Computational Statistics and Data Analysis 55 (2011) 1215–1225

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Weighted and robust archetypal analysis

Manuel J.A. Eugster *, Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, 80539 München, Germany

ARTICLE INFO

Article history:
Received 3 May 2010
Received in revised form 13 October 2010
Accepted 14 October 2010
Available online 29 October 2010

Keywords:
Robust archetypal analysis
M-estimator
Breakdown point
Iteratively reweighted least squares

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. The original algorithm is adapted to be a robust M-estimator and an iteratively reweighted least squares fitting algorithm is presented. As a required first step, the weighted archetypal problem is formulated and solved. The algorithm is demonstrated using an artificial example, a real world example and a detailed simulation study.

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 is primarily 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

* Corresponding author. Tel.: +49 89 2180 6254; fax: +49 89 2180 5040.
E-mail addresses: Manuel.Eugster@stat.uni-muenchen.de (M.J.A. Eugster), Friedrich.Leisch@stat.uni-muenchen.de (F. Leisch).
URLs: http://www.statistik.lmu.de/~eugster (M.J.A. Eugster), http://www.statistik.lmu.de/~leisch (F. Leisch).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.10.017

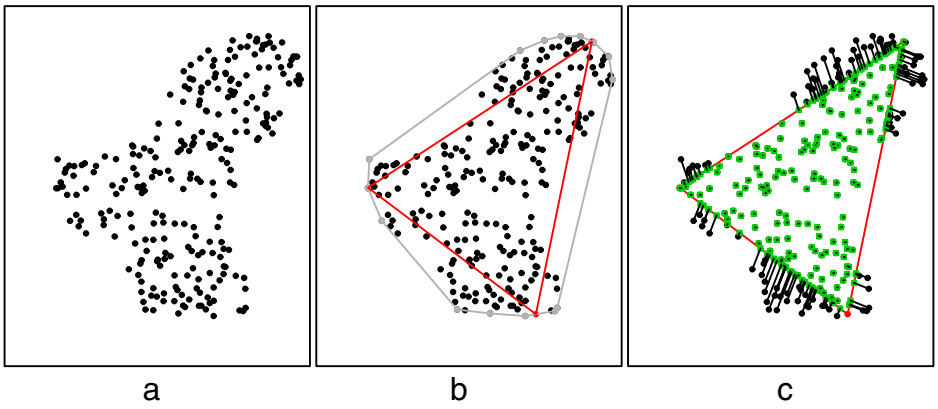

Fig. 1. (a) Artificial toy data set. (b) Approximation of the convex hull (outer polygon) by three archetypes (inner triangle). (c) Approximation of the data through the three archetypes and the corresponding α values.

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). Section 6 presents a structured simulation study to analyze the algorithm's robustness and convergence with respect to data dimension, number of outliers and distance of outliers to the majority of data. Finally, in Section 7 the conclusions and future work 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 km-dimensional archetypes. More precisely, to find the two $n \times k$ coefficient matrices α and β which minimize the residual sum of squares

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

subject to the constraints

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

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

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.

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.

Fig. 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. Fig. 1(b) shows the archetypes, their approximation of the convex hull (the inner triangle) and the convex hull of the data (outer polygon). Fig. 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.

Here, we discuss the breakdown of the archetypes matrix Z, i.e., estimator $T = Z$. For a given k archetypal analysis reaches the worst possible value, $\epsilon_n^*(Z, X) = 1/n$ 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; Fig. 2 illustrates this scenario. Note how one of the archetypes has gone on to "catch" this outlier observation when the cross data 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

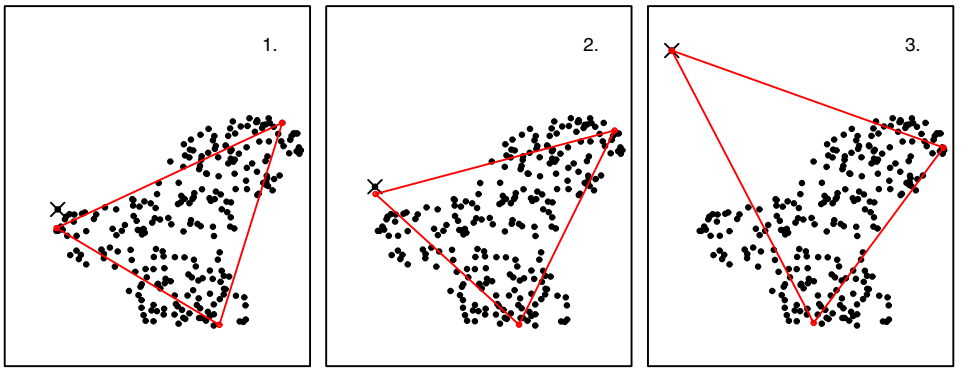

Fig. 2. Behavior of the archetypes (triangle) when one data point (cross) moves away.

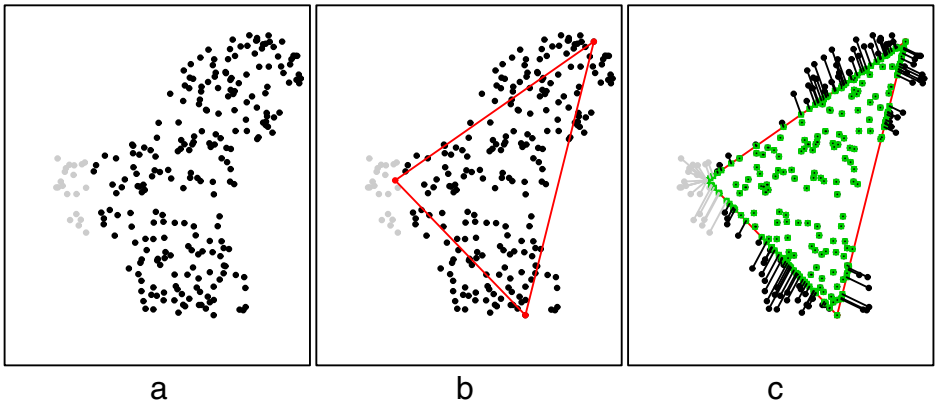

Fig. 3. Weighted archetypal analysis where gray data points weigh 0.8 and black data points 1.

outlier is approximated well, than 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, Eq. (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

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

Weighting the residuals is equivalent to weighting the data set:

$$\begin{aligned} W(X - \alpha Z^\top) &= W(X - \alpha(X^\top\beta)^\top) = W(X - \alpha\beta^\top X) \\ &= WX - W(\alpha\beta^\top X) = WX - (W\alpha)(\beta^\top W^{-1})(WX) \\ &= \tilde{X} - \tilde{\alpha}\tilde{\beta}\tilde{X} = \tilde{X} - \tilde{\alpha}\tilde{Z}^\top. \end{aligned}$$

Therefore the problem can be reformulated as minimizing

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

This reformulation allows the usage of the original algorithm with the additional pre-processing step to calculate $\tilde{X}$ and the additional post-processing step to recalculate α and β 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 represent 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 the case of archetypal analysis. Furthermore, W can be an arbitrary square matrix, for example, a matrix to decorrelate the observations.

Fig. 3 illustrates the weighted archetypal analysis of the toy data set for $k = 3$. (a) Gray data points weigh 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 weighted RSS.

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 the square; this yields 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 ($i = 1, \ldots, n$) 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 (generalized) *Bisquare* objective $\rho(\cdot)$ and weight $w(\cdot)$ functions are defined as $\rho(R) = \sum_{i=1}^{n} \tilde{\rho}(R_i)$ and $w(R) = \mathrm{diag}(\tilde{w}(R_i))$, $i = 1, \ldots, n$ with R_i the m dimensional residual of the ith observation and

$$\tilde{\rho}(R_i) = \begin{cases} \dfrac{c^2}{6}\left(1 - \left(1 - \left\|\dfrac{R_i}{c}\right\|_1^2\right)^3\right), & \text{for } \|R_i\|_1 < c \\ \dfrac{c^2}{6}, & \text{for } \|R_i\|_1 \geq c, \end{cases}$$

$$\tilde{w}(R_i) = \begin{cases} \left(1 - \left\|\dfrac{R_i}{c}\right\|_1^2\right)^2, & \text{for } \|R_i\|_1 < c \\ 0, & \text{for } \|R_i\|_1 \geq c. \end{cases}$$

The value c is a tuning parameter; practical application showed that $c = 6s$ with s the median of the residuals unequal to zero works well (following Cleveland, 1979). $\|\cdot\|_1$ denotes the 1-norm; on the lines of the absolute value in the original one-dimensional function definitions.

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

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

or according to Eq. (3),

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

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, providing 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.

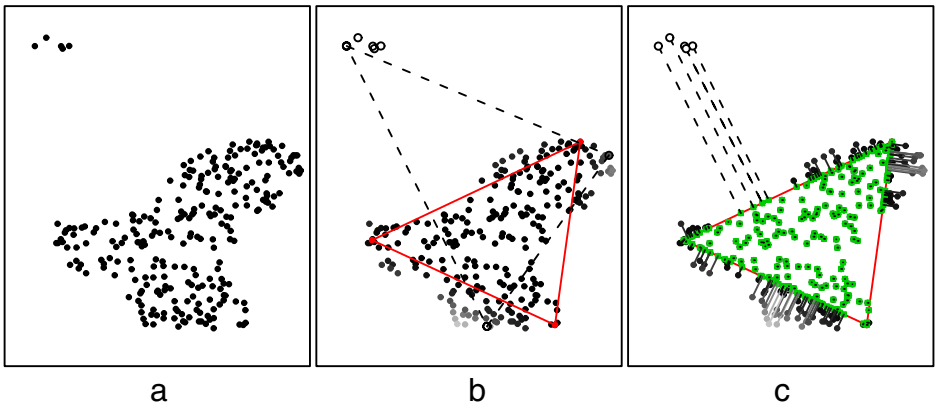

Fig. 4. Robust archetypal analysis; the gray scale of the data points indicate their final weights. Note that the outliers have weight 0 (unfilled).

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.

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 randomly 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. Note that these initialization methods do not ensure a good initialization (i.e., no outliers as initial archetypes) and several starts with different initializations are recommended.

Recalculation. Step 3 (the loop of the algorithm) computes the coefficient matrices α and β and the archetypes Z on the weighted data set X^t. For the final result the coefficient matrices have to be recalculated for the unweighted data set X. Step 4 again (as in each iteration of the loop) solves n and k convex least squares problems to find the best α and the best β given the set of archetypes Z and now the unweighted data set X (see Eugster and Leisch, 2009, for details on the convex least squares problems).

Computational complexity. The complexity of the algorithm is determined by the complexity of the underlying non-negative convex least squares method ($n + k$ problems per iteration and for the robust algorithm additional $n + k$ problems in Step 4) and the number of iterations until its convergence. In the concrete implementation we use the iterative NNLS algorithm defined by Lawson and Hanson (1974). The authors prove the convergence of the algorithm, but the number of iterations is dependent on the concrete problem. This behavior propagates to the archetypal analysis algorithm for which Cutler and Breiman (1994) also prove its convergence, but again the number of iterations is not fixed. Currently, we are not able to determine the theoretical computational complexity (i.e., Big O notation) for the algorithm. Therefore, Eugster and Leisch (2009) provide a simulation study to show how the (original) algorithm scales with numbers of observations, attributes and archetypes; and Section 6 provides a simulation study to compare the convergence of the original and robust algorithms. Note that the implementation is flexible and allows us to replace NNLS with other algorithms.

Fig. 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 (Fig. 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 contribute to the overall RSS (indicated with the dotted lines).

Fig. 5 illustrates individual algorithm iterations of the archetypal analysis which leads to the solution presented in Fig. 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 (close to 0) and decrease to 0 at the final iteration.

5. Application example

In this section we apply a 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

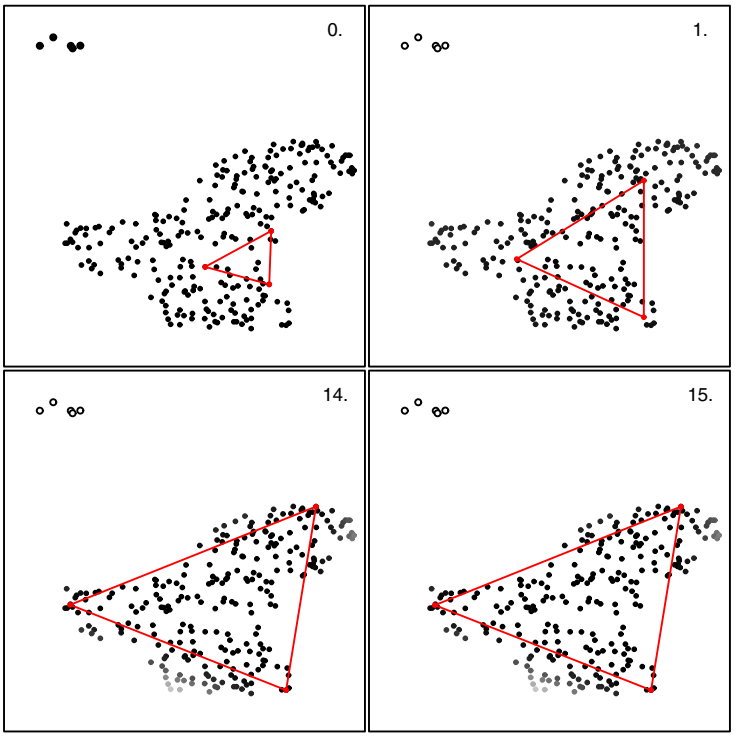

Fig. 5. Individual iterations of the robust archetypal analysis which led to the solution presented in Fig. 4.

Table 1
Percentile profiles of the archetypes computed by the original algorithm and the robust algorithm on the original data set and the outlier data set.

Data set	Air-Pollution						Air-Pollution+Outliers					
Algorithm	Original			Robust			Original			Robust		
Archetypes	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
OZONE	91	12	3	91	12	12	76	100	3	89	12	12
500MH	96	45	5	92	45	6	64	100	6	91	64	7
WDSP	43	8	91	43	8	91	27	100	63	43	8	89
HMDTY	78	11	74	78	12	81	50	100	19	77	11	63
STMP	95	15	6	95	16	11	73	100	5	91	21	11
INVHT	7	67	100	11	66	71	1	100	99	10	63	70
PRGRT	55	2	95	57	5	93	38	100	36	56	2	91
INVTMP	95	30	3	93	30	5	79	100	5	92	40	5
VZBLTY	15	88	77	15	88	77	9	100	87	15	76	76

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 consists of measurements of data relevant to air pollution in the 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; the 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). The percentile value indicates in which percentile of the data set an archetype in a specific variable is. For example, the OZONE value in archetype 1 is 91, i.e., the archetype 1 OZONE value is in the 91th percentile of the 330 OZONE cases in the data. Cutler and Breiman interpret the archetypes as follows: "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 $x * \mathrm{MAX} + \mathrm{IQR}$ 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 (the right part of Table 1). Fig. 6 shows the *panorama plot*, a simple diagnostic tool to inspect arbitrary high-dimensional archetypes solutions. For each archetype (individual panels) the

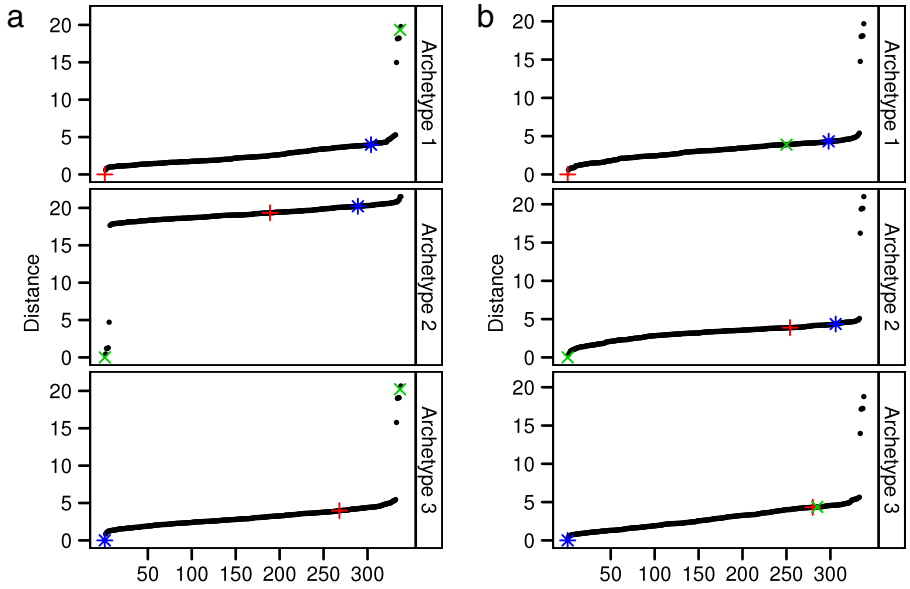

Fig. 6. Panorama plots: The distance between each archetype and each data point in case of the (a) original algorithm and (b) robust algorithm.

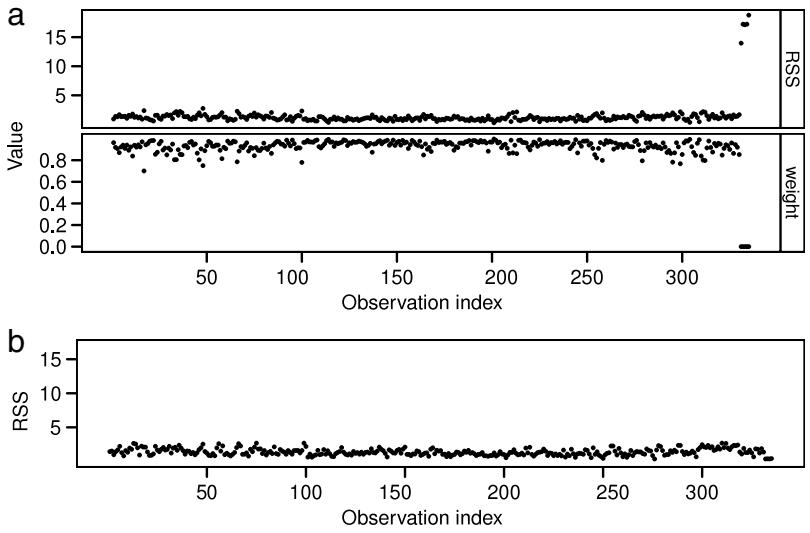

Fig. 7. The residual length and weight of each individual data point. (a) In case of the robust algorithm the majority of the data have low residuals and high weights. The outliers have high residuals and low weights. (b) In case of the original algorithm all data points have low residuals (and weights 1).

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 the case of the original algorithm, Fig. 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 algorithm, Fig. 6(b), focuses on the majority of the data points — no archetype is in the neighborhood of the outlying observations.

In the sense of the adapted citation of Rousseeuw and Leroy (2003) on robust analysis in the introduction, Fig. 7(a, top panel) 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, Fig. 7(a, bottom panel), fit accordingly: the majority of the data has high weights, the outliers low weights. By comparison, all data points have low residuals (and weights 1) in case of the original algorithm, Fig. 7(b).

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).

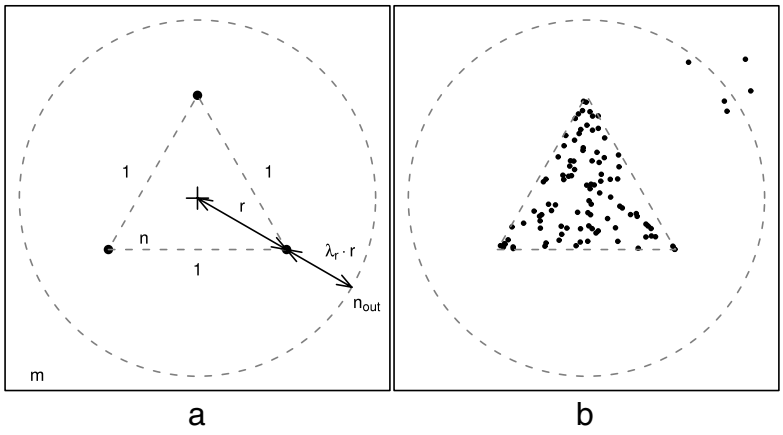

Fig. 8. Schematic representation of the simulation setup: (a) Basis is an m-simplex as true data generating process, enclosed by an m-sphere as mean for the outlier generating process. (b) An exemplar data set with $m = 2$, $n = 100$, $n_{\text{out}} = 5$, $\lambda_r = 2$, $\sigma = 0.05$.

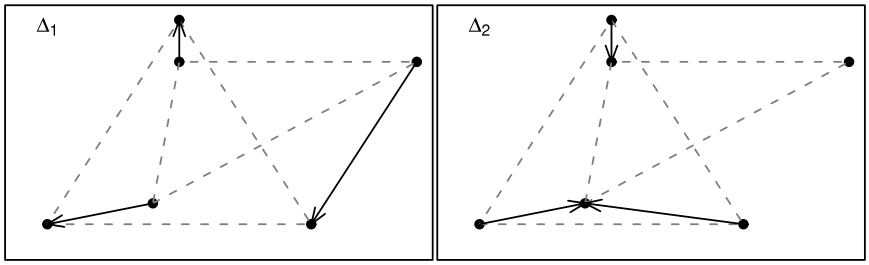

Fig. 9. The performance measure is the total distance between the computed archetypes and the nearest true archetypes (Δ_1) and the total distance between the true archetypes and the nearest computed archetypes (Δ_2).

6. Simulation study

So far, two exemplar data sets were used to present robust archetypal analysis and to demonstrate its proper functioning. This section now analyzes the algorithm's robustness and convergence with respect to data dimension, number of outliers and distance of outliers to the majority of data in a structured way. The simulation setup follows Hothorn et al. (2005, cf. the *simulation problem*), the analysis follows Eugster et al. (2008). The two crucial points of this simulation study are (1) to define a basis setup which enables the measurement of robustness without the effects of other algorithm characteristics like structural stability or convergence, and (2) to define a performance measure which reflects robustness in numbers — as the residual sum of squares (RSS) is not meaningful for this case. Note that this section presents selected results of the simulation study, the complete results are available in the supplemental material (see the section on computational details).

We define a uniformly distributed regular m-simplex (an m-dimensional polytope of $m + 1$ vertices and equally long edges; note that the common name n-simplex conflicts with our notation) as a true data generating process. The outlier generating process is defined as the multivariate normal distribution with the covariance matrix $\Sigma = \sigma I_m$ and the mean μ on a uniformly distributed m-sphere (the generalization of the surface of an ordinary sphere to dimension m) of radius $r + \lambda_r \cdot r$, with r the distance between the center and an m-simplex vertex and λ_r an expansion factor. n observations and n_{out} outliers are drawn from the data and outlier generating process, respectively. Fig. 8(a) illustrates the setup, Fig. 8(b) shows an exemplar data set with $m = 2$, $n = 100$, $n_{\text{out}} = 5$, $\lambda_r = 2$, $\sigma = 0.05$. We define the performance measure as the total Euclidean distance between the computed archetypes and the nearest true archetype in each case (Δ_1), and the total Euclidean distance between the true archetypes and the nearest computed archetypes in each case (Δ_2). Note that this definition must not yield in a one-to-one assignment; Fig. 9 illustrates a potential case. Then, a good solution has small and nearly similar Δ_1 and Δ_2 (necessary condition). A perfect solution is a one-to-one assignment with Δ_1 and Δ_2 of value zero.

The first simulation investigates the original and robust algorithm when the correct number of archetypes k (i.e., $k = m + 1$) is known: A data set is generated for each combination of $m = 2, 3, 10, 15$ attributes, $n = 2000$ observations, $n_{\text{out}} = 5, 20, 100$ outliers, $\lambda_r = 2, 3, 5, 15$ radius expansion factors, $\sigma = 0.05$ covariance and 100 replications in each case. Each configuration is fitted with randomly chosen initial archetypes, stop criteria are 100 iterations or an improvement less than the square root of the machine epsilon ($\sqrt{(2.22 \cdot 10^{-16})}$). For each of the $m \times n \times n_{\text{out}} \times \lambda_r \times \Sigma \times (1, \ldots, 100)$ original and robust fits the two distance measures, the number of iterations, the computation time and the RSS are reported (all in all 48 000 measured values). The constant number of observations n allows the discussion of the algorithms in view of the curse of dimensionality.

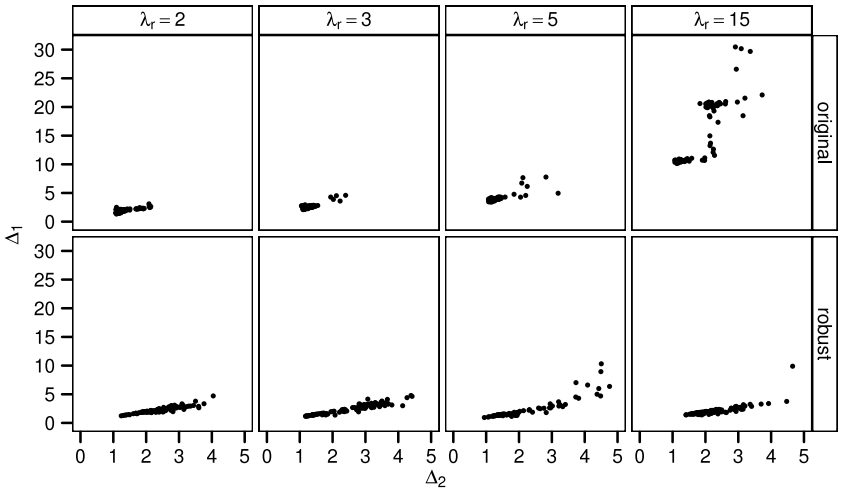

Fig. 10. Distances Δ_1 versus Δ_2 for sample fits of both algorithms in case of the configuration $m = 10$, $n_{\text{out}} = 100$, $\lambda_r = 2, 3, 5, 15$.

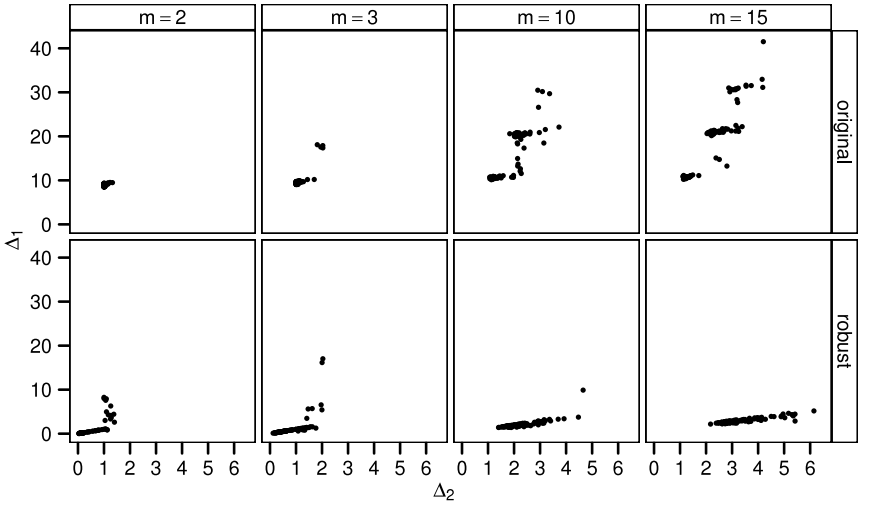

Fig. 11. Distances for each dimension $m = 2, 3, 10, 15$ and $n_{\text{out}} = 100$, $\lambda_r = 15$. Note that the y-axis scale is different per distance.

Fig. 10 shows the distances Δ_1 (y-axis, scale from 0 to 30) versus Δ_2 (x-axis, scale from 0 to 5) for each fit of the original and the robust algorithm in case of the configuration $m = 10$, $n_{\text{out}} = 100$, $\lambda_r = 2, 3, 5, 15$. In case of the original algorithm, Δ_1 increases with increasing outlier distance λ_r; one computed archetype always goes for the outlier group. One the other side, Δ_2 stays small, so the remaining computed archetypes fit well to the true archetypes. In case of the robust algorithm, both distances remain small with increasing outlier distance; the robust algorithm stays robust and finds good solutions for the majority of the data (with a few exceptions). Noticeable is that the robust algorithm's distances are more variable and that Δ_2 is slightly higher. The first fact indicates less stable solutions (part of our ongoing research, see discussion in Section 7). The second fact is, amongst other things, explored in the following paragraph. The described patterns appear for all other configurations as well.

Fig. 11 shows the distances for each dimension $m = 2, 3, 10, 15$ and $n_{\text{out}} = 100$, $\lambda_r = 15$. Δ_1 also remains small over the number of dimensions for the robust algorithm and increases for the original algorithm (as expected). Interesting is the fact that Δ_2 increases over the number of dimensions for the robust algorithm. This indicates that the computed solution is smaller than the true m-simplex; observations near the m-simplex boundary are weighted down even though they are from the true data generating process. This is an effect of the curse of dimensionality, in higher dimensions the robust algorithm finds the majority of the data's true shape but not in the exact size.

As already stated, the archetypal analysis algorithm is computer-intensive and a fast convergence is desired. Fig. 12 compares the median number of iterations between the original (dashed line) and the robust (solid line) algorithm for each configuration. Ignoring $m = 2$, the robust algorithm needs less or a nearly equal number of iterations in most of the cases. In the case $m = 2$ both algorithms generally converge much slower, and the robust algorithm needs twice as many iterations than the original algorithm. One supposable reason could be the large number of observations ($n = 2000$) in relation to

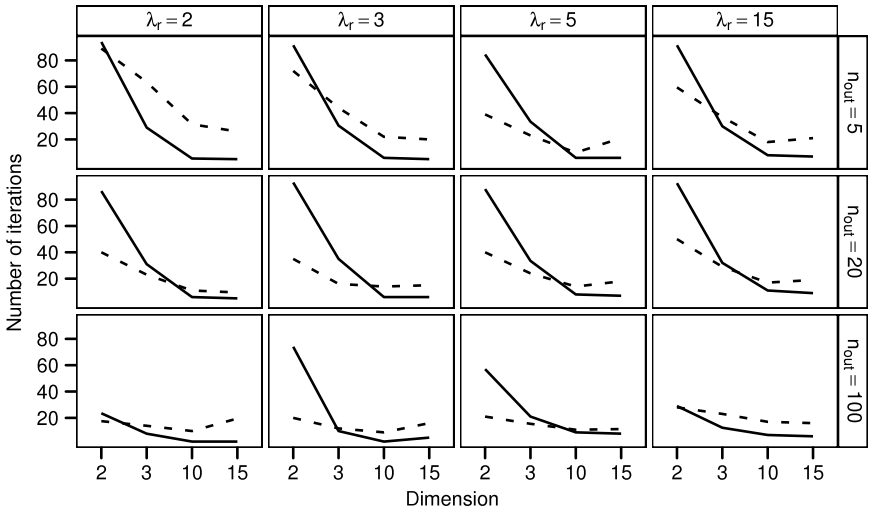

Fig. 12. Median number of iterations for the original (dashed line) and the robust algorithm (solid line).

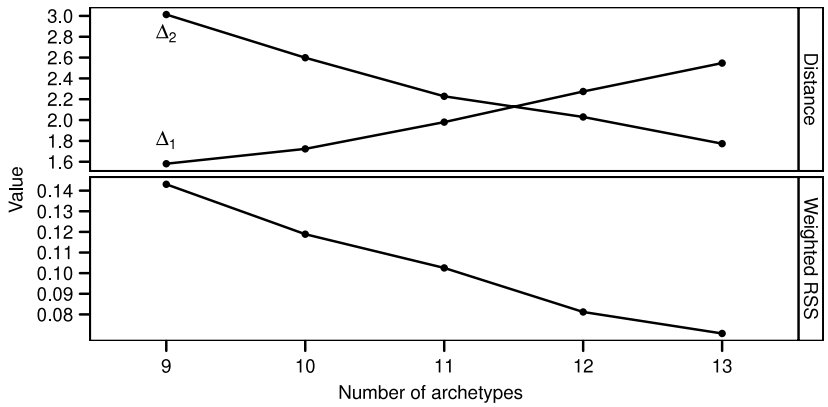

Fig. 13. Median distances Δ_1 and Δ_2 and median weighted RSS for the robust algorithm in case of the configuration $m = 10, n_{\text{out}} = 100, \lambda_r = 2, 3, 5, 15$, $k = 9, \ldots, 13$.

the space of the m-simplex with side length 1 and the consequential closeness of the observations; but detailed simulations need to be done to analyze this phenomenon.

The second simulation investigates the robustness of the algorithm when the number of correct archetypes k (i.e., $k = m + 1$) is unknown: The simulation setup is equivalent to the first one, but an additional fit is computed for each $k = m - 1, m, m + 1, m + 2, m + 3$. Fig. 13 (top panel) shows the median distances Δ_1 and Δ_2 of the 100 replications for the robust algorithm and $m = 10$, $n_{\text{out}} = 100$, $\lambda_r = 15$ and $k = 9, \ldots, 13$. The distances are inside a reasonable range, the algorithm is robust in case of wrong k. Δ_1 increases and Δ_2 decreases with increasing number of archetypes; the true solution is near their intersection point. This reflects the different assignment scenarios: (1) less computed archetypes than true archetypes — a one-to-one assignment with the remaining true archetypes for Δ_1 and a one-to-many assignment for Δ_2; (2) as many computed archetypes as true archetypes — a one-to-one assignment; (3) more computed archetypes than true archetypes — a one-to-many assignment for Δ_1 and one-to-one assignment with the remaining computed archetypes for Δ_2. Obviously, this performance measure is only usable in simulation studies where the data generating process and the true archetypes are known. In real world applications the weighted RSS is a possible performance measure; Fig. 13 (bottom panel) shows the median weighted RSS. However, as this screeplot indicates, this performance measure often allows no well-defined decision. This is a problem of great complexity — algorithm characteristics, like the structural stability or the convergence, play a major role. Currently we have no sound solution and this is part of our ongoing research (see discussion in Section 7).

7. Summary

The present paper adapts the archetypal analysis estimator by Cutler and Breiman (1994) to allow weighted and robust archetypal analysis. Weighted archetypes enable us 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 the presence of outliers the robust algorithm gives reliable archetypes which are greatly similar to the archetypes calculated on the same data set without outliers. The simulation study analyzes the algorithm with respect to data dimension, number of outliers, distance of outliers to the majority of data and number of archetypes in a structured way. The study shows that the algorithm is highly robust and often converges faster than the original algorithm.

Major future work is the evaluation of the algorithm according to its structural stability, especially in the case of unknown and wrong number of archetypes. The goal is the development of a framework along the lines of the framework for "structure and reproducibility of cluster solutions" by Dolnicar and Leisch (2010). This includes, amongst other things, the definition of a criterion for stability of archetypal solutions.

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). The simulation study is done using the package **benchmark** 0.3-1 (Eugster et al., 2008). **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 (version $>$ 2.0-1). 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")
```

The full analysis of the simulation study is available via:

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

The source code file for a demo is accessible via (replace `***` with `toy`, `ozone` and `simulation`):

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

Acknowledgements

The authors would like to thank Sebastian Kaiser for valuable discussions according to the simulation study and two anonymous referees for constructive comments and suggestions.

References

Bauckhage, C., Thurau, C., 2009. Making archetypal analysis practical. In: Proceedings of the 31st DAGM Symposium on Pattern Recognition, pp. 272–281.
Chan, B.H.P., Mitchell, D.A., Cram, L.E., 2003. Archetypal analysis of galaxy spectra. Monthly Notice of the Royal Astronomical Society 338, 790–795.
Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74 (368), 829–836.
Cutler, A., Breiman, L., 1994. Archetypal analysis. Technometrics 36 (4), 338–347.
Dolnicar, S., Leisch, F., 2010. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. Marketing Letters 21, 83–101.
Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: A Festschrift for Erich Lehmann. pp. 157–184.
Eugster, M.J.A., Hothorn, T., Leisch, F., 2008. Exploratory and inferential analysis of benchmark experiments. Technical Report 30, Institut für Statistik, Ludwig-Maximilians-Universität München, Germany. http://epub.ub.uni-muenchen.de/4134.
Eugster, M.J.A., Leisch, F., 2009. From Spider-man to Hero — archetypal analysis in R. Journal of Statistical Software 30 (8), 1–23.
Hothorn, T., Leisch, F., Zeileis, A., Hornik, K., 2005. The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics 14 (3), 675–699.
Huber, P.J., Ronchetti, E.M., 2009. Robust Statistics, 2nd ed. John Wiley & Sons, Inc.
Lawson, C.L., Hanson, R.J., 1974. Solving Least Squares Problems. Prentice-Hall.
Li, S., Wang, P., Louviere, J., Carson, R., 2003. 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, pp. 1674–1679.
Morgenthaler, S., 2007. A survey of robust statistics. Statistical Methods and Applications 15 (3), 271–293.
Mullen, K.M., van Stokkum, I.H.M., 2010. **nnls**: The Lawson–Hanson algorithm for non-negative least squares (NNLS). **R** package version 1.2. http://CRAN.R-Project.org/package=nnls.
Porzio, G.C., Ragozini, G., Vistocco, D., 2008. On the use of archetypes as benchmarks. Applied Stochastic Models in Business and Industry 24 (5), 419–437.
Quackenbush, J., 2002. Microarray data normalization and transformation. Nature Genetics 32, 496–501.
R Development Core Team. 2009. **R**: A Language and Environment for Statistical Computing. **R** Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.
Rousseeuw, P.J., Leroy, A.M., 2003. Robust Regression and Outlier Detection. John Wiley & Sons, Inc.