Archetypal Analysis

Soo-Young Kim

Colorado State University

December 2, 2015

Overview

Archetypal Analysis

- **Archetypal Analysis** approximates data points by prototypes that are themselves linear combinations of data points.

- Data: $x_1, \cdots, x_n$ be m-dimensional data points
- Archetypes: $z_1, \cdots, z_p$ are mixtures of the data values $\{x_i\}$

Archetypal Analysis

- The problem is to find $z_1, \cdots, z_p$ where

$$z_k = \sum_{j=1}^{n} \beta_{kj} x_j, k = 1, \cdots, p$$

- Need to find α_{ik} and β_{kj} that minimize (using a convex optimization)

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right|\right|^2 \\ &= \sum_{i=1}^{n} \left|\left| x_i - \sum_{k=1}^{p} \alpha_{ik} \sum_{j=1}^{n} \beta_{kj} x_j \right|\right|^2 \end{aligned}$$

subject to constraints

$\alpha_{ik} \geq 0$ and $\sum_{k=1}^{p} \alpha_{ik} = 1$

$\beta_{kj} \geq 0$ and $\sum_{j=1}^{p} \beta_{kj} = 1$

Archetypal Analysis

Proposition [Cultler, 1994]

Let C be the convex hull of $x_1, \cdots, x_n$. Let S be the set of data points on the boundary of C, and N be the cardinality of S.
If $1 < p < N$, there is a set of archetypes $z_1, \cdots, z_p$ on the boundary of C that minimize RSS

- For $p > 1$, the archetypes fall on the convex hull of the data.
- Thus, the archetypes are extreme data values such that all of the data can be well represented as convex mixtures of the archetypes.
- The overall problem is not convex, however, and so the algorithm converges to a local minimum of the criterion.
  (z_i's are constrained to be a mixture of data points)

Archetypal Analysis: Swiss Army Head-Dimension data

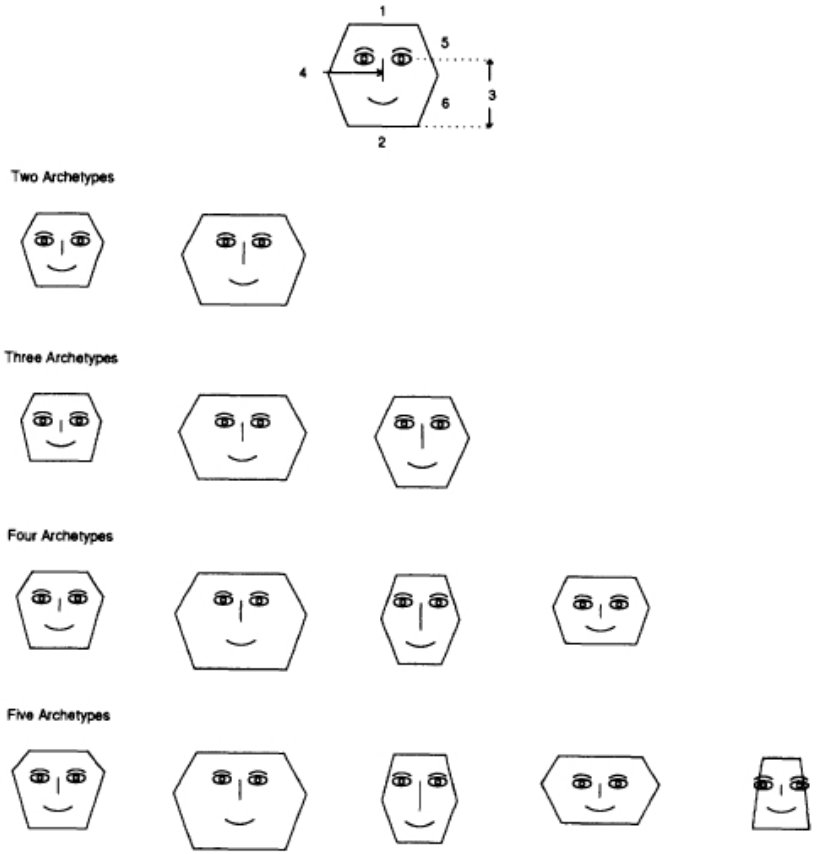


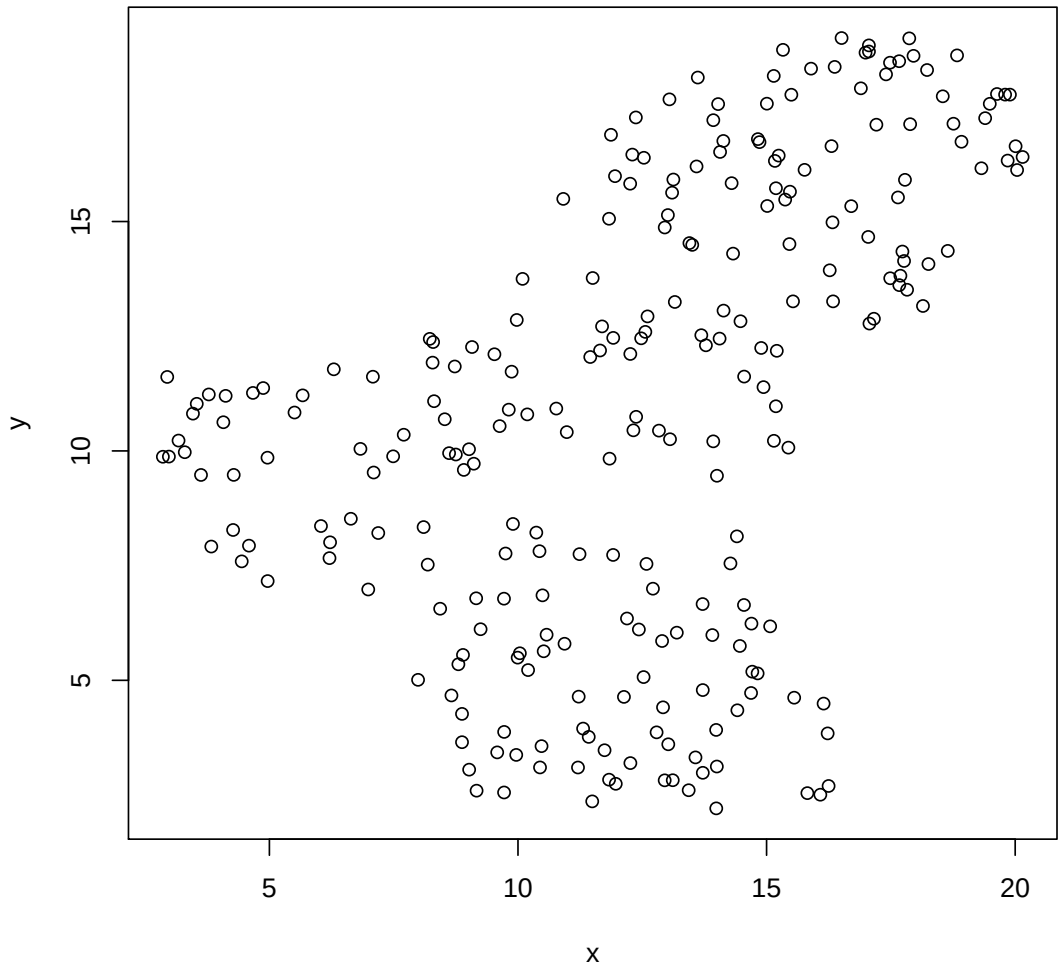
Figure 1. Archetypes for Head-Dimension Data.

The data consists of 6 measurement on each head. Idea: Each real individual can be well approximated by a mixture of the pure types or archetypes

Archetypal Analysis in R

Archetypal Analysis in R

Toy data



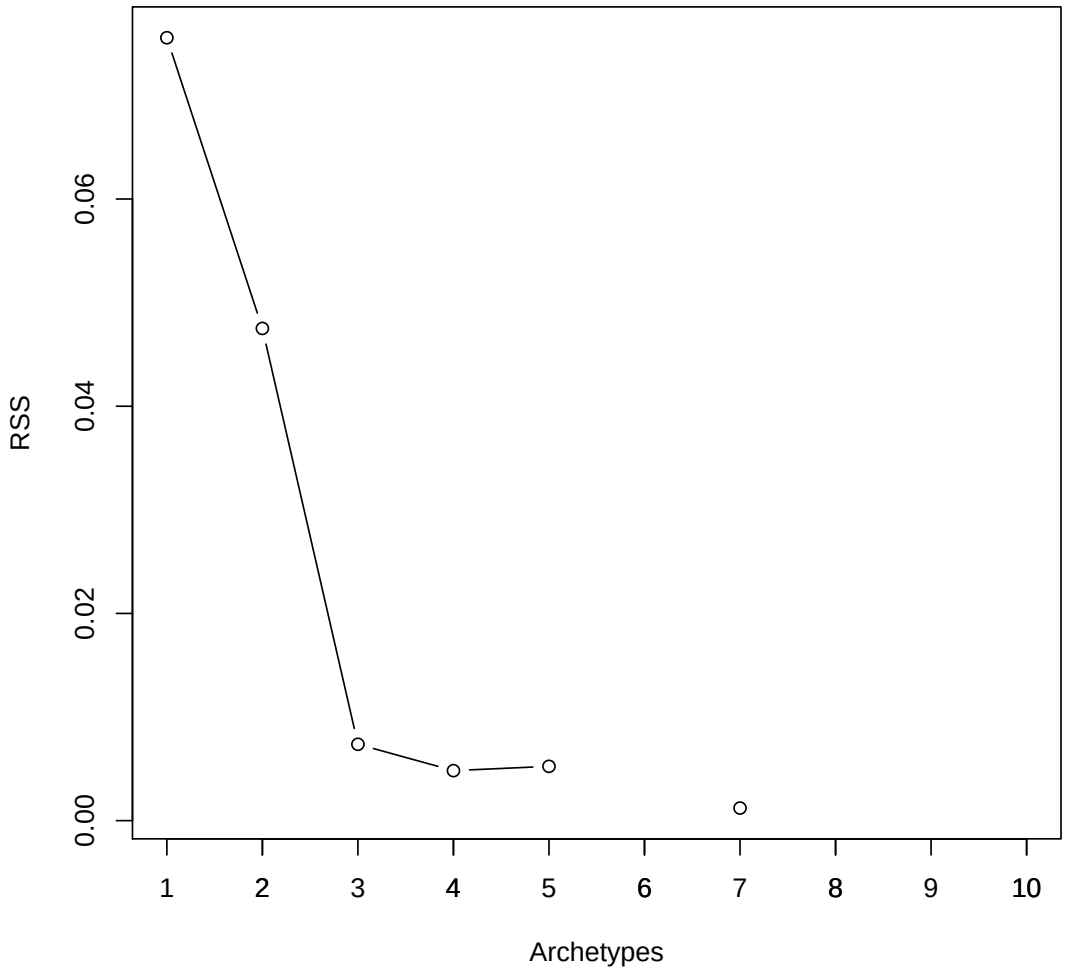
Archetypal Analysis in R

Example (R code)

```
library(archetypes)
data("toy")
set.seed(1986)
as <- stepArchetypes(data = toy, k = 1:10, verbose = FALSE,
 nrep = 4)
screeplot(as)
a7=bestModel(as[[7]])

a$archetypes
#plots
simplexplot(a)
xyplot(a, toy, chull = chull(toy)) #show convex hall
xyplot(a, toy, adata.show = TRUE)  #show approximated data
```

Archetypal Analysis in R



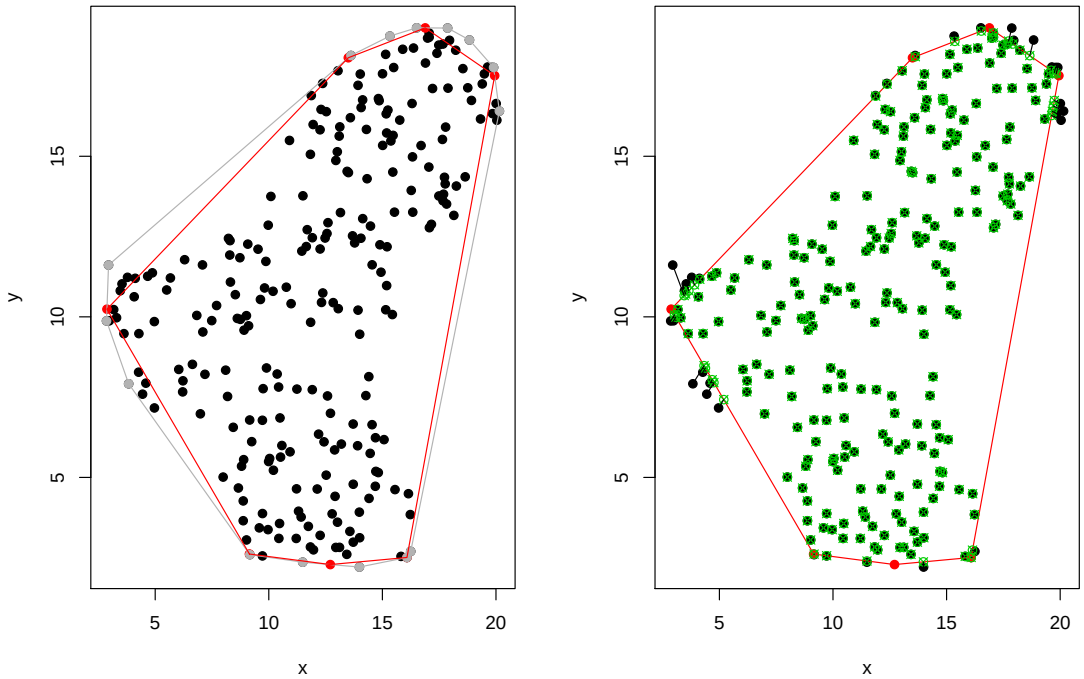
Archetypal Analysis in R

Example (R code)

```
R> a7
Archetypes object
Convergence after 100 iterations
with RSS = 0.001216349.

#seven final archetypes
R> a7$archetypes
              x         y
[1,] 16.081116  2.507586
[2,]  2.876206 10.239522
[3,]  9.147667  2.614262
[4,] 13.500297 18.067922
[5,] 16.884172 18.998137
[6,] 12.708133  2.286835
[7,] 19.942246 17.511102
```

Archetypal Analysis in R



The left plot: the archetypes, their approximation of the convex hull (red) and the convex hull (grey) of the data.
The right plot: the approximation of the data through the archetypes and the corresponding values (black)

Archetypal Analysis

- Archetypes are "extreme" or "pure" types of patterns such that each real data point can be well approximated by a mixture of the pure types or archetypes.
- Since archetypes are located on the prototypes on the convex hull of the data, the procedure can be sensitive to outliers.

References

Adele Cutler and Leo Breiman (1994)

Archetypal Analysis

Technometris 36(4), 338 – 347.

Manuel J. A. Eugster and Friedrich Leisch (2009)

Archetypal Analysis in R

Journal of Statistical Software 30(8).

The End