Visualising Incomplete Data with Subset Multiple Correspondence Analysis

Johané Nienkemper-Swanepoel, Niël J. le Roux and Sugnet Gardner-Lubbe

Abstract Determining the cause of missing values is a challenge, but an important task in order to select correct analysis techniques for missing data. This paper presents a new approach to identify the missing data mechanism (MDM) by applying cluster analysis to biplots of data having missing observations. Subset multiple correspondence analysis (sMCA) enables an isolated analysis of a chosen subset while preserving the scaffolding of the original data set. Multivariate categorical data sets are frequently represented in a coded dummy matrix, referred to as an indicator matrix. Additional category levels can be created for the indicator matrix to account for the unobserved information which has the advantage of not forfeiting any observed information. The extended indicator matrix easily partitions a data set into observed and unobserved subsets. sMCA biplots are used for the visual exploration of the subsets. Configurations of the incomplete subsets enable the recognition of non-response patterns which could aid in the identification of a particular MDM. The missing at random (MAR) MDM refers to missing responses that are dependent on the observed

Johané Nienkemper-Swanepoel · Niël J. le Roux · Sugnet Gardner-Lubbe
Stellenbosch University, Private Bag X1, Matieland, 7600, South Africa
✉ nienkemperj@sun.ac.za
✉ njlr@sun.ac.za
✉ slubbe@sun.ac.za

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/33
ISSN 2363-9881

information and is expected to be identified by patterns and groupings occurring in the incomplete sMCA biplot. The missing completely at random (MCAR) MDM states that all observations have the same probability of not being captured which could be identified by a random cloud of points in the incomplete sMCA biplot. The partitioning around mediods (pam) clustering technique is used to establish the number of available clusters in an incomplete sMCA biplot. A simulation study confirmed that there is a difference in the number of sufficient clusters that can by identified from MAR and MCAR simulated data sets. A real data set is also explored and the MDM is identified using the results of the simulation study as guidelines.

1 Introduction

Techniques to handle missing values have been well documented over the past decades, identifying multiple imputation (MI) as the preferred approach to handle missing values in a majority of applications (Rubin, 2003). MI replaces each missing observations with multiple plausible response values, resulting in multiple completed data sets to use for standard analysis (Rubin, 1987). Most missing data techniques require an initial assumption to be made regarding the cause of the missing values. Missing values are believed to occur due to a random process referred to as the missing data mechanism (MDM) (Van Buuren, 2012). Missing values commonly occur in questionnaires, typically containing categorical variables, which could be due to the deliberate omission of sensitive questions which in some cases are related to completed questions in the questionnaire. This is an example of observations that are classified as being missing at random (MAR), since the missing values are conditional on observed responses in the data sets (García-Laencina et al., 2009; Schafer and Olsen, 1998). The missing completely at random (MCAR) MDM refers to data entries with the same probability to be unobserved. This implies that the cause of missingness is independent of the observed and unobserved data (Van Buuren, 2012). The MDM that will not be considered, is the missing not at random (MNAR) MDM. In this scenario missing values are conditional on unobserved responses, therefore, related to values that are not captured by the data.

A few general comments have to be made before stating the aim of this paper. Let us consider a general multivariate categorical data set in which a set of individuals (referred to as samples) are required to answer a list of categorical

questions (referred to as variables). Measurements on a categorical variable can only be one of a finite number of qualities like “agree”, “disagree” or “dont know”. These finite qualities are referred to as the category levels (CLs) of a variable. Typically, the categorical variables are represented as the columns of a matrix having the samples as its rows. Multiple correspondence analysis (MCA) is a multivariate categorical technique which enables the simultaneous exploration of the associations among samples and their categorical responses, which can be displayed in a biplot (Greenacre, 2010). Biplots consist of coordinates for the samples, one point for each sample in the data set, and category level points (CLPs), one for each CL per variable in the data set. Short distances between the sample points and CLPs indicate strong associations. Therefore, the similarities between samples and their specific responses can be explored in a single display (Gower et al., 2011).

The input data matrix can be adapted (Section 3.1) to distinguish between observed and missing information. The subset of missing information can be used in subset MCA (sMCA) which allows an isolated view of associations between the samples and missing CLPs (Greenacre and Pardo, 2006).

The aim of this paper is to determine whether the MDM can be identified as to be of type MAR or MCAR by using techniques to cluster the missing subsets in the sMCA biplot. It is hypothesized that if substantial clustering structures are identified between the CLPs of the sMCA of missing responses, this could be an indication of dependence which is caused by the MAR MDM. On the contrary, poorly classified clusters could be an indication of independence caused by the MCAR MDM.

2 Data

Simulated data sets are crucial for the evaluation of missing data techniques. The simulation protocol followed in this paper is presented in Section 2.1 while a real data set (Section 2.2) will be used to illustrate the application of the proposed methods.

2.1 Simulation Study

Categorical responses are simulated from a uniform distribution by dividing the generated response values into the number of required category levels (CLs) per variable. Consider the example in Table 1 for the allocation of CLs from simulated uniformly distributed values that are allocated to two CLs:

Fifteen possible data combinations are simulated using different sample sizes (100, 500, 1000, 2500 and 3000), numbers of response variables (5, 10 and 15) and randomly varying the number of CLs per variable between two and five. A thousand simulations per combination are generated. Missing values are inserted using the MAR and MCAR MDMs with 10 %, 30 % and 50 % missing values.

Table 1: Allocating CLs from continuous values.

Uniform Values		**Categorical Responses**
0.2655		A
0.3721	$(0; 0.5] \rightarrow$ A	A
0.5729	$(0.5; 1] \rightarrow$ B	B
0.8984		B
0.4543		A

The following conditions were used to insert the missing values with the MAR MDM:

- If CLs 1 or 2 are recorded in variable 1, delete the corresponding samples in variables 3 to $(p - 2)$, where p denotes the total number of variables.
- If the last two CLs are recorded in variable $p/2$, delete the corresponding samples in variables $(p/2 + 1)$.
- If the "middle" CL is recorded in variable 2, delete the corresponding samples in variables $(p - 2)$ to p with increments of 2.
- If the "middle" CL is recorded in variable 3, delete the corresponding samples in variables 2 to p with increments of 2.
- If the last CL is recorded in variable p, delete the corresponding samples in variable 2.

- If the same CLs are recorded for a sample in variables 1 and p, delete the corresponding samples in variable p.

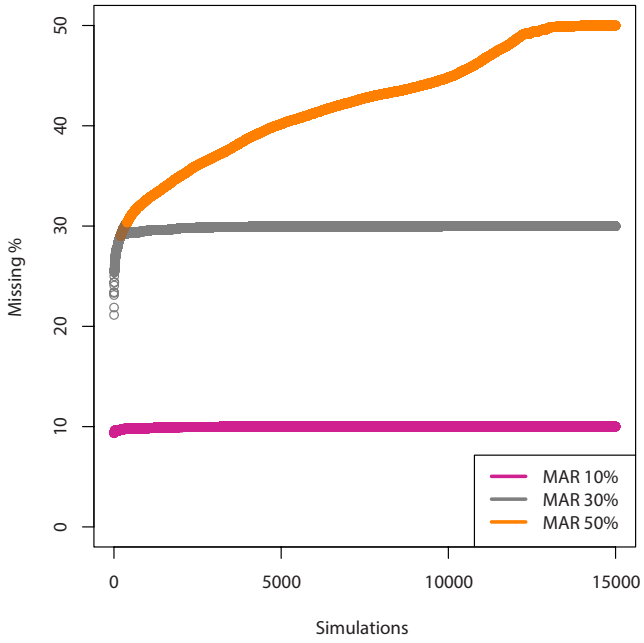

Figure 1: True percentages of missing values using MAR MDM in 15000 simulated data sets.

The true percentages of missing values that were obtained over 15000 data sets (1000 simulations for each of the 15 combinations) are shown in Figure 1. The percentages have been ordered to ease visual inspection.

Figure 1 provides an overall trend of the true percentages that are missing in the simulated sets. Since the percentage of missing values for a MAR MDM will depend on whether the given conditions are satisfied in a particular data set, the true percentage is expected to be underestimated in some of the missing data sets. The 10 % (on average true 9.97 %) and 30 % (on average true 29.88 %) missing values are correctly reflected in a majority of cases, as can be observed from the purple (bottom) and grey (middle) lines of Figure 1, whereas the conditions might not be flexible enough when the percentage of missing values increases to 50 % (on average true 42.2 %; see the orange (top) line of Figure 1).

Further inspection shows that in cases, where the number of variables increases, the true percentage of missingness is less likely to be reflected. This could be due to the additional variety of possible responses with lower frequency which results in the conditions not being satisfied. The MCAR MDM is inserted by drawing a random sample of the required percentage of missing values and resulted in a true reflection of the intended missing percentage.

2.2 Real Data Set

The International Social Survey Program (ISSP 1994) is considered for the real application. This survey investigated the family perspectives of changing gender roles in Germany using 11 questions with three possible CLs (Agree, Neutral and Disagree), as well as 5 demographic variables: region, gender, age, marital status and education. The adapted survey as used by Greenacre and Pardo (2006) is available from `www.carme-n.org/?sec=data`. The results presented in this paper are based on the survey consisting of 3291 samples of which 811 samples contain non-responses. Missing values occur in 25 % of the samples with an overall percentage of missing values of 5 %.

3 Methodology

In this section, we describe the main elements of the methodology, namely indicator matrices (Section 3.1), subset multiple correspondence analysis (sMCA; Section 3.2), and partitioning around medoids (pam; Section 3.3).

3.1 Indicator Matrix

The data matrix of a multivariate categorical data set is commonly coded as an indicator matrix ($\mathbf{G}$) of zeros and ones. The number of columns of the indicator matrix is dependent on the total number of observed categories (CLs). A one is allocated to an observed CL and zeros to the unobserved CLs per variable. A missing response will result in multiple zero elements in the indicator matrix. Since we are interested in the visualisation of the missing subset, the indicator

matrix must be adjusted to incorporate the unobserved CLs. There are two approaches to this particular handling of missing values: single active and multiple active. Single active handling creates one additional CL per variable for the missing responses, whereas multiple active handling creates a unique CL for each sample with a missing response for a particular variable. Consider the following example of a single categorical variable with three possible CLs, shown in Table 2.

Table 2: Multivariate categorical toy data set.

	Variable 1
Sample 1	?
Sample 2	3
Sample 3	1
Sample 4	?

Single active handling of the missing responses is illustrated in $\mathbf{G}_{sing}$ and multiple active handling in $\mathbf{G}_{mult}$ (missing values displayed in **bold** numbers):

$$\mathbf{G}_{sing} = \begin{bmatrix} V1:1 & V1:2 & V1:3 & V1:? \\ 0 & 0 & 0 & \mathbf{1} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{bmatrix} \tag{1}$$

$$\mathbf{G}_{mult} = \begin{bmatrix} V1:1 & V1:2 & V1:3 & V1:s1? & V1:s4? \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix} \tag{2}$$

The single active method assumes that all samples with missing values for a particular variable are similar by pooling the missing responses in the same CL. This might be a biased representation of the samples in the data set. Multiple active handling resolves this problem, but has the disadvantage of creating a large number of CLs with low frequency. The decision of single- and multiple active data handling should be based upon the data and the

specific research aim (Van de Geer, 1993). Furthermore, it is to be noted that the extended indicator matrix does not mean that missing values are imputed but rather they are accommodated by treating them as additional levels (categories) of the original categorical variables.

3.2 Subset Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) is applied by performing Correspondence Analysis (CA) on the subset of missing responses from the adjusted indicator matrix ($\mathbf{G}_h$). The indicator matrix is transformed by the diagonal matrices containing the row ($\mathbf{R}$) and column ($\mathbf{C}$) margins before applying the singular value decomposition (SVD). After transforming $\mathbf{G}$:

$$\mathbf{S} = \mathbf{R}^{-1/2}_{(n\times n)}\mathbf{G}_{(n\times h)}\mathbf{C}^{-1/2}_{(h\times h)} \tag{3}$$

the SVD of $\mathbf{S}$ follows as

$$\mathbf{S} = \mathbf{U}_{(n\times r)}\mathbf{\Sigma}_{(r\times r)}\mathbf{V}^{T}_{(r\times h)}\,, \tag{4}$$

where n is the number of samples, r the reduced dimension size and h the number of CLs in the missing subset. The singular vectors are represented in $\mathbf{U}$ and $\mathbf{V}$ with the singular values represented in decreasing order in $\mathbf{\Sigma}$ (Greenacre, 2017; Greenacre and Pardo, 2006).

The sMCA biplot is constructed by plotting the samples using the first two columns of $\mathbf{R}^{-1/2}_{(n\times n)}\mathbf{U}_{(n\times 2)}\mathbf{\Sigma}_{(2\times 2)}$ and the CLPs using the first two columns of $\mathbf{C}^{-1/2}_{(h\times h)}\mathbf{V}_{(h\times 2)}$ (Gower et al., 2011). The original column and row masses for the calculation of the distances are maintained when using a subset of the indicator matrix. Therefore, the total variation (inertia) is partitioned into components associated with the various subsets and no interpretable information is lost (Greenacre and Pardo, 2006). The *ca* package in R can be used for the calculations of MCA and sMCA (Nenadić and Greenacre, 2007; R Core Team, 2018).

3.3 Partitioning Around Medoids

The well-known partitioning around medoids (pam) method (Kaufman and Rousseeuw, 1987) is implemented to identify distinguishable clusters between the CLPs in the sMCA biplots of the missing subsets. A medoid is referred to as a representative object which has the shortest average distance to the other data points of interest. The data points closest to the medoid form a cluster. The pam method is applied to dissimilarities (distances) and therefore does not rely on distributional constraints, as is the case with the k-means method (Kaufman and Rousseeuw, 1987; Struyf et al., 1997). Moreover, in minimizing a sum of squared dissimilarities instead of a sum of squares it is a more robust method and importantly, it allows the algebraic selection of the number of clusters. We are interested in determining whether a sufficient clustering structure exists for the CLPs, since this could lead to emphasizing the association between missing responses and subsequently identifying the MDM. Since cluster analysis is applied on the reduced dimension sMCA solution, this is regarded as a tandem clustering approach (Mitsuhiro and Yadohisa, 2015).

In order to determine whether the number of predetermined medoids successfully discriminates between the clusters, the average silhouette width is evaluated. The silhouette value is obtained by first calculating the average dissimilarity of all objects in a specific cluster, say **C1**, to its medoid and then by identifying the closest neighbouring cluster, say **C2**, for each object in **C1**. The silhouette value provides a ratio between the distance to the medoid of the allocated cluster and the second-best option for each object (CLP in our context). Silhouette values are calculated for all data points and then averaged to provide a global measure of fit, referred to as the average silhouette width or silhouette coefficient, $s(i)$. Silhouette coefficients are between -1 and 1 with the following interpretation (Struyf et al., 1997):

- $s(i) \approx -1$, the classified CLP is closer to the second-best medoid than the allocated medoid resulting in unsuccessful classification.
- $s(i) \approx 0$, the classified CLP lies between two medoids.
- $s(i) \approx 1$, the CLP is well classified.

Guidelines to decide whether a silhouette value reflects substantial clustering structures are not available, but $s(i) > 0.5$ is regarded as an above average

measure reflecting the efficient identification of clustering structures. Struyf et al. (1997) advise that a silhouette value below 0.25 is an indication that no notable clusters are present in a data set. Carrying forward, a $s(i) \geq 0.5$ will be regarded as an indication of well separated clusters, which illustrates dependency between missing response CLs. A $s(i) < 0.5$ will be indicative of no substantial clustering structures, and, therefore, independence of missing response CLs. The pam method is available in the *cluster* package (Maechler et al., 2018) in R.

4 Simulation Results and Discussion

Only the single active handling approach to missing values is applied in the simulation study due to the high computational power required to cluster the missing subsets using multiple active handling. However, a comparison between the single- and active handling approaches of two cases is illustrated in Figure 5, which will appear later in this section. The average silhouette widths obtained from the CLPs of the missing sMCA solutions using two and three medoids over the fifteen simulated combinations are shown in Figures 2 and 3.

The plotting characters and colours distinguish between the number of variables in the simulated data sets. Five groupings within each plot appear to follow similar trends, each of these groupings consist of a particular sample size. From left to right the sample sizes increase from 100 to 3000 (Section 2.1) for each of the display windows. The sample size does not have an effect on identifying clusters in the missing CLPs. The number of variables do however impact the successful identification of clustering structures, it appears that for a lower number of variables ($p = 5$), the silhouette values are dispersed for both MAR and MCAR MDMs. As the number of variables increases ($p = 10$ or $p = 15$) the silhouette values stabilise with a majority of values above 0.5 for the MAR simulations and a majority below 0.5 for the MCAR simulations. A frequency distribution of the silhouette coefficients is presented in Table 3.

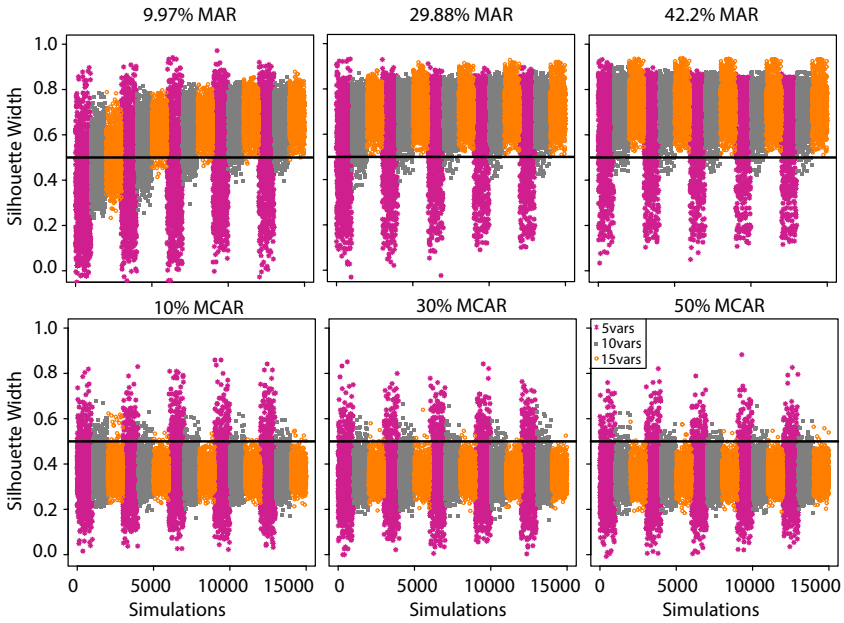

Figure 2: Average silhouette widths obtained from pam with 2 clusters over 15000 simulations.

From Table 3 it can be observed that a majority of silhouette coefficients occur in the [0.5 ; 0.75) interval for MAR simulations, whereas a majority of silhouette coefficients occur in a range with lower values between [0.25 ; 0.5) for the MCAR simulations. Figure 3 shows the silhouette coefficients obtained when specifying three medoids. First considering the MAR silhouette values, it can be seen that a lower number of variables ($p = 5$) result in silhouette values below 0.5. The performance of the cluster allocations seems to stabilise across sample sizes as the percentage of missingness increases. The MCAR simulations result in consistent trends irrespective of sample size or percentage of missing values. When comparing the silhouette values obtained for MCAR in Figure 2 with those of Figure 3, it is clear that more silhouette values occur above the 0.5 line when the number of medoids is increased. This is an indication that with an increase in the number of clusters, MCAR resulted in more substantial clusters than before.

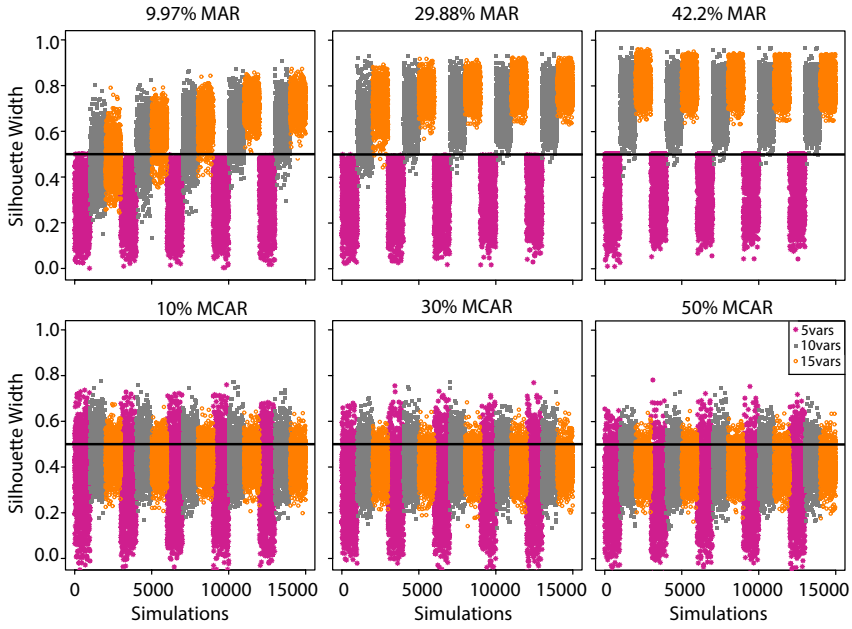

Figure 3: Average silhouette widths obtained from pam with 3 clusters over 15000 simulations.

Again, the MAR simulations result in larger proportions of higher silhouette coefficients and resulted in even better silhouette values when approximately 42.2 % of values were missing. As observed from Tables 3 and 4, there is a slight improvement in the proportion of silhouette values that range between [0.5 ; 0.75) when the number of clusters is increased from two to three for the MCAR MDM, while a decrease in the proportion of values in the range between [0.5 ; 0.75) are observed for MAR cases. This leads to the conclusion that a lower number of clusters resulting in higher silhouette coefficients could be indicative of dependence with well separated clusters agreeing with the description of MAR. Whereas a higher number of clusters with high silhouette coefficients could be an indication of independence, since more separation and unique groupings occur.

Table 3: Frequency distribution of silhouette coefficients of two clusters with highest frequencies indicated in **bold**.

	9.97 %MAR		**29.88 %MAR**		**42.2 %MAR**	
[-0.25 ; 0)	22	0.15 %	3	0.02 %	0	0 %
[0 ; 0.25)	901	6.01 %	339	2.26 %	245	1.63 %
[0.25 ; 0.50)	3383	22.55 %	1391	9.27 %	944	6.29 %
→ **[0.50 ; 0.75)**	**9335**	**62.23 %**	**9028**	**60.19 %**	**8294**	**55.29 %**
[0.75 ; 1.00)	1359	9.06 %	4239	28.26 %	5517	36.78 %
Total	15000	100.00 %	15000	100.00 %	15000	100.00 %
	10 %MCAR		**30 %MCAR**		**50 %MCAR**	
[-0.25 ; 0)	0	0 %	0	0 %	1	0.01 %
[0 ; 0.25)	1118	7.45 %	1299	8.66 %	1308	8.72 %
→ **[0.25 ; 0.50)**	**13070**	**87.13 %**	**13022**	**86.81 %**	**13135**	**87.57 %**
[0.50 ; 0.75)	783	5.22 %	668	4.45 %	546	3.64 %
[0.75 ; 1.00)	29	0.19 %	11	0.07 %	10	0.07 %
Total	15000	100.00 %	15000	100.00 %	15000	100.00 %

Table 4: Frequency distribution of silhouette coefficients of three clusters with highest frequencies indicated in **bold**.

	9.97 %MAR		**29.88 %MAR**		**42.2 %MAR**	
[-0.25 ; 0)	0	0 %	0	0 %	0	0 %
[0 ; 0.25)	1867	12.45 %	1435	9.57 %	1259	8.39 %
[0.25 ; 0.50)	5333	35.55 %	3694	24.63 %	3765	25.10 %
→ **[0.50 ; 0.75)**	**6974**	**46.49 %**	**5522**	**36.81 %**	4163	27.75 %
→ **[0.75 ; 1.00)**	826	5.51 %	4349	28.99 %	**55813**	**38.75 %**
Total	15000	100.00 %	15000	100.00 %	15000	100.00 %
	10 %MCAR		**30 %MCAR**		**50 %MCAR**	
[-0.25 ; 0)	31	0.21 %	37	0.25 %	33	0.22 %
[0 ; 0.25)	1519	10.13 %	1791	11.94 %	1840	12.27 %
→ **[0.25 ; 0.50)**	**10389**	**69.26 %**	**11071**	**73.81 %**	**11298**	**75.32 %**
[0.50 ; 0.75)	3055	20.37 %	2097	13.98 %	1827	12.18 %
[0.75 ; 1.00)	6	0.04 %	4	0.03 %	1	0.01 %
Total	15000	100 %	15000	100 %	15000	100 %

The visualisations and frequency distributions showed that there is a clear difference between the success of identifying clusters in MAR and MCAR MDMs. In order to distinguish between the two MDMs based on silhouette coefficients, a cut-off value of 0.6 has been identified from the tendencies observed in Figure 2 and 3. A majority of silhouette coefficients were observed above 0.6 in the MAR scenarios, which could be indicative of an MAR MDM. The percentages of occurrences above 0.6 are given in Table 5.

Table 5: Percentage of silhouette coefficients above 0.6.

	Two Clusters		**Three Clusters**	
$s(i) > 0.6$	MAR	MCAR	MAR	MCAR
$\approx 10\,\%$ missing	49.61 %	1.52 %	35.25 %	4.06 %
$\approx 30\,\%$ missing	72.43 %	1.22 %	58.82 %	2.15 %
$\approx 50\,\%$ missing	76.89 %	0.81 %	62.76 %	1.44 %

The sMCA biplots of missing CLPs and samples are presented for both the MAR MDM (Figure 4, left panel) and MCAR MDM (Figure 4, right panel) of a selected simulation with 30 % missing values.

The CLPs are illustrated with a triangle plotting character and open circle plotting characters illustrate the sample points. Three groupings appear between the CLPs in the left panel of Figure 4, which resulted in an average silhouette width ($s(i)$) of $s(i) = 0.7132$ and $s(i) = 0.6594$ when two clusters were specified. Two less distinctive groupings appear in the right panel of Figure 4, which resulted in $s(i) = 0.4149$ with three clusters and $s(i) = 0.4599$ when specifying two clusters. Since, both silhouette widths are less than 0.5 when attempting to cluster the biplot from the MCAR MDM, it confirms that the clustering structure is not sufficient. The contrary is confirmed by the silhouette widths obtained from the MAR MDM, which resulted in measures both greater than 0.6. The visual differences are subtle and therefore a measure of fit used in combination with the visualisation is useful to confirm the anticipated MDM.

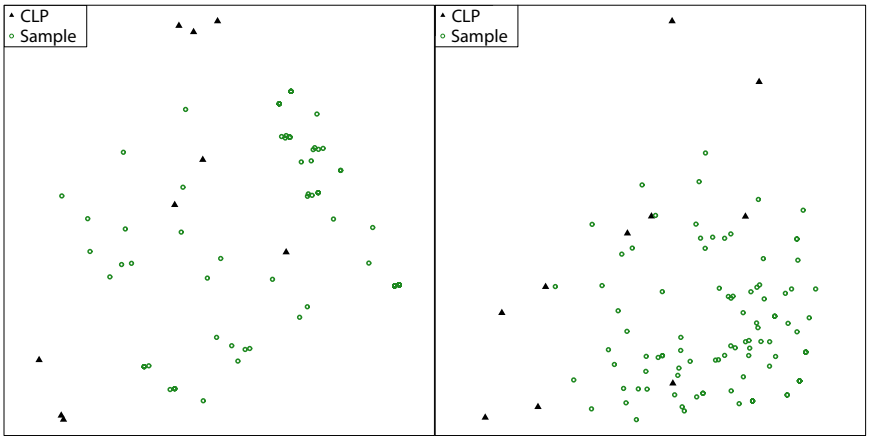

Figure 4: sMCA biplots (single active) with 30 % missing values. Left panel: MAR MDM. Right panel: MCAR MDM.

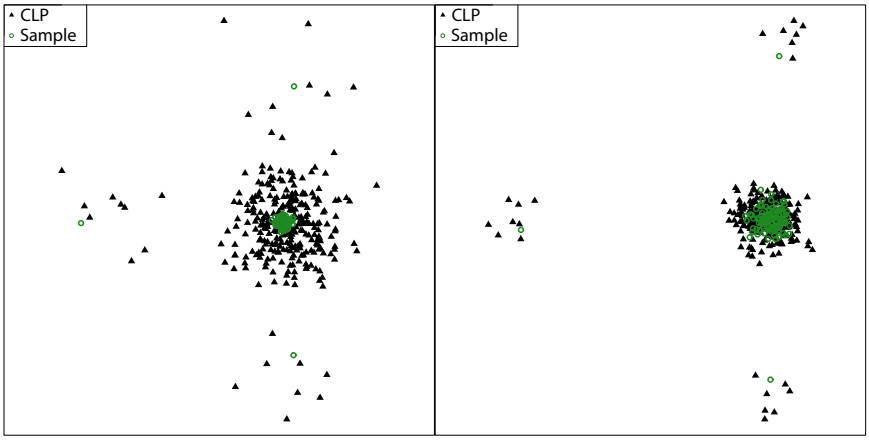

Figure 5: sMCA biplots (multiple active) with 30 % missing values. Left panel: MAR MDM. Right panel: MCAR MDM.

The difference between the MDMs are less evident when using multiple active handling of the missing CLs. There are however four discernible groupings in the MAR sMCA biplot (Figure 5, left panel) with less distinctive groupings appearing in the MCAR sMCA biplot (Figure 5, right panel). This confirms the hypothesis that MAR CLPs result in more clustering structures than MCAR MDMs.

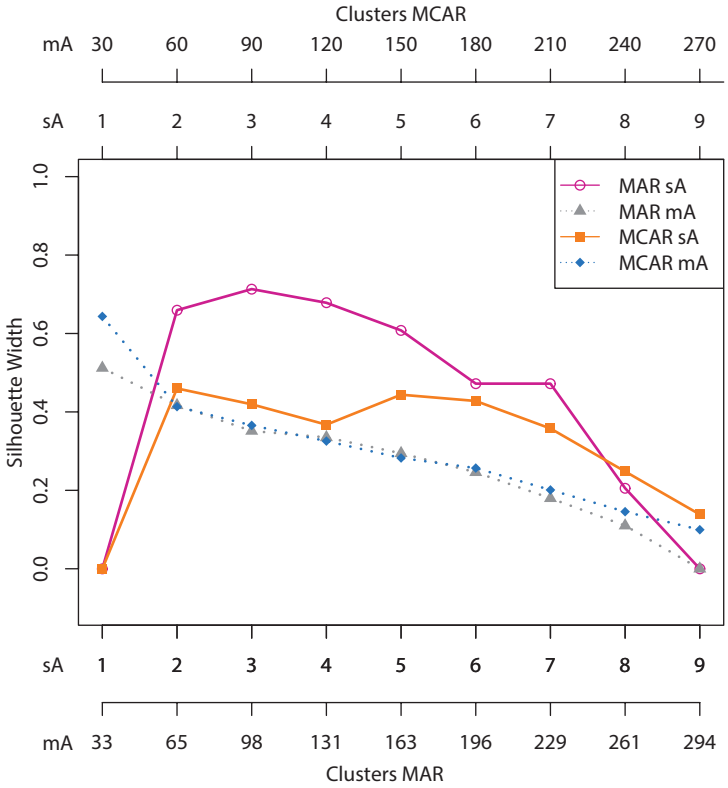

Figure 6: Comparison of silhouette widths for multiple active (mA) and single active (sA) sMCA biplots for MAR and MCAR MDMs.

It is not a fair comparison to use the same number of clusters for single active and multiple active sMCA biplots, since there is a substantial difference between the number of CLPs in the respective biplots. The number of clusters specified for multiple active sMCA biplots are chosen proportionally to the number of clusters in the single active sMCA biplots and taking the total number of CLPs available for clustering into consideration. As shown in Figure 6 using single active handling results in different clustering structures for MAR and MCAR MDMs, with the MCAR MDM achieving lower silhouette widths than the MAR MDM except when the number of clusters approaches the total number of CLPs. Albeit, resulting in low silhouette widths as the number of clusters increase, this is another indication of the independence of the MCAR MDM. The argument

is that if the CLPs can be distinctively clustered in isolation from the other CLPs, it confirms that the association is not strong. There are slight differences between the MDMs using multiple active which do not provide conclusive interpretations of the MDMs.

5 Real Application

There are two and three groupings identified in the left panel of Figure 7 which resulted in $s(i) = 0.7832$, for two clusters, and $s(i) = 0.5016$, for three clusters. The concentrated sample points show high association to a number of closely positioned CLPs, which is in accordance with the hypothesis of a MAR MDM, as stated earlier. The clusters specified for multiple active are again chosen proportionally which all resulted in lower silhouette widths as the number of clusters exceeded two groupings for single active as presented in Figure 8.

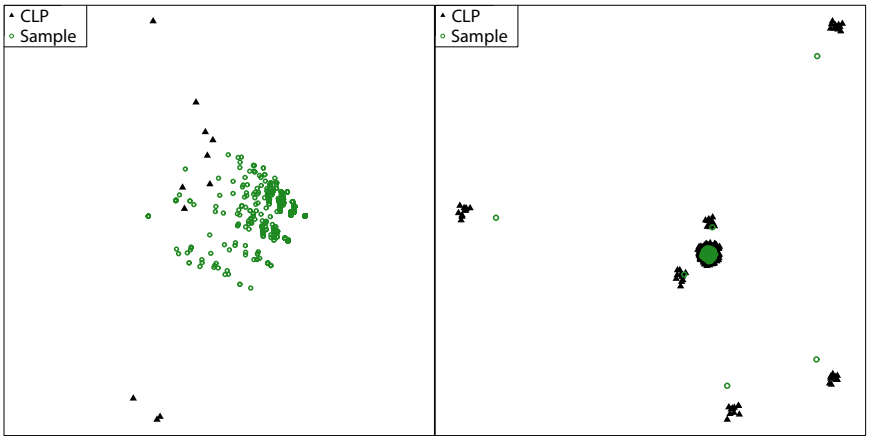

Figure 7: Real application: sMCA biplot of missing CLPs. Left panel: single active. Right panel: multiple active.

The real application results are consistent with the findings in the simulation study (Section 4). The decrease in the silhouette coefficient with an increase in the number of clusters, agrees with the MAR MDM structure. Also, the

silhouette coefficient above 0.6 confirms that the missing values might be due to the MAR MDM.

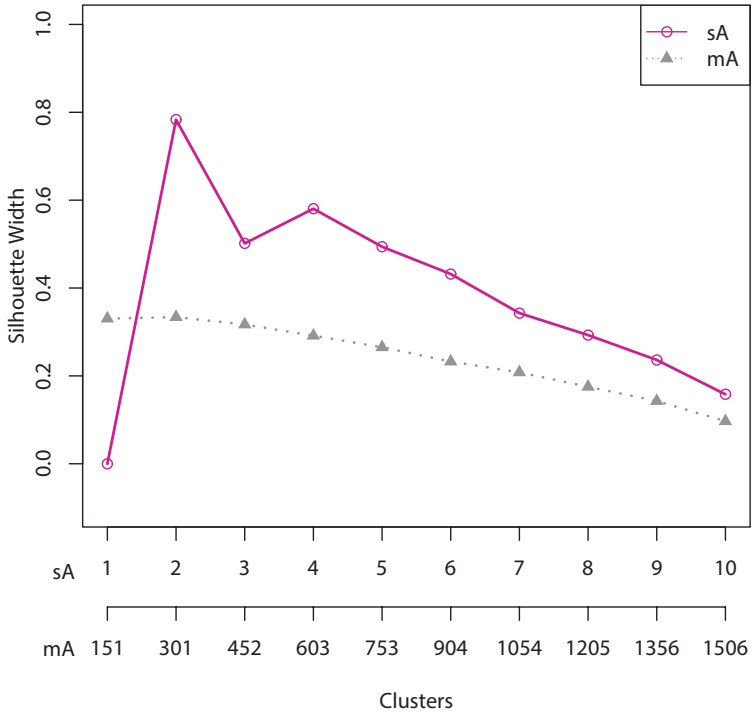

Figure 8: Comparison of silhouette widths for multiple- and single active sMCA biplots.

6 Concluding Remarks

It has been confirmed that there is a structural difference between the missing CLPs of the simulated MAR and MCAR MDMs. The MAR simulations resulted in satisfactory clusters with higher average silhouette coefficients than MCAR. The results suggest that silhouette coefficients above 0.6 could be an indication of an underlying MAR MDM. The effectiveness of multiple active handling of missing values to determine the MDM has to be further investigated. The recommendation for users is to use single active handling for sMCA biplots

along with a clustering technique that can provide measures of fit, in order to confirm the MDM with more certainty.

References

García-Laencina PJ, Sancho-Gómez J, Figueuras-Vidal AR (2009) Pattern Classification With Missing Data: A Review. Neural Computing and Applications 19(2):263–282. DOI: 10.1007/s00521-009-0295-6.

Gower J, Lubbe S, Le Roux NJ (2011) Understanding Biplots. John Wiley & Sons, Hoboken (USA). ISBN: 978-0-470012-55-0.

Greenacre M (2010) Biplots in Practice. Fundación BBBV, Bilbao (Spain). ISBN: 978-8-492384-68-6.

Greenacre M (2017) Correspondence Analysis in Practice, 3rd edn. Chapman & Hall/CRC, New York (USA). DOI: 10.1201/9781315369983.

Greenacre M, Pardo R (2006) Multiple Correspondence Analysis of Subsets of Response Categories. In: Greenacre M, Pardo R (eds.), Multiple Correspondence Analysis and Related Methods, Chapman & Hall/CRC, New York (USA), pp. 197–217. DOI: 10.1201/9781420011319.

Kaufman L, Rousseeuw P (1987) Clustering By Means of Medoids. In: Statistical Data Analysis Based on the L1 Norm and Related Methods, pp. 405–416. North-Holland, Amsterdam (The Netherlands).

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2018) cluster: Cluster Analysis Basics and Extensions. URL: `https://cran.r-project.org/web/packages/cluster/index.html`. R package version 2.0.7-1.

Mitsuhiro M, Yadohisa H (2015) Reduced k-means Clustering With MCA in a Low-dimensional Space. Computational Statistics 30(2):463–475, Kluwer Academic Publishers. DOI: 10.1007/s00180-014-0544-8.

Nenadić O, Greenacre M (2007) Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca package. Journal of Statistical Software 20(3):1–13. DOI: 10.18637/jss.v020.i03.

R Core Team (2018) R: A Language and Environment for Statistical Computing. Vienna, Austria. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Hoboken (USA). DOI: 10.1002/9780470316696.fmatter.

Rubin DB (2003) Disscussion on Multiple Imputation. International Statistical Review 71(3):619–625. DOI: 10.1111/j.1751-5823.2003.tb00216.x

Schafer JL, Olsen MK (1998) Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research 33(4):545–571. DOI: 10.1207/s15327906mbr3304_5.

Struyf A, Hubert M, Rousseeuw PJ (1997) Integrating robust clustering techniques in S-PLUS. Computational Statistics & Data Analysis 26(1):17–37. DOI: 10.1016/S0167-9473(97)00020-0.

Van Buuren S (2012) Flexible Imputation of Missing Data, 1st edn. Chapman & Hall/CRC, New York (USA). ISBN: 978-0-429065-40-8, DOI: 10.1201/b11826.

Van de Geer JP (1993) Multivariate Analysis of Categorical Data: Theory. Sage Publications, Newbury Park (USA). ISBN: 978-0-803945-65-4.