Supplementary material for “Archetypal shapes based on landmarks and extension to handle missing data”

Irene Epifanio and M. Victoria Ibáñez and Amelia Simó
Dept. Matemàtiques and IMAC
Universitat Jaume I
Castelló, 12071
Spain

Abstract

We carry out a comparison study of AA with missing data in the multivariate case and shape case. Our proposal provides very competitive performance in such comparisons.

1 Comparison study for AA with missing data in the multivariate case

In this section, two experiments have been considered to compare the performance of the proposed procedure for AA with missing data in the multivariate case. In section 1.1, we use a small, simple artificial, two-dimensional data set to illustrate the results and to compare our procedure with the previous attempt to compute multivariate AA with missing data, which was introduced by [6]. In section 1.2, AA results with different strategies, such as erasing the cases with missing data or imputation, are compared versus our proposal, with a well-known benchmark data set. The results show that our new procedure is the best alternative of those considered.

1.1 Toy example

A previous attempt to compute AA with missing data was proposed by [6]. In that paper, they considered a different objective function for minimizing RSS in AA and a different minimization algorithm. In their RSS for missing data, β values were also used as a dividend plus a certain quantity (ϵ), so the original RSS from AA was changed. As we will see, this means that their results do not fit with the expected theoretical results.

Let X be the matrix composed of the following four two-dimensional data: (1 2), (5 NaN), (6 6), (4 7). We apply AA for $p = 1$ and $p = 2$ for illustration purposes. Our methodology is referred to as AAM, while the method proposed by [6] is referred to as AAMOHAN.

On the one hand, the archetype for $p = 1$ should be the mean. In this example, the mean is (4 5). This is the result obtained by AAM, but for AAMOHAN the archetype that returns the lowest RSS defined by [6] after 20 repetitions is: (2.61 4.06), quite far away from the expected result.

The data are plotted in black in Fig. S1 together with the archetypes returned by AAM (red stars) and those returned by AAMOHAN (blue triangles) with p =2. The archetypes obtained by AAM are (1.05 2.63) and (5.47 6.23), while the best archetypes provided by AAMOHAN after 20 repetitions are (5.06 6.46) and (1.01 2.01). These archetypes are the same as those that would have been obtained by AA if the case with missing data had been erased, i.e. the same as if the information given by the data (5 NaN) had been discarded.

1.2 Simulation study with waveform data

When missing values are present, different strategies can be used to handle them in order to apply AA. Three strategies are considered for the comparison. The simplest is to remove the cases with missing values and work only with complete cases. This strategy is referred to as COM. Obviously, some (valuable) information is wasted with the COM strategy. Another strategy is to estimate the missing values and work with all the cases. There are many missing value estimators. These rank from the simplest one, such as using the mean values of the non-missing values in the corresponding variable, to more sophisticated ones that use the information from other variables. Multiple imputation using additive regression, bootstrapping and predictive mean matching have been chosen, which is implemented with the *aregImpute* func-

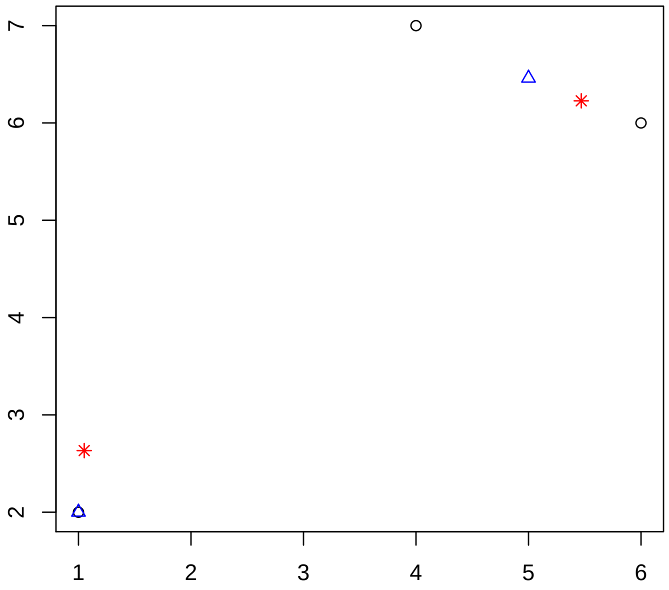


Figure S1: Toy example. See text for details. The point (5 NaN) is not plotted.

tion in the R package **Hmisc** ([4]). This is a well-known imputation method. A total of 5 (the default number in *aregImpute*) imputations are obtained. To combine the 5 imputed data sets, they are appended. For instance, if we have 21 variables with 150 observations, as the imputation frequency is 5, then the final data set will be composed of 21 variables with 750 observations. This strategy is referred to as IMP. The third strategy is to work with the missing values and apply the methodology proposed in the manuscript. As before, this strategy is referred to as AAM.

The outline of the experiment carried out to compare the AA results obtained with the three strategies is as follows. Each of the following steps is repeated 100 times.

1. A data set is generated. The data set is called ds_i ($i = 1, ..., 100$). A benchmark data set, the waveform data set defined by [2, pp. 49-55], has been considered. The reason why this set has been chosen is because three archetypes (bases) are the sources from which the data is generated. Data sets are generated using the function *mlbench.waveform* from the R package **mlbench** [5]. Each data set is formed by $n = 150$ samples and $r = 21$ continuous variables and a variable showing the 3 classes (33% for each of the 3 classes). Each class is generated from a convex combination of 2 of 3 "base" waves. The categorical variable is discarded. AA is applied to ds_i with $p = 3$ and 20 repetitions. The archetypes that give the lowest RSS, which is the best model or repetition, are kept. This strategy is referred to as ORG, as it uses the original data without missing values.

2. Some values are removed from ds_i. The data sets with missing values are called dsm_i ($i = 1, ..., 100$). In shapes, the missing landmarks could not be completely random but, for example, at the end of extremities or in fragile areas, i.e. in specific zones. Therefore, although this is the multivariate case, the removing procedure is not completely random. Only three variables (the 5th, 10th and 15th) are selected as candidates for removing their values. Then 50% of the values of these variables are randomly removed. Therefore, of the cases with missing values, it could happen that they would have one, two or three missing values. The percentage 50% was chosen because it is a difficult situation with a high percentage of missing values.

3. The different strategies (COM, IMP, AAM) are applied to each data set

dsm_i and archetypes are obtained for $p = 3$. Archetypes corresponding to the best of the 20 repetitions are kept.

4. The matrix α that approximates the original ds_i (without missing values) using the archetypes kept for each strategy is obtained. Then the RSS is computed with ds_i and those archetypes and α. The idea is to judge the capacity of each method to recover the original data.

A summary (mean and standard deviation) of the RSS from each strategy can be seen in Table S1. The gold standard reference is ORG, as it uses all the information, without missing values. We can see that our proposal, AAM, reports results that are very close to ORG, despite working with missing values. As expected, the worst is COM, as it discards information.

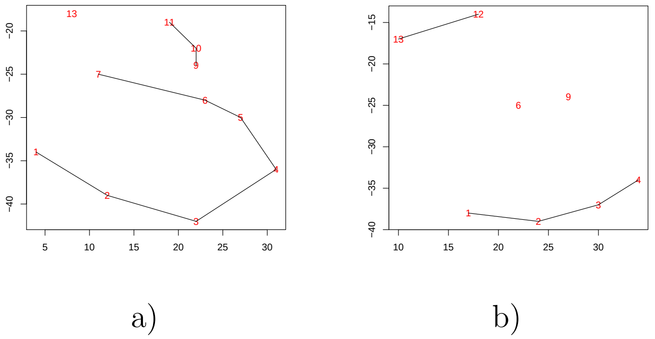
Table S1: Mean RSS (st. deviation in brackets) for each strategy for waveform data. Details are given in the text.

ORG	COM	IMP	AAM
19.40 (0.55)	24.02 (1.94)	21.12 (0.79)	19.70 (0.55)

2 Comparison study for AA with missing landmarks

We are going to consider the dataset *digit3.dat* again (without the first outlier digit) in the R package **shapes** ([3]). Although the data set includes the coordinates of all the landmarks, the R package **LOST** ([1]) can be used to simulate missing landmarks randomly. In particular, the function *missing.specimens* randomly selects a predetermined number of individuals and removes some of their landmarks. In our case, it has randomly chosen 40% of the individuals and has randomly removed $2, 3, 4$ or 5 landmarks from each of them. Fig S2 shows some of the individuals with missing landmarks.

As in Section 3 of the main manuscript, 4 archetypal shapes are computed. As previously, ORG-S denotes the results for AA that use the original data without missing values. COM-S uses only the complete cases for AA with shapes. Three strategies for landmark imputation are considered, which are implemented with the *MissingGeoMorph* function of the R package **LOST** ([1]) . The archetypal shapes obtained by previously using each of these



a) b)

Figure S2: a) Individual with two missing landmarks; b) individual with 5 missing landmarks.

imputation techniques are denoted by: IMP-mean, which uses mean substitution, IMP-BPCA, which uses Bayesian principal component analysis, and IMP-REG, whose values are estimated based on the most strongly correlated variable available. Our proposal is referred to as AAM-S. The archetypes returned for each strategy can be seen in Fig. S3. Then the matrix α that approximates the original dataset *digit3.dat* (without missing values) using the archetypes returned for each strategy is obtained. Afterward, RSS is computed with those archetypes and α, and the results are displayed in Table S2. As before, the gold standard reference is ORG-S, as it uses all the information, without missing values. The largest RSS is for COM-S, since it discards information. The RSS for IMP-BPCA, IMP-mean and AAM-S are very similar. The RSS for IMP-REG is the closest to ORG-S.

Table S2: RSS for each strategy for the *digit3.dat* dataset.

ORG-S	COM-S	AAM-S	IMP-BPCA	IMP-mean	IMP-REG
0.02155	0.02375	0.02261	0.02242	0.02264	0.02188

References

[1] J. Arbour and C. Brown. *LOST: Missing Morphometric Data Simulation and Estimation*, 2017. R package version 1.3.

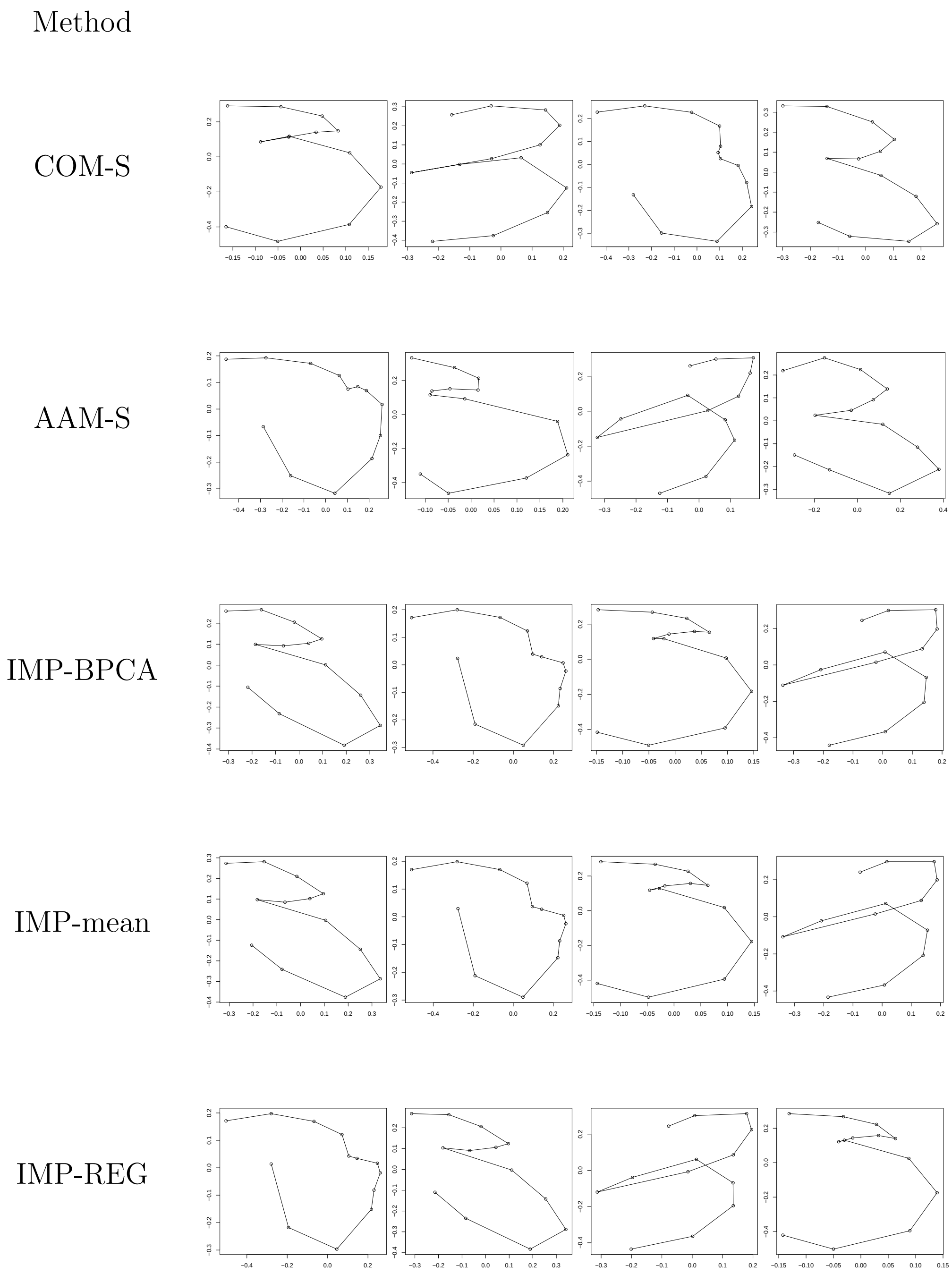


Figure S3: Archetypes obtained for each strategy for handling missing landmarks in digit 3s.

[2] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, Monterey, 1984.

[3] I. L. Dryden. *shapes: Statistical Shape Analysis*, 2015. R package version 1.1-11.

[4] F. E. Harrell Jr, with contributions from Charles Dupont, and many others. *Hmisc: Harrell Miscellaneous*, 2016. R package version 4.0-2.

[5] F. Leisch and E. Dimitriadou. *mlbench: Machine Learning Benchmark Problems*, 2010. R package version 2.1-1.

[6] M. Mørup and L. K. Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012.