Supplementary material for "Archetypal analysis with missing data: See all by looking at a few based on extreme profiles"

Abstract

A toy example illustrates our proposed methodology and the flaws of the previous approaches.

Keywords: Incomplete data set, Archetype Analysis, Multidimensional Scaling, Partial Distance Strategy

1 Toy example

Let $\mathbf{X}$ be the matrix composed of the following four two-dimensional data: (1 2), (7 NA), (6 6), (4 7). Note that the instance (7 NA) is certainly outside the convex hull of the rest of the data, as its first coordinate, 7 is higher than the first coordinate of the other data. AA is computed for the different approaches and Table 1 shows the archetypes and RSS/n for k 1 to 4. For cMDS only 2 of the first 3 (n - 1) eigenvalues are positive, i.e. $\mathbf{D}$ is not Euclidean, and therefore the original dissimilarities cannot be completely recovered in a Euclidean space.

The theoretical archetype for $k = 1$ is the mean, (4.5 5), which is returned by all methods except MAAEIS and AAEDcMDS. The theoretical βs should be a 4-dimensional vector with values equal to 0.25 in each position. Nevertheless, those βs are only returned by AAK, AAEDHP and AAPDSHP (see Table 2). For $k = 2$, the lowest RSS is achieved by AAPDSHP. The convex hull of the configuration points with cMDS has only three vertices (the data point (6 6) is projected inside the convex hull of the other points). This explains the solution found for $k = 3$ with AAEDcMDS, and the impossibility of computing a solution for $k = 4$. Although the RSS is zero for AAEDcMDS with $k = 3$, it is not comparable with the other methods due to one of the archetypes having missing values, which influences the PDS estimates. For the rest of the methods, with no missing values in the archetypes, the method with the lowest RSS is AAII, followed by MAAEIS and AAPDSHP. Note also that the second coordinates of some archetypes returned by AAEIS are larger than the highest datum (and also lower than the smallest datum), which is theoretically impossible if archetypes are a mixture of data. For $k = 4$, as the data set compromises four points that are vertices (i.e. each of them is not in the convex hull of the other points), the archetypes should theoretically coincide with the data set. This only happens with AAEDHP and AAPDSHP. For some procedures, no solution is returned. The AAII solution cannot be computed with $k = 4$ because we need 4 initial complete archetypes to start with. If we impute, for example, the minimum value of the variable to the missing entry to build the starting archetypes, while for the rest of the algorithm this imputation is discarded, then AAII returns the four points, which is the right solution. Note that only the procedures that employ equation 2 of the main manuscript to define

the archetypes can return archetypes with NAs, i.e. AAEDcMDS, AAEDHP, AAPDSHP, MAAMOHAN, MAAEIS, AAII and AAK.

An overview of the solutions for the different ks suggests that AAPDSHP is the best method for this data set, together with AAII.

1.1 Expanded toy example

We will see that AAMOHAN does not fulfill the theory and can return archetypes outside the convex hull of the data. Let $\mathbf{X}$ be the same matrix as the previous Section, but now the point (7 NA) is substituted by (7 1). These points constitute 4 archetypes from which we generate 50 sample points as $\mathbf{x}_i = \mathbf{X}\mathbf{h}_i$, where $\mathbf{h}_i$ is a random vector sampled from a Dirichlet distribution with $\alpha = (0.2,\ 0.2,\ 0.2,\ 0.2)$ and $i = 1,\ ...,\ 50$. Then, if the first coordinate is higher than 6, the second coordinate is removed. If AAMOHAN is applied to this data set with $k = 4$, one of the archetypes returned is (6.54 0.14), which is clearly outside the convex hull of the data, since the value of the second coordinate of the data set cannot be below one, taking into account the model that generated the data.

Table 1: Archetypes and RSS for the toy example.

No. archetypes	$k = 1$			$k = 2$			$k = 3$			$k = 4$		
Methods	Arch.		RSS/n	Arch.		RSS/n	Arch.		RSS/n	Arch.		RSS/n
AAMOHAN	4.5	5	13.4	6.39	6.97	1.1	4	6.99	9e-3	1	2	1e-2
ϵ = 1e-3				1.05	2.08		1	2		4.73	2.04	
							6.91	5.93		4.02	6.99	
										6.89	5.94	
AAMOHAN	4.5	5	13.4	6.43	7	1.0	4	7	1e-2	4.05	6.99	3e-3
ϵ = 1e-9				1.05	2.08		1	2		6.95	6	
							6.90	6		1	2	
										6.63	6.04	
MAAMOHAN	4.51	5	13.4	6.39	7	1.1	6.91	6.	7e-3	6.89	6	1e-2
ϵ = 1e-3				1.05	2.09		4	7		4.02	7	
							1	2		1	2	
										4.74	2.04	
MAAMOHAN	4.5	5	13.4	6.43	7	1.0	6.9	6	1e-2	6.95	6	3e-3
ϵ = 1e-9				1.05	2.08		4	7		4.05	6.99	
							1	2		1	2	
										6.63	6.04	
AAEIS	4.5	5	13.4	1.33	3.41	2.5	4.01	9.36	0.28	4.04	9.41	0.2
				6.15	5.6		6.99	0.06		4.78	8.83	
							1	2.67		7	0	
										1	2.67	
MAAEIS	4.5	4.44	13.67	1.33	2.55	2.0	4	7	6e-5	4	7	0
				6.15	5.86		6.99	6		4.77	6.61	
							1	2		7	NA	
										1	2	
AAII	4.5	5	13.4	6.47	7	1.0	7	5.99	1e-9	Not	comp.	-
				1.02	2.03		4	7				
							1	2				
AAK	4.5	5	13.4	5.74	6.43	2.3	6.78	6	5e-2	Not	comp.	-
				1.09	2.14		4.07	6.96				
							1	2				
AAEDcMDS	4.36	4.54	14.2	6.33	7	1.1	1	2	0	Not	comp.	-
				1.06	2.09		7	NA				
							4	7				
AAEDHP	4.5	5	13.4	1.06	2.09	1.7	1	2	8e-3	7	NA	0
				6.06	6.51		4.01	6.98		6	6	
							6.91	6		1	2	
										4	7	
AAPDSHP	4.51	5	13.4	6.85	7	0.9	1	2	2e-4	4	7	0
				1	2		6.99	6		7	NA	
							4.02	6.99		1	2	
										6	6	

Table 2: β values for the toy example with $k = 1$.

Methods				
AAMOHAN (ϵ = 1e-3)	0.261	0.091	0.508	0.140
AAMOHAN (ϵ = 1e-9)	0.262	0.092	0.506	0.140
AAEIS	0.331	0.155	0.515	0
AAII	0.269	0	0.654	0.077
AAK	0.250	0.250	0.251	0.250
AAEDcMDS	0.290	0.410	0	0.300
AAEDHP	0.250	0.250	0.250	0.250
AAPDSHP	0.250	0.250	0.250	0.250