LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Manuel J. A. Eugster & Friedrich Leisch

Weighted and Robust Archetypal Analysis

Technical Report Number 082, 2010
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Weighted and Robust Archetypal Analysis

Manuel J. A. Eugster and Friedrich Leisch
Department of Statistics, LMU München, München, Germany
Firstname.Lastname@stat.uni-muenchen.de

Abstract

Archetypal analysis represents observations in a multivariate data set as convex combinations of a few extremal points lying on the boundary of the convex hull. Data points which vary from the majority have great influence on the solution; in fact one outlier can break down the archetype solution. This paper adapts the original algorithm to be a robust M-estimator and presents an iteratively reweighted least squares fitting algorithm. As required first step, the weighted archetypal problem is formulated and solved. The algorithm is demonstrated using both an artificial and a real world example.

1 Introduction

Archetypal analysis has the aim to represent observations in a multivariate data set as convex combinations of a few, not necessarily observed, extremal points (archetypes). The archetypes themselves are restricted to being convex combinations of the individuals in the data set and lie on the data set boundary, i.e., the convex hull. This statistical method was first introduced by Cutler and Breiman (1994) and has found applications in different areas, e.g., in economics (Li et al., 2003; Porzio et al., 2008), astrophysics (Chan et al., 2003) and pattern recognition (Bauckhage and Thurau, 2009).

Archetypal analysis approximates the convex hull of the data set – this suggests itself that data points which "behave differently from the large majority of the other points" (Morgenthaler, 2007) have a great influence on the solution. In fact, the farther a data point is from the center of the data set the more influence it has on the solution. Although archetypal analysis is about the data set boundary, practice has shown that in many cases one primarily is interested in the archetypes of the large majority than of the totality. For example, Li et al. (2003) look at extreme consumers in segmenting markets

– it is obvious that the extreme consumers should not be total outliers but related to the majority of the consumers. The present paper adapts the original archetypes estimator to be a robust M-estimator (Huber and Ronchetti, 2009) and presents an iteratively reweighted least squares (IRLS) fitting algorithm. This enables a robust analysis in terms of Rousseeuw and Leroy (2003, defined for robust regression): "A robust analysis first wants to fit *an archetypal analysis* to the majority of the data and then to discover the outliers as those points which possess large residuals from that robust solution."

Robust archetypal analysis formulated in this way is based on weighting the residuals and observations respectively. On this account, the paper formulates and solves the weighted archetypal problem in a first step. Weighted archetypal analysis enables to represent additional information available from the data set, like the importance of observations or the correlation between observations.

The paper is organized as follows. In Section 2, the original archetypal analysis is briefly introduced and its breakdown point discussed. In Section 3 the weighted archetypal problem is solved. Based on that, Section 4 introduces the robust M-estimator, the corresponding iteratively reweighted least squares problem and the fitting algorithm. Each step is illustrated using an artificial toy example. In Section 5 the robust algorithm is applied on the Air-Pollution data set (slightly modified to contain outliers) which is already used in the original archetypal analysis paper by Cutler and Breiman (1994). Finally, in Section 6 the conclusions are given.

2 Archetypal analysis

Consider an $n \times m$ matrix X representing a multivariate data set with n observations and m attributes. For given k the archetypal problem is to find the matrix Z of k m-dimensional archetypes. More precisely, to find the two $n \times k$ coefficient matrices α and β which minimize the residual sum of squares

$$\text{RSS} = \|X - \alpha Z^\top\|_2 \text{ with } Z = X^\top \beta \tag{1}$$

subject to the constraints

$$\sum_{j=1}^{k} \alpha_{ij} = 1 \text{ with } \alpha_{ij} \geq 0 \text{ and } i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \beta_{ji} = 1 \text{ with } \beta_{ji} \geq 0 \text{ and } j = 1, \ldots, k.$$

The constraints imply that (1) the approximated data are convex combinations of the archetypes, i.e., $X = \alpha Z^\top$, and (2) the archetypes are convex combinations of the data points, i.e., $Z = X^\top \beta$. $\| \cdot \|_2$ denotes the Euclidean matrix norm.

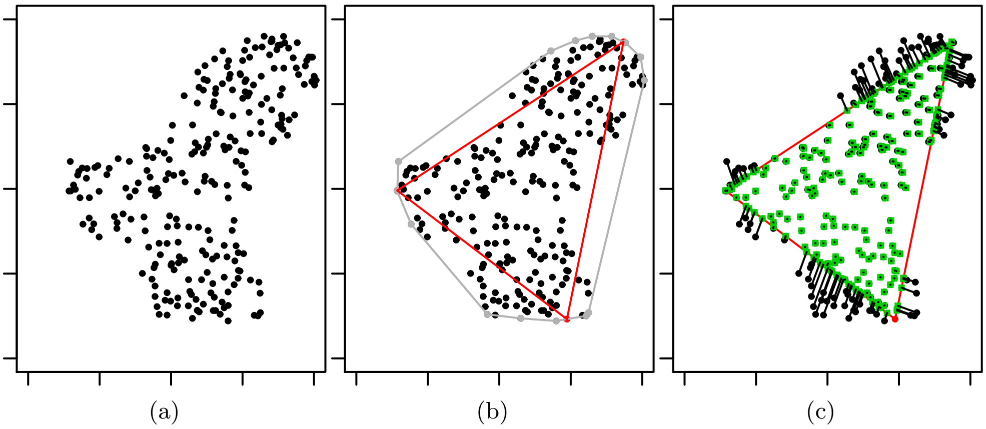


Figure 1: (a) Artificial toy data set. (b) Approximation of the convex hull (outer polygon) by three archetypes (inner triangle). (c) Approximation of the data through the three archetypes and the corresponding α values.

Cutler and Breiman (1994) present an alternating constrained least squares algorithm to solve the problem: it alternates between finding the best α for given archetypes Z and finding the best archetypes Z for given α; at each step several convex least squares problems are solved, the overall RSS is reduced successively. Section 4 provides details.

Figure 1(a) shows an artificial two-dimensional toy data set. The advantage of such a simple problem is that we can visualize the result, Section 5 shows a more realistic example. The toy data set consists of two attributes x and y, and 250 observations. It is generated in a way such that $k = 3$ archetypes are the optimal solution. Figure 1(b) shows the archetypes, their approximation of the convex hull (the inner triangle) and the convex hull of the data (outer polygon). Figure 1(c) shows the approximation of the data through the archetypes and the corresponding α values; as we can see, all data points outside the approximated convex hull are mapped on its boundary, all data points inside are mapped exactly.

The breakdown point of archetypal analysis The breakdown point is the smallest amount of contamination that may cause an estimator to take arbitrary large values. We follow the sample version defined by Donoho and Huber (1983): Given the data set X with n observations, and T, an estimator based on X, we let $\epsilon_n^*(T, X)$ denote the smallest fraction of contaminated observations needed to break down the estimator T.

For a given k archetypal analysis reaches the worst possible value, $\epsilon_n^*(T, X) = 1/n$ for every X; which converges to 0 as $n \to \infty$. To check this fact, suppose that one data point of the toy data set moves away; Figure 2 illustrates this scenario. Note how one of the archetypes has gone on to "catch" this outlier observation when the cross data

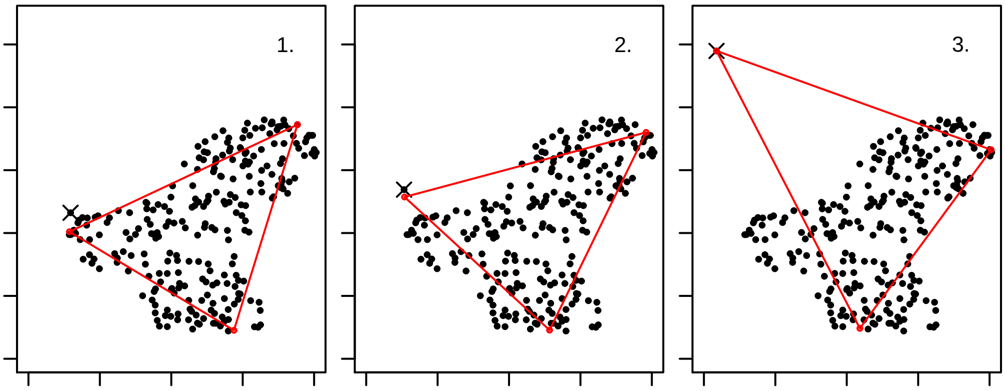


Figure 2: Behavior of the archetypes (triangle) when one data point (cross) moves away.

point moves away from the majority of the data. In terms of the minimization problem this means that at one point (related to the distance of the outlier) the RSS is more reduced if the outlier is approximated well, then the remaining data points. Now, take the outlier to infinity to break down the archetype solution with one single outlier.

3 Weighted archetypes

In the original archetypal problem, equation (1), each observation and therefore each residual contributes to the solution with equal weight. Remember that X is an $n \times m$ matrix and let W be a corresponding $n \times n$ square matrix of weights. The weighted archetypal problem is then the minimization of

$$\mathrm{RSS} = \|W(X - \alpha Z^\top)\|_2 \text{ with } Z = X^\top \beta. \tag{2}$$

Weighting the residuals is equivalent to weighting the data set:

$$\begin{aligned} W(X - \alpha Z^\top) &= W(X - \alpha(X^\top\beta)^\top) = W(X - \alpha\beta^\top X) = \\ &= WX - \alpha\beta^\top WX = WX - \alpha(X^\top W^\top \beta)^\top = \\ &= WX - \alpha((WX)^\top\beta)^\top = \tilde{X} - \alpha\tilde{Z}^\top \end{aligned}$$

Therefore the problem can be reformulated as minimizing

$$\mathrm{RSS} = \|\tilde{X} - \alpha\tilde{Z}^\top\|_2 \text{ with } \tilde{Z} = \tilde{X}^\top\beta \text{ and } \tilde{X} = WX. \tag{3}$$

This reformulation allows the usage of the original algorithm with the additional prepprocessing step to calculate $\tilde{X}$ and the additional post-processing step to recalculate α for the data set X given the archetypes $\tilde{Z}$.

The weight matrix W can express different aspects. In case of W a diagonal matrix the weights represents some kind of importance of the observations. The weight values are rescaled to the range $[0, 1]$ – values greater than one disperse the data points, and therefore the data set boundary, which is not meaningful in case of archetypal analysis. Furthermore, W can be an arbitrary square matrix, for example, a matrix to decorrelate the observations.

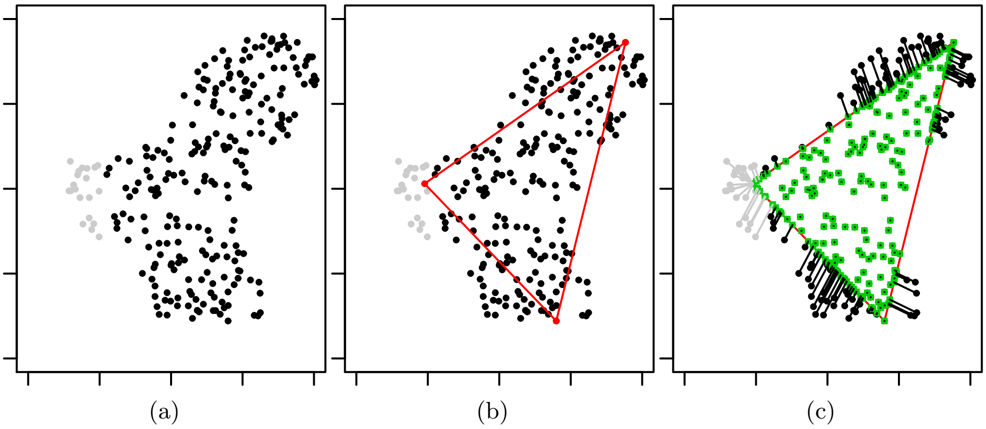


Figure 3: Weighted archetypal analysis where gray data points weight 0.8 and black data points 1.

Figure 3 illustrates the weighted archetypal analysis of the toy data set for $k = 3$. (a) Gray data points weight 0.8 and black data points 1). (b) As expected, on the side of the lower weighted data points the corresponding archetype is inside the data set boundary). (c) These data points are mapped on the approximated convex hull boundary, their residuals contribute to the overall RSS weighted.

4 Robust archetypes

A popular robust technique is using M-estimators instead of least squares estimators. Let $R = (X - \alpha Z^\top)$ be the matrix of residuals. The standard archetypal analysis tries to minimize the Euclidean (matrix) norm of the residuals, i.e., $\min \|R\|_2$. Here, large residuals have large effects, which privileges outliers. M-estimators try to reduce the effect of outliers by replacing the squared residuals by another function $\rho(\cdot)$ less increasing than square; this yields to the optimization problem $\min \rho(R)$. Such a problem can be reformulated as an iterated reweighted least squares one, i.e., in the tth iteration $\min \|w(R^{(t-1)})R\|_2$ is solved with $w(\cdot)$ a weight function depending on the residuals of the $(t-1)$th iteration. (For general details on transforming the object function into the influence and weight functions we refer to, for example, Huber and Ronchetti, 2009)

There is a large set of suitable objective functions $\rho(\cdot)$ and corresponding weight functions $w(\cdot)$ available – used, for example, in robust regression (Rousseeuw and Leroy, 2003) and locally weighted regression and scatterplot smoothing (Cleveland, 1979). Note that here the residual R_i of observation i ($i = 1, \dots, n$) is of dimension m, therefore the one-dimensional distance calculations in the original functions are replaced by the corresponding norm functions. For an example, the *Bisquare* objective $\rho(\cdot)$ and weight $w(\cdot)$ functions are defined as $\rho(R) = \sum_{i=1}^{n} \tilde{\rho}(R_i)$ and $w(R) = \mathrm{diag}(\tilde{w}(R_i))$, $i = 1, \dots, n$ with R_i the m dimensional residual of the ith observation and

$$\tilde{\rho}(R_i) = \begin{cases} \frac{k^2}{6}(1 - (1 - \|\frac{R_i}{k}\|_2)^3), & \text{for } \|R_i\|_1 < k \\ \frac{k^2}{6}, & \text{for } \|R_i\|_1 \geq k \end{cases},$$

$$\tilde{w}(R_i) = \begin{cases} (1 - \|\frac{R_i}{k}\|_2)^2, & \text{for } \|R_i\|_1 < k \\ 0, & \text{for } \|R_i\|_1 \geq k \end{cases}.$$

The value k is a tuning parameter; practical application showed that $k = 6s$ with s the median of the residuals unequal to zero works well (following Cleveland, 1979). Reducing the Bisquare weight function to binary weights, i.e., Bisquare weights greater than a threshold are 1, otherwise 0, proved to be useful as well.

The iterative reweighted least squares algorithm at step t involves solving the weighted archetypal minimization problem

$$\begin{aligned} R^{(t)} &= \underset{R}{\mathrm{argmin}} \|w(R^{(t-1)})\ R\|_2 \\ &\quad \text{with } R = (X - \alpha Z^\top) \text{ and } Z = X^\top \beta, \end{aligned} \tag{4}$$

or according to equation (3),

$$\begin{aligned} R^{(t)} &= \underset{R}{\mathrm{argmin}} \|R\|_2 \\ &\quad \text{with } R = (X^t - \alpha Z^{t\top}) \\ &\quad \text{and } Z^t = X^{t\top}\beta,\ X^t = w(R^{(t-1)})X. \end{aligned} \tag{5}$$

The original algorithm proposed by Cutler and Breiman (1994) is an iterative alternating constrained least squares algorithm: it alternates between finding the best α for given archetypes Z and finding the best archetypes Z for given α. The algorithm has to deal with several numerical issues, e.g., each step requires the solution of several convex least squares problems. Eugster and Leisch (2009) describe the algorithm in detail, provide different numerical solutions for individual steps and investigate its stability and computational complexity. Here, we focus on the additional steps needed to enable weighted and robust archetypal analysis (marked with * in the following listing). Given the number of archetypes k and a weight matrix W (weighted archetypes) and a weight function $w(\cdot)$ (robust archetypes) the algorithm consists of the following steps:

*1. Data preparation: standardize and weight data, $X^0 = WX$.

2. Initialization: define α and β in a way that the constraints are fulfilled to calculate the starting archetypes Z.

3. Loop until RSS reduction is sufficiently small or the number of maximum iterations is reached:

   *3.1. Reweight data: $X^t = w(R^{(t-1)})X$.

   3.... Calculate Z, i.e., α and β given the data X^t.

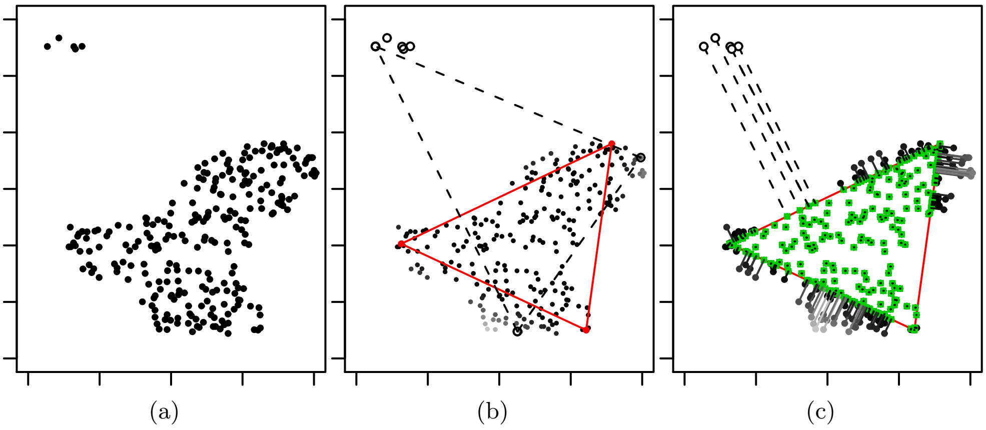
   3.6. Calculate residuals R^t and residual sum of squares RSS.

*4. Recalculate α and β for the given set of archetypes Z and X.

5. Post-processing: rescale archetypes.

Convergence Cutler and Breiman (1994) show that the algorithm converges in all cases, but not necessarily to a global minimum. Hence, the algorithm should be started several times with different initial archetypes.

Standardization Step 1 standardizes the data set to mean 0 and standard deviation 1. The mean is not robust and if outliers are in the data set available, a normalization toward the median is more suitable. Scale and median normalization (e.g., Quackenbush, 2002) is one simple approach we use: Transform the m attributes such that their distributions or their medians are equivalent.


(a)

(b)

(c)

Figure 4: Robust archetypal analysis; the gray scale of the data points indicate their final weights. Note that the outliers have weight 0 (unfilled).

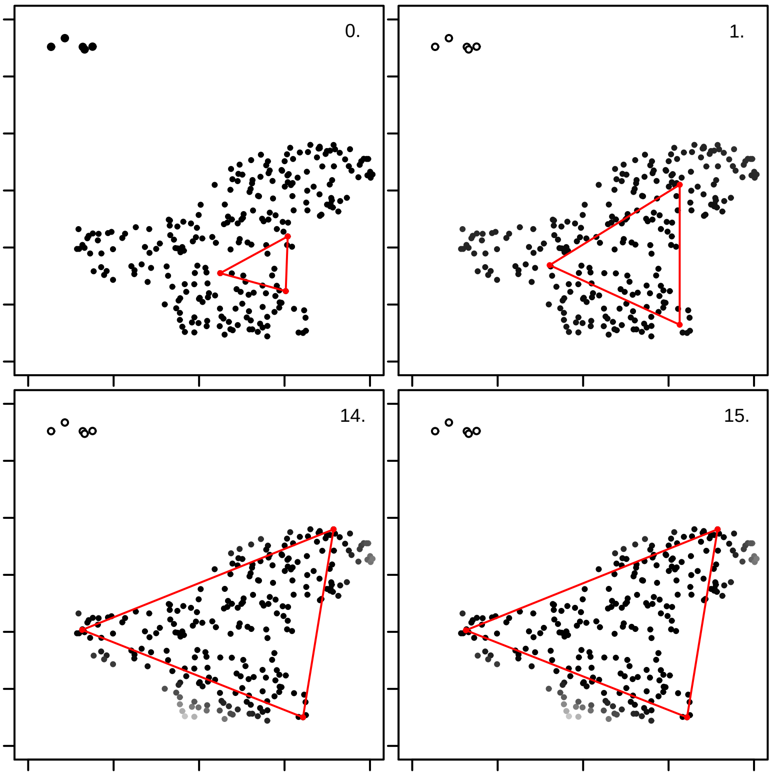


Figure 5: Individual iterations of the robust archetypal analysis which led to the solution presented in Figure 4.

Initialization Step 2 initializes the archetypes; a good initialization is important as a bad selection can cause slow convergence, convergence to a local minimum or even a non-robust solution. A common approach is to draw the initial archetypes from the complete data set. This can lead to the selection of an outlier as initial archetype. Approaches to select initial archetypes from the majority of the data are, for example, to draw them from the subset of data points which are inside some quantiles in each attribute or which are in the neighborhood of the median.

Figure 4(a) shows the robust archetypal analysis of the toy data set extended with five outliers. (b) The dotted line indicates the solution of the original $k = 3$ archetype algorithm; one archetype has gone on to “catch” the outliers. The $k = 3$ Bisquare archetypes solution is similar to the solution on the data set without outliers (Figure 1). (c) The gray scale of the data points indicate their final weights. The outliers have weight 0 (therefore filled with white color), their residuals do not contributed to the overall RSS (indicated with the dotted lines).

Figure 5 illustrates individual algorithm iterations of the archetypal analysis which leads to the solution presented in Figure 4. The algorithm converges in fifteen iterations, the individual plots show the initial setup (randomly initialized archetypes), the first, fourteenth and final iteration. The gray scale of the data points indicate their current weights. Note that in the initial setup all data points have weight 1. Already in the first iteration the weights of the outliers are very low (closely to 0) and decrease to 0 at the final iteration.

5 Application example

In this section we apply robust archetypal analysis on the Air-Pollution data set used by Cutler and Breiman in the original archetypal analysis paper (where they declare this problem as the "initial spark" for their study on archetypal analysis). Using a data set which already has been extensively studied allows us to compare the robust solutions with a well known solution.

The data consist of measurements of data relevant to air pollution in Los Angeles Basin in 1979. There are 330 cases consisting of daily measurements on the attributes ozone (OZONE), 500 millibar height (500MH), wind speed (WDSP), humidity (HMDTY), surface temperature (STMP), inversion base height (INVHT), pressure gradient (PRGRT), inversion base temperature (INVTMP), and visibility (VZBLTY). These data were standardized to have mean 0 and variance 1. Cutler and Breiman (1994) focus on three archetypes; left part of Table 1 lists the percentile value of each variable in an archetype as compared to the data (Figure 4 in Cutler and Breiman, 1994). Their interpretation of the archetypes is : "Archetype 1 is high in OZONE, 500MH, HMDTY, STMP, and INVTMP and low in INVHT and VZBLTY. This indicates a typical hot summer day. The nature of the other two archetypes is less clear; Archetype 3 seems to represent cooler days toward winter."

We contaminate the data set with a group of 5 outliers. The attributes of the outliers are calculated by $x * \text{MAX} + \text{IQR}$ with MAX the maximum and IQR the interquartile range of the attribute and x randomly drawn from $[1.5, 2]$. Three archetypes are computed with the original and the robust algorithm. Figure 6 shows the *panorama plot*, a simple diagnostic tool to inspect arbitrary high-dimensional archetypes solutions. For each archetype (individual panels) the Euclidean distance (y-axis) between the archetype and each data point is shown in ascending order (x-axis); other archetypes are shown as cross symbols. The underlying idea is to look at the data from the viewpoint of an archetype ("to watch its panorama"). This uncovers archetypes having only a few near data points – which then can be considered as candidates for outliers. In case of the original algorithm, Figure 6(a), the second archetype is the archetype gone on to "catch" the outliers – it is the archetype near to the outliers. In contrast, the robust

	Original data set			Outlier data set		
	A1	A2	A3	A1	A2	A3
OZONE	12	3	91	12	12	90
500MH	45	5	96	64	7	91
WDSP	8	91	43	8	89	43
HMDTY	11	74	78	11	65	77
STMP	15	6	95	21	11	93
INVHT	67	100	7	63	70	8
PRGRT	2	95	55	2	91	54
INVTMP	30	3	95	40	6	92
VZBLTY	88	77	15	76	76	15

Table 1: Percentile profiles of the archetypes computed on the original data set and the outlier data set.

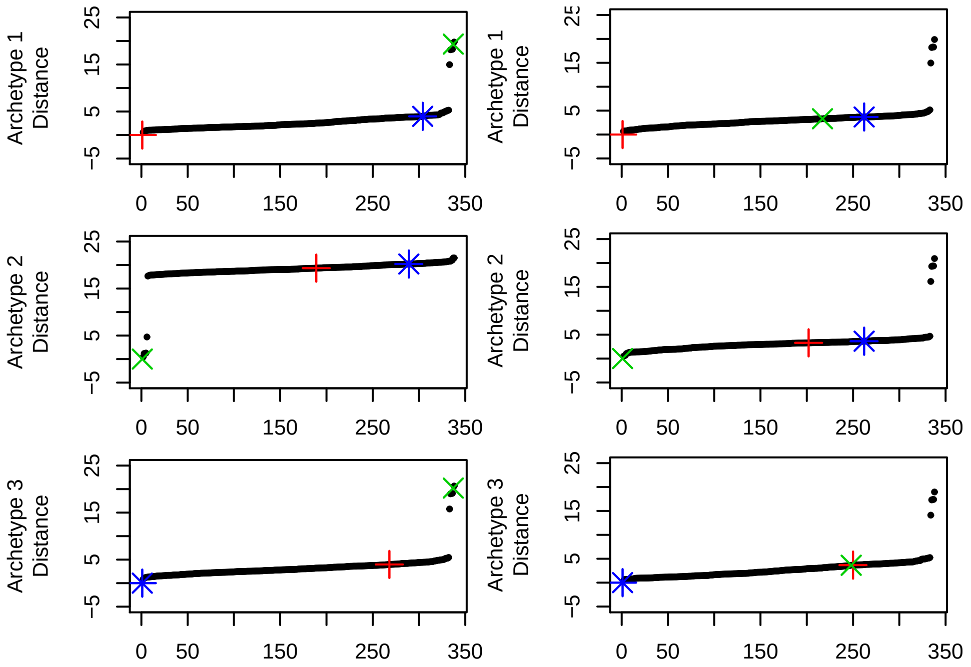
algorithm, Figure 6(b), focuses on the majority of the data points – no archetype is in the neighborhood of the outlier observations.

In the sense of the adapted citation of Rousseeuw and Leroy (2003) on robust analysis in the introduction, Figure 7(a) shows for each individual data point its residual length. The majority of the data has low residuals (note that variations from zero can occur due to numerical issues), whereas the five outliers stand out with high residuals. The calculated weights of the data points, Figure 7(b), fit accordingly: the majority of the data has high weights, the outliers low weights.

Finally, taking a look at the concrete robust archetypes values, right part of Table 1, shows that they are very similar to the archetypes calculated on the original data set (without the outliers).

6 Summary

The present paper adapts the archetypal analysis estimator by Cutler and Breiman (1994) to allow weighted and robust archetypal analysis. Weighted archetypes enables to represent additional information like importance of and correlation between observations. Robust archetypes focus on the majority of the data set; data points which behave differently from the large majority achieve less weight in the fitting process. The proposed estimator is an M-estimator whose minimization problem is solved by an iteratively reweighted least squares fitting algorithm. The artificial toy example and the real world application example shows that in presence of outliers the robust algorithm gives reliable archetypes which are greatly similar to the archetypes calculated on the same data set without outliers.



(a) (b)

Figure 6: Panorama plots: The distance between each archetype and each data point in case of the (a) original algorithm and (b) robust algorithm.

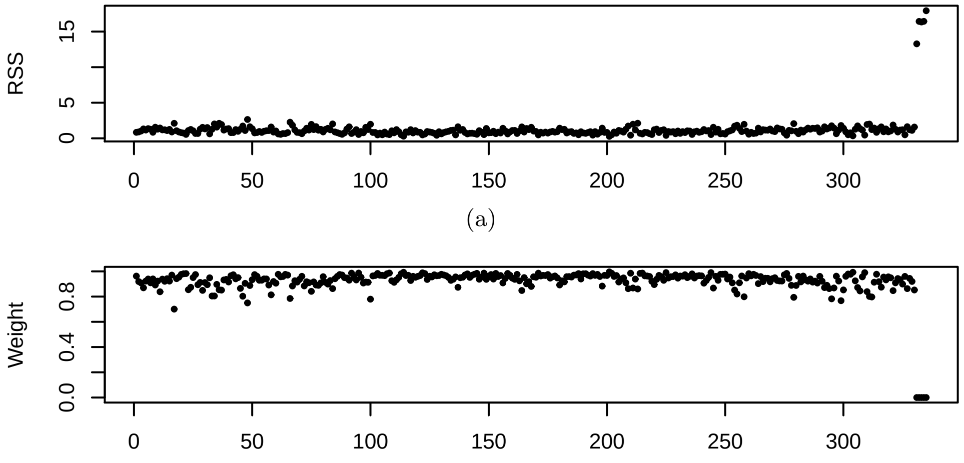
Computational details

All computations and graphics have been done using **R** 2.10.1 (**R** Development Core Team, 2009) and the package **archetypes** 2.0 (Eugster and Leisch, 2009) which implements weighted and robust archetypal analysis as introduced in this paper. It relies on package **nnls** (Mullen and van Stokkum, 2010) for non-negative least squares (NNLS). **R** itself and all packages used are freely available under the terms of the General Public License from the Comprehensive **R** Archive Network at `http://CRAN.R-project.org/`. Code for replicating our analysis is available in the **archetypes** package. The toy data set analysis is executed via:

```
R> demo("robust-toy", package = "archetypes")
```

The Air-Pollution data set analysis is executed via:

```
R> demo("robust-ozone", package = "archetypes")
```



(a)

(b)

Figure 7: The residual length and weight of each individual data point. The majority of the data has (a) low residuals and (b) high weights. The outliers have (a) high residuals and (b) low weights.

The source code files are accessible via:

```
R> edit(file = system.file("demo", "robust-toy.R",
+                             package = "archetypes"))
```

```
R> edit(file = system.file("demo", "robust-ozone.R",
+                             package = "archetypes"))
```

References

Christian Bauckhage and Christian Thurau. Making archetypal analysis practical. In *Proceedings of the 31st DAGM Symposium on Pattern Recognition*, pages 272–281, 2009.

Ben H. P. Chan, Daniel A. Mitchell, and Lawrence E. Cram. Archetypal analysis of galaxy spectra. *Monthly Notice of the Royal Astronomical Society*, 338:790–795, 2003.

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

David L. Donoho and Peter J. Huber. The notion of breakdown point. In *A Festschrift for Erich Lehmann*, pages 157–184, 1983.

Manuel J. A. Eugster and Friedrich Leisch. From Spider-man to Hero – archetypal analysis in R. *Journal of Statistical Software*, 30(8):1–23, 2009.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics.* John Wiley & Sons, Inc., 2 edition, 2009.

Shan Li, Paul Wang, Jordan Louviere, and Richard Carson. Archetypal analysis: A new way to segment markets based on extreme individuals. In *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution. Proceedings of the ANZMAC 2003 Conference, December 1-3, 2003*, pages 1674–1679, 2003.

Stephan Morgenthaler. A survey of robust statistics. *Statistical Methods and Applications*, 15(3):271–293, 2007.

Katharine M. Mullen and Ivo H. M. van Stokkum. ***nnls**: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*, 2010. URL `http://CRAN.R-Project.org/package=nnls`. **R** package version 1.2.

Giovanni C. Porzio, Giancarlo Ragozini, and Domenico Vistocco. On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry*, 24 (5):419–437, 2008.

John Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

R Development Core Team. ***R**: A Language and Environment for Statistical Computing.* **R** Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, Inc., 2003.