is an $N \times N$ identity matrix, and $I_N$ is an $N \times 1$ vector of ones. When we estimate parameters, the number of clusters $K_{c_1=\cdots=c_H}$ and $p$ must be prespecified.

Unlike existing approaches incorporating external information, a two-level hierarchical structure is defined on $U$, with known classes in the first level and unknown clusters in each class in the second level.

Table 1. Some notations in MCCCA.

| Symbol | Description |
|--------|-------------|
| $N$    | the number of individuals. |
| $J$    | the number of active variables. |
| $H$    | the number of external variables. |
| $q_j$  | the number of categories for the $j$th active variable. |
| $Q$    | the total number of categories among $J$ active variables. |
| $C_h$  | the number of categories for the $h$th external variable. |
| $C$    | the total number of classes (e.g., if $C_1 = 2, C_2 = 2, C = 4$), the number of categories in the first and second external variables) |
| $p$    | the index for the categories of the $h$th external variable. |
| $Z_j$  | (e.g., the "$c_1 = 1$ and $c_2 = 1$" class indicates the class consisting of the first category of the first and second external variables) |
| $B_j$  | the number of clusters for the $(c_1, \ldots, c_H)$ class. |
| $U$    | the total number of clusters among all classes. |
| $G$    | the number of dimensions. |
| $G_{c_1 \cdots c_H}$ | the $N \times p$ quantification matrix for the $j$th active variable. |
| $G_{c_1 \cdots c_H}$ | the $N \times K$ cluster indicator matrix, where each $U_{c_1 c_2 \cdots c_H}$ is an $(N \times K_{c_1 c_2 \cdots c_H})$ matrix. |
| $K_{c_1 \cdots c_H}$ | the $K \times p$ quantification matrix for cluster centers, where each $G_{c_1 c_2 \cdots c_H}$ is a $(K_{c_1 c_2 \cdots c_H} \times p)$ matrix. |
To illustrate the construction of $U$ we use a small example. Suppose that we have $N = 8$ observations and two external variables (that is, $H = 2$), corresponding to two binary variables. In this case $C_1 = C_2 = 2$. Let the category indicator matrices for the two external variables be

$$V_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}. $$

Here, each column in $V_1$, and $V_2$ corresponds to the external categories of each individual variable. Considering all combinations of these external categories results in four classes, because $C = C_1 C_2 = 4$. Also, $U$ in Equation (2) will be $U = (U_{11}, U_{21}, U_{12}, U_{22})$, and $U_{11}$ (equivalently, $U_{21}$ for $c_1 = 1, c_2 = 1$) indicates the cluster indicator for the class corresponding to the first categories of both variables. Suppose that we want to find two clusters for all classes, so that $K_{c_1 c_2} = 2$ for $c_1 = 1, 2$ and $c_2 = 1, 2$ (i.e., $K_{11} = K_{21} = K_{12} = K_{22} = 2$). Then, each $U_{c_1 c_2}$ ($c_1 = 1, 2; c_2 = 1, 2$) in $U$ is $(N \times 2)$ matrix. Let $u_i = (u_i^{(1)}, u_i^{(2)}, u_i^{(12)}, u_i^{(22)})$ denote the $i$th row of $U$. Since, according to $V_1$ and $V_2$, the first observation $i = 1$ corresponds to the $c_1 = 1$ and $c_2 = 1$ class,

$$u_i^{(1)} = \begin{cases} (1, 0) & \text{if observation } i = 1 \text{ belongs to} \\
(0, 1) & \text{if observation } i = 1 \text{ belongs to} \end{cases}$$

2nd cluster of the $c_1 = 1$ and $c_2 = 1$ class

while the parts corresponding to other classes are $u_i^{(2)} = u_i^{(12)} = u_i^{(22)} = (0, 0)$. Similarly, if the sixth observation ($i = 6$), who is in the class corresponding to the second categories for both variables belongs to the second cluster in this class, the sixth row of $U$ will be

$$u_6 = (u_6^{(1)}, u_6^{(2)}, u_6^{(12)}, u_6^{(22)}) = (0, 0, 0, 0, 0, 0, 1)' .$$

Thus, the external variables in MCCCA partition $U$ in Equation (2) into different classes, after which class-specific clusters can be obtained.

In Figure 1, we present a schematic overview of how MCCCA incorporates the external categories for the small example above. As can be seen, in MCCCA, the external categories are used to split the data into different classes so that class-specific clusters are obtained. Then all class-specific clusters and categories are plotted in a single biplot, that allows for a depiction of the heterogeneous tendencies in each class.

If there are no external variables, there will be no first level with known classes in Figure 1, and, consequently, no class-specific clusters. In that case, $U$ in Equation (2) is a single cluster indicator matrix (i.e., $U = (U_{11}))$, and Equation (1) reduces to cluster CA (van de Velden, D’Enza, and Palumbo 2017, p. 160).

Hence, MCCCA is an extension of cluster CA in which cluster allocations are simultaneously obtained for each class in a common low-dimensional space, where the category quantities $B_j (j = 1, \ldots, f)$ are optimally estimated for all clusters.

In the averaging approach, as described in Section 1, there is no clustering at the second level of Figure 1. Instead, the data are partitioned using external variables, and the average quantification for all individuals in each class is obtained.

### 2.2. Algorithm

To estimate the parameters $U, G$, and $B_j (j = 1, \ldots, f)$, we use an alternating least squares algorithm. The updating formulas come from a direct extension of cluster CA (van de Velden, D’Enza, and Palumbo 2017).

**Step 1: Initialization.** Determine $K_{c_1 \ldots c_H}$ ($c_h = 1, \ldots, C_h; h = 1, \ldots, H$) and $p$. Set the number of iterations to $t = 0$, and set a convergence criterion $\epsilon$. Then, randomly generate initial clusters for each class.

**Step 2: Update $B$.** Let $B = (B_1', \ldots, B_f')'$ and $Z = (Z_1, \ldots, Z_f)$. Then find $B^{[t+1]}$ as

$$B^{[t+1]} = \sqrt{N}D^{-1/2}B' ,$$

where

$$\frac{1}{N} D^{-1/2}Z'M_N U^{(t)}U^{(t)} - U^{(t)}M_NZD^{-1/2} = B A B' ,$$

$$D = \tilde{Z}Z, \quad \tilde{Z} = b-diag(Z_1, \ldots, Z_f) .$$

**Step 3: Update $G$.** Obtain $G^{[t+1]}$ as follows:

$$G^{[t+1]} = \frac{1}{f} (U^{(t)}U^{(t)})^{-1}U^{(t)}M_N ZB^{[t+1]} .$$

**Step 4: Update $U$.** To obtain $U^{[t+1]}$, the update proceeds row-wise. Specifically, suppose individual $i$ belongs to the $(c_1, \ldots, c_H)$ class. In that case, the $i$th row of $U^{[t+1]}$ is updated as

$$u_{ik}^{(c_1 \ldots c_H)} = \begin{cases} 1 & \text{if } k = \arg \min_{l \in \{1, \ldots, K_{c_1 \ldots c_H}\}} \|f_j - g_l^{(c_1 \ldots c_H)}\|_2^2, \\
0 & \text{otherwise} \end{cases} ,$$

$k = 1, \ldots, K_{c_1 \ldots c_H}$.
where $\mathbf{g}_{k}^{(i)}$ is the $k$th element of the $i$th rows of $\mathbf{U}_{c_{1}, \ldots, c_{t}}$, while the $i$th row of all other cluster indicator matrices are set to 0.

So, for the small example described in Section 2.1, if $i$ belongs to the $c_{1} = 1$ and $c_{2} = 1$ class, the $i$th row of $\mathbf{U}_{11}$ is updated according to Equation (3). (Note that in this case, $g_{k}^{(i)} = 0$ indicates the $\ell$th class quantification for the $c_{1} = 1$ and $c_{2} = 1$ class). On the other hand, the $i$th row of the other cluster indicator matrices, $\mathbf{U}_{21}, \mathbf{U}_{12}, \mathbf{U}_{22}$, are all set to 0.

**Step 5: Convergence test** Compute $\phi^{[t]}$, the value of the objective function from Equation (1), using updated parameters. For $t > 1$, if $\phi^{[t]} - \phi^{[t-1]} < \epsilon$, terminate; otherwise, let $t = t + 1$ and return to Step 2.

### 2.3. Biplots

In this section, we show how MCCA can be used to construct a biplot. In van de Velden, D’Enza, and Palumbo (2017), cluster CA is formulated as a maximization problem. Accordingly, the MCCA in Equation (1) can also be rewritten as the following maximization problem:

$$
\max_{\mathbf{U}, \mathbf{B}} \psi(\mathbf{U}, \mathbf{B} \mid \mathbf{Z}) = \text{trace}(\mathbf{Z}' \mathbf{M}_{N} \mathbf{U}'(\mathbf{U}' \mathbf{U})^{-1} \mathbf{U}' \mathbf{M}_{N} \mathbf{ZB}) \quad \text{(4)}
$$

s.t. 
$$
\frac{1}{N} \sum_{j=1}^{I} \mathbf{B}'_{j} \mathbf{Z}_{j} \mathbf{B}_{j} = I_{p}.
$$

The proof for the equivalence of two optimization problems, Equations (1) and (4), is in Proposition B.1 of Appendix B in the supplementary material. When we leave $\mathbf{U}$ fixed, maximizing Equation (4) is equivalent to minimizing

$$
\min_{\mathbf{G}, \mathbf{B}} \phi^{\text{CA}}(\mathbf{G}, \mathbf{B} \mid \mathbf{Z}, \mathbf{V}, \mathbf{U}) = \|\tilde{\mathbf{P}} - \mathbf{D}_{k}^{1/2} \mathbf{G}^{T} \mathbf{D}_{k}^{-1/2}\|^{2} \quad \text{(5)}
$$

s.t. 
$$
\frac{1}{N} \mathbf{B}' \mathbf{D}_{k} \mathbf{B} = I_{p},
$$

where $\tilde{\mathbf{P}} = \mathbf{D}_{k}^{-1/2} (\mathbf{P} - \mathbf{r} \mathbf{c}') \mathbf{D}_{k}^{-1/2}$

$$
\mathbf{P} = \frac{1}{N} \mathbf{U}' \mathbf{Z}, \quad \mathbf{r} = \mathbf{P} \mathbf{1}_{Q}, \quad \mathbf{c} = \mathbf{P} \mathbf{1}_{K},
$$

$$
\mathbf{D}_{k} = \text{diag}(\mathbf{r}), \quad \mathbf{D}_{\ell} = \text{diag}(\mathbf{c}).
$$

The proof of the equivalence is available from van de Velden, D’Enza, and Palumbo (2017). Here, $\mathbf{P}$ indicates a $K \times Q$ scaled contingency table of clusters for each class (row) and category (column), and each element in $\mathbf{r}'$, $C_{\ell}C_{k}(k = 1, \ldots, K; \ell = 1, \ldots, Q)$, indicates the scaled expected frequency with an assumption of independence between the $k$th cluster and the $\ell$th category. Thus, the matrix $\tilde{\mathbf{P}}$ represents the standardized deviations from the assumption of independence between cluster membership and the categorical variables.

From Equation (5), it follows that the inner product of $\mathbf{D}_{k}^{1/2} \mathbf{G}$ and $\mathbf{D}_{k}^{-1/2} \mathbf{B}$ approximates the matrix of standardized deviations from independence, $\tilde{\mathbf{P}}$. That is, in MCCA, we can use $\mathbf{G}$ and $\mathbf{B}$ to construct a biplot in which the greater the inner product of the $k$th row vector of $\mathbf{G}$ and the $\ell$th row vector in $\mathbf{B}$ generally indicates a stronger association between the $k$th cluster and the $\ell$th category.

Note that in the resulting biplot, the points of the row and column are not necessarily similarly spread (e.g., Gower, Groenen, and 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, D’Enza, and Palumbo (2017) for detail.

### 2.4. Relationship to the 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, in a linear row constraint framework, which they call GCMCA. Note that this GCMCA framework can be considered as a special casR of canonical correspondence analysis Ter Braak (1986) and Takane and Hwang (2002). In this section, we briefly review these constrained MCA approaches and their relationship to MCCA.

To add linear row constraints in MCA, the following objective function can be formulated:

$$
\min_{\mathbf{F}, \mathbf{B}} \phi^{\text{const}}(\mathbf{F}, \mathbf{B} \mid \mathbf{Z}_{p}, \mathbf{V}_{h}) = \frac{1}{N} \sum_{j=1}^{I} \|\mathbf{RF} - \mathbf{Z}_{j} \mathbf{B}_{j}\|^{2} \quad \text{(7)}
$$

s.t. 
$$
\frac{1}{N} \sum_{j=1}^{I} \mathbf{B}_{j}' \mathbf{Z}_{j} \mathbf{B}_{j} = I_{p}, \quad \mathbf{M}_{N} \mathbf{RF} = \mathbf{RF},
$$

where $\mathbf{R}$ is the $N \times N$ matrix that contains linear row constraints for the quantifications. If $\mathbf{R} = \mathbf{I}$, the problem reduces to the homogeneity formulation of MCA. The choice of $\mathbf{R}$ depends on the objective that underlies the incorporation of the external information.

For example, let the $(N \times C)$ matrix $\mathbf{W}$ denote a class indicator matrix (that is, if an individual $i$ belongs to the $r$th class, the $(i, r)$th element of $\mathbf{W}$ is one whereas all elements in that row are zero). If we use $\mathbf{R} = \mathbf{W}(\mathbf{W}' \mathbf{W})^{-1} \mathbf{W}'$, then Equation (7) amounts to the averaging approach described previously. To see this, note that the class (a combination of external categories) gets represented by the average quantification of individuals corresponding to that class.

Alternatively, if we aim to “remove” the effect of external information from a biplot, we can choose $\mathbf{R} = I_{N} - \mathbf{W}(\mathbf{W}' \mathbf{W})^{-1} \mathbf{W}'$ (e.g., Takane and Shibayama 1991; Takane and Hwang 2002; Hwang and Takane 2002), which is equivalent to deducting the class conditional means from the data. For example, if as external variable we have gender, the mean of all males is deducted from all male observations.

Although MCCA follows a different motivation 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 $\mathbf{U}$, the MCCA objective function in Equation (1) can be rewritten as a minimization problem:

$$
\min_{\mathbf{F}, \mathbf{B}} \phi^{\text{MCCA}}(\mathbf{F}, \mathbf{B} \mid \mathbf{Z}, \mathbf{U}, \mathbf{V}) = \frac{1}{N} \sum_{j=1}^{I} \|\mathbf{RF} - \mathbf{Z}_{j} \mathbf{B}_{j}\|^{2} \quad \text{(8)}
$$

s.t. 
$$
\frac{1}{N} \sum_{j=1}^{I} \mathbf{B}_{j}' \mathbf{Z}_{j} \mathbf{B}_{j} = I_{p}, \quad \mathbf{M}_{N} \mathbf{RF} = \mathbf{RF},
$$

where $\mathbf{R} = \mathbf{U}(\mathbf{U}' \mathbf{U})^{-1} \mathbf{U}'$. 


and $U$ features the hierarchical cluster structure constraint explained in Section 2.1. From this formulation, it immediately follows that MCCCA represents a special case of Equation (7), with $R = U(U'U)^{-1}U'$. Proposition B.2 in Appendix B in the supplementary material offers a proof of the equivalence of Equations (7) and (8).

3. Numerical Illustration of an MCCCA Biplot

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

To start, we generate categorical data for 390 individuals that represent two categorical variables (meal and drink preference), and two external variables (nationality and gender). Table 2 contains the variables and corresponding categories, and Table 3 shows the contingency table with frequencies of meal and drink choices. 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 generated the data in such a way that there are four true clusters in the full dataset. Individuals in the first cluster choose “Meat” for the meal variable, and “Fruit juice” for the drink variable (M&J), those in the second cluster choose “Fish” and “Tea” (F&T), those in the third cluster, individuals choose “Rice” and “Tea” (R&T), and in the fourth cluster, choose “Rice” and “Milk” (R&M). The frequency distributions of the generated artificial data for each combination of external categories for each true cluster, are shown in Figure 2. We can see that there is one cluster for American males, and that there are two clusters for American females, Japanese males and Japanese females.

The biplot that results from the averaging approach, in Figure 3 (left), clearly reveals overall tendencies of many Americans and Japanese consumers, strongly associated with M&J, and F&T and R&T, respectively. However, the “Milk” category is located far from others, which makes it difficult to clearly specify who (i.e., which nationality or gender) makes this choice. In addition, it is difficult to specify the group of American females who choose F&T in the biplot, because it is not the largest group.

In contrast, by obtaining clusters for each class, the MCCCA biplot makes the tendencies of this relatively small number of individuals visible. When we use the correct number of clusters for each class, the MCCCA biplot result in Figure 3 (right) clearly reveals that a small number of female Japanese choose “Milk.” Similarly, a group (cluster) of American females who choose F&T is identified as “AF2” in the biplot. In addition, this biplot still depicts the tendencies of the larger groups, as obtained in the averaging approach. That is, MCCCA reveals the tendencies of small groups, without losing the information about the tendencies of larger groups.

In conclusion, while the averaging approach only reveals one kind of association between classes and main categories (e.g., the American female class is associated with Meat and Juice), MCCCA reveals more than one association between classes and categories (e.g., a cluster of American females showing association with Meat and Juice, and another, smaller cluster of American females associated with Fish and Tea). In addition, for the MCCCA result, we can see which active (nonexternal) categories are close to which class-specific clusters. Finally, by considering the sizes of these class-specific clusters, we can roughly infer the proportion of individuals in a class that have chosen those categories.

In Appendix C in the supplementary material we use a CA framework to provide some additional insights into how the MCCCA and averaging approaches differ with respect to their depiction of heterogeneous tendencies.

4. Simulation

We conducted a simulation study to evaluate the performance of MCCCA in different scenarios. In particular, the objective of the simulation is to know how the nature of external variables (e.g., the number of categories $C_i$, and whether the distribution over the categories is balanced) affects the accuracy of the clustering and the biplots obtained using MCCCA.
4.1. Data Generation

The data-generation process consists of two steps: generating an $N \times J$ data matrix, and generating $N \times H$ matrix of external variables. First, we start by dividing the $J$ variables into two groups: 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 nonnoise 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 $\hat{p} = (p_\ell) (\ell = 1, \ldots, q - 1)$, where $\hat{p} = ((1 - 0.8) \times (p_1, \ldots, p_{q-1})/ \sum_{\ell=1}^{q-1} p_\ell)$ 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 nonnoise to noise variables to $1 : 1$.

Second, to generate the data matrix corresponding to the $H$ external 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 $C_{h}/S$, where $C_{h}$ denotes the number of categories for the external variable, and $S = \sum_{h=1}^{H} S_{h}, (h = 1, \ldots, H)$.

4.2. Simulation Study Design

To assess the performance of the methods in different settings, we fix the number of observations $N = 500$ and the number of variables $J = 10$. 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 external variables $H = 1, 2$; and the number of categories for the external variables $C_{h} = 2, 3$. Finally, for the external variables we note the balanced and unbalanced scenarios. For each combination of parameters in the simulation, we randomly generate 50 different $N \times J$ data matrices and $N \times H$ external variable matrices. For each dataset, we apply MCCCA using 100 random initial values.

4.3. 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 Adjusted Rand Index (ARI, Hubert and Arabie 1985). The ARI assesses the similarity between two cluster allocations, so it takes a value of 1 for a perfect recovery, and this value decreases as performance worsens. 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 and Ten Berge 2006). The GF between configurations $Y$ and $H$ is defined as

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

Therefore, we calculate the GF between $Y$ and $H$, where $H = GB'$ (with $G$ and $B$ as the MCCCA solutions) and $Y = \tilde{P}'_{\text{true}} = D_y (P'_{\text{true} - rc}) D_y$, such that $P'_{\text{true}} = U'Z$ and $U$ 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.

4.4. Result

The results for the GF index in Figure 4 indicate that it tends to decrease as the number of categories $q$ increases. The number of external variables $H$ does not substantially affect the GF.
Rather, the GF tends to be somewhat better when there are fewer categories $C_h$ in the external variables and when the distribution over the categories is balanced.

The cluster retrieval results are shown in Figure 5. We see that the number of the external variables $H$ appears to affect the variance in the ARI results. In particular, the variance of the ARI for $H = 2$ is greater than that for $H = 1$. In addition, when there are more external categories, especially in the unbalanced scenarios, we observe more outlying results.

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

5. Application

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 MCCCA, the averaging approach, and cluster CA, to establish how each method would visualize the relationships.

5.1. Data and Setting

The data were obtained from the U.K. Department for Transport’s road safety statistics (https://www.gov.uk/government/collections/road-accidents-and-safety-statistics). 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 dataset contains $N = 3026$ observations.

Regarding the circumstances of the accident, we consider four (i.e., $J = 4$) variables: lighting conditions, weather conditions, road surface conditions, and speed limit. For the types of accident, we select two ($H = 2$) external variables: Casualty class and Area. Table 4 summarizes the variables and their categories.

To determine the dimensionality of the solution, we inspect the adjusted inertias obtained by applying MCA to the full dataset, as shown in Table 5. In this case, as the first and second dimensions account for around 61.3% of the adjusted total inertia. (Note that here we use the adjusted total inertia Greenacre (2017), which does not sum up to 100%, as it takes
Figure 5. Boxplot of ARI for each case.

Table 4. Categories for each variable and their corresponding labels in biplots and descriptions.

| Variable type          | Variable name | Label | Description                                      |
|------------------------|---------------|-------|--------------------------------------------------|
| Analysis variables     | Light conditions | L0    | Daylight                                         |
|                        |               | L1    | Darkness: street lights present and lit          |
|                        |               | L2    | Darkness: street lights present but unlit         |
|                        |               | L3    | 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 3 cm deep)             |
| Speed limit            | S30           |       | Speed limit is up to 30 km/h                      |
|                        | S70           |       | Speed limit is from 30 km/h to 70 km/h           |
| External variables     | Casualty class | Driver      | Casualty is one driver                           |
|                        |               | Ped     | Casualty is one pedestrian                        |
|                        | Area          | Urban   | Occurring in urban area                           |
|                        |               | Rural   | Occurring in rural area                           |

Table 5. Absolute and relative adjusted inertias per dimension for the MCA solution.

| Dimension | 1   | 2   | 3   | 4   | 5   | 6   | 7   | Total  |
|-----------|-----|-----|-----|-----|-----|-----|-----|--------|
| Adjusted inertias | 0.087 | 0.039 | 0.028 | 0.003 | 0.000 | 0.000 | 0.000 | 0.157  |
| Proportion of explained adjusted total inertia | 0.424 | 0.190 | 0.136 | 0.014 | 0.002 | 0.000 | 0.000 | 0.765  |
into account the unexplained inertia). Although three dimensions would explain more inertia, we choose \( p = 2 \) as this allows a straightforward visualization whilst still capturing a substantial part of the inertia.

In MCCA, the number of clusters must be prespecified for each class \( K_{c_1, \ldots, c_H} \) \((c_h = 1, \ldots, C_h; h = 1, \ldots, H)\). As is the case in any cluster analysis method, determining the number of clusters is a complex task due to the lack of a single clear and objective criterion to do so. Typically, the choice depends on subjective (e.g., interpretability) criteria combined with the use of internal cluster validity measures. For this study, we employed a pragmatic approach using the Krzanowski–Lai index (KL index, Krzanowski and Lai 1988). Specifically, in each class, we separately apply k-means to individual quantifications obtained by the MCA on the full data (i.e., the same MCA result as the one used to determine \( p \) mentioned above). Next, we select the number of class-specific clusters \( K_{c_1, \ldots, c_H} \) that corresponds to the optimal KL index. For these data, this procedure results in six clusters for the driver and urban class, six clusters for the pedestrian and urban class, three clusters in the driver and rural class, and three clusters for the pedestrian and rural class (i.e., \( K_{11} = 6, K_{12} = 6, K_{21} = 3, \) and \( K_{22} = 3 \)). From here on, we refer to a cluster from the driver and urban class as a D&U cluster, clusters from the pedestrian and rural class are P&R clusters, and so on.

To appraise the differences and added value of our MCCA approach, we also consider results from the averaging approach as well as cluster CA. Note that, as cluster CA is not a method that allows the incorporation of external variables, we include the external variables as active variables. That is, the Casualty type and Area variables play a role in the determination of clusters and quantifications. To select the number of clusters in the cluster CA approach, we also used the KL index resulting in \( K = 7 \) clusters.

### 5.2. Results

#### 5.2.1. MCCA

In the biplot for the MCCA solution (Figure 6), we see that categories located on the positive side of the horizontal axis correspond to “good” driving conditions (e.g., “Fine” and “~30”), while categories on the negative side indicate “bad” driving conditions (e.g., “Snow” and “Dark3”). Moreover, on the negative side of both horizontal and vertical axes, we find categories corresponding to “Rainy” conditions, such as “Rain” and “Wet.”

In addition, note that in the same direction as categories corresponding to “good” driving conditions (such as “Fine,” “Dark0,” and “Dry”), we find the largest P&U, P&R, and D&R clusters. This indicates that, in general, “fine weather” conditions occur more often than “bad conditions” (such as rainy and snowy) and, consequently, in absolute terms, more accidents tend to happen in such conditions. Especially, accidents involving pedestrian casualties have a strong association with good driving conditions because of this size effect.

On the other hand, the second largest D&R cluster (with a cluster size that is very similar to the largest one) is related to categories such as “Snow_w (weather conditions),” “Snow (road surface)” and “Fog.” This indicates that accidents in rural areas resulting in driver casualties is more likely to occur in “bad” conditions than for the other classes.

In addition, around the “Rain” and “Wet” categories, we find several smaller clusters for all classes, indicating that for all classes some accidents occurred in rainy conditions.

By inspecting the MCCA 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. However, accidents in rural areas involving drivers are

---

**Figure 6.** Results using MCCA. The numbered bubbles indicate cluster points with DU indicating clusters in the driver and urban class, PU corresponding to clusters in the pedestrian and urban class, DR to the driver and rural class, and PR to pedestrian and rural class clusters. The numbers and sizes of the bubbles reflect the size of each cluster within its class (e.g., “DU1” indicates the largest cluster in the driver and urban class). Character labels correspond to categories of light conditions (L), weather conditions (W), road surface conditions (R), and speed limits (S).
more strongly related to bad conditions, especially conditions involving snow.

5.2.2. Averaging

The results using the averaging approach are in Figure 7. Although we can still interpret the information regarding classes with respect to categories, the averaging of the results limits the available information. In particular, we see that D&R clusters are related to categories indicating bad driving conditions, while clusters in the other classes appear to be related to categories corresponding to good driving conditions. It is, however, difficult to interpret the heterogeneous tendencies within each class. For example, in Figure 7, it seems that no classes have an association with the categories “Snow_w” and “Dark3.” This may be because, even though the D&R class is in fact associated with both categories, it gets located in the middle of these categories and thus, the association is not clearly depicted. On the other hand, the MCCCA result in Figure 6 clearly shows that the large D&R cluster appears to have a strong association with both snowy and dark conditions. Similarly, in Figure 7, we cannot distinguish “good” from “rainy” conditions, possibly because accidents are likely to occur in both conditions and thus, categories related to both conditions are located close to each other. Thus, the averaging approach does not allow us to interpret specific tendencies within each class.

5.3. Summary of Results

In this section, we 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 MCCCA were we able to differentiate heterogeneous conditions in which many or few accidents occurred. Specifically, though many accidents in most classes occur even in good conditions, only by using MCCCA we were able to differentiate heterogeneous conditions in which many or few accidents occurred. Specifically, though many accidents in most classes occur even in good conditions, simply because good conditions occur more frequently, MCCCA reveals that the association of bad conditions with accidents in the D&R class is stronger than with other classes. Moreover, for most classes, we uncover relatively small clusters of accidents that relate strongly to rainy conditions.

Note that our method is primarily a descriptive tool that can help in gaining additional insights in a categorical dataset.
Cautions should be taken when trying to generalize the findings beyond the sample.

6. Conclusion

We proposed a new approach to incorporate and interpret external information in a biplot for categorical data. Specifically, we introduce a multiple-class extension to cluster CA, MCCCA, that can visually establish the relationship between external information and categories. In MCCCA, the class-specific clusters obtained make it possible to identify heterogeneous tendencies within each class. We also show how MCCCA relates to a linear row constraint framework.

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 external variables \( H \) has an effect on cluster accuracies. In contrast, the biplot accuracy is better if the external variables feature few categories and a balanced distribution over categories.

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

We implemented MCCCA in the R package “mccca” for the statistical computing environment R (R Core Team 2021), which can be downloaded from the comprehensive R archive network (Takagishi 2022, https://CRAN.R-project.org/package=mccca.)

Finally, MCCCA can be applied in different settings. In particular, it could be adopted in a three-way setting to depict the relationship among multiple two-way datasets. For example, if we have \( N \times J \) categorical datasets corresponding to \( T \) different time points, we could use MCCCA to reveal the relationships among clusters at different times.

Supplementary Materials

In supplementary materials, you can obtain:

Data: the datasets used in Section 3.
MCCCA_app.pdf: contains a short description of the datasets used in Section 3, and Appendix B and C.
The R package of MCCCA is available in https://cran.r-project.org/package=mccca.

Acknowledgments

The authors thank the Editor and the anonymous reviewer for their helpful comments and suggestions which have significantly improved the article.
The authors also wish to express our thanks to the U.K. Department for Transport's road safety statistics for providing data.

**Funding**

This work was supported by the Japan Society for the Promotion of Science KAKENHI grants 20K19755.

**ORCID**

Mariko Takagishi https://orcid.org/0000-0002-2984-8991

**References**

Böckenholt, U., and Böckenholt, I. (1990), "Canonical Analysis of Contingency Tables with Linear Constraints," *Psychometrika*, 55, 633–639.

Böckenholt, U., and Takane, Y. (1994), "Linear Constraints in Correspondence Analysis," in *Correspondence Analysis in Social Sciences*, eds. M. J. Greenacre and J. Blasius, pp. 112–127, London: Academic Press.

Gabriel, K. R. (2002), "Goodness of Fit of Biplots and Correspondence Analysis," *Biometrika*, 89, 423–436.

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

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

Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press.

——— (1993), "Biplots in Correspondence Analysis," *Journal of Applied Statistics*, 20, 251–269.

——— (2017), *Correspondence Analysis in Practice*, Boca Raton, FL: CRC press.

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

Hwang, H., and Takane, Y. (2002), "Generalized Constrained Multiple Correspondence Analysis," *Psychometrika*, 67, 211–224.

Hwang, H., Yang, B., and 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.

Krzanowski, W. J., and Lai, Y. (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., and Ten Berge, J. M. (2006), "Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity," *Methodology*, 2, 57–64.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/

Takagishi, M. (2022). "mccca: Visualizing Class Specific Heterogeneous Tendencies in Categorical Data." Available at https://CRAN.R-project.org/package=mccca

Takane, Y., and Shibayama, T. (1991), "Principal Component Analysis with External Information on Both Subjects and Variables," *Psychometrika*, 56, 97–120.

Takane, Y., Yanai, H., and Mayekawa, S. (1991), "Relationships Among Several Methods of Linearly Constrained Correspondence Analysis," *Psychometrika*, 56, 667–684.

Ter Braak, C. J. (1986), "Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis," *Ecology*, 67, 1167–1179.

Van Buuren, S., and Heiser, W. J. (1989), "Clustering n Objects into k Groups Under Optimal Scaling of Variables," *Psychometrika*, 54, 699–706.

Van Buuren, S., and de Leeuw, I. (1992), "Equality Constraints in Multiple Correspondence Analysis," *Multivariate Behavioral Research*, 27, 567–583.

van de Velden, M., D’Enza, A. L., and Palumbo, F. (2017), "Cluster Correspondence Analysis," *Psychometrika*, 82, 158–185.

Yanai, H. (1986), "Some Generalizations of Correspondence Analysis in Terms of Projectors," in *Data Analysis and Informatics IV*, eds. E. Diday, Y. Escoufier, L. Lebart, J. E. Pages, Y. Schektman, and R. Thomassone, pp. 193–207, Amsterdam: North Holland.

Yanai, H. (1988), "Partial Correspondence Analysis and Its Properties," in *Recent Developments in Clustering and Data Analysis*, eds. C. Hayashi, M. Jambu, E. Diday, and N. Ohsumi, pp. 259–266, Boston: Academic Press.

Yanai, H., and Maeda, T. (2002), "Partial Multiple Correspondence Analysis," in *Measurement and Multivariate Analysis*, eds. S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji, pp. 57–68, Tokyo: Springer.