Archetypal Analysis

Mining the Extreme

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximiliams-Universität München

HIIT seminar, Helsinki Institute for Information Technology, 2012

Bachelor's degree: Software & Information Engineering, Vienna University of Technology, Austria

A Neural Network Toolbox in R

Master's degree: Computational Intelligence, Vienna University of Technology, Austria

Meta-Learning – Machine Learning with no Idea?

PhD: Working Group Computational Statistics, Department of Statistics, Ludwig-Maximilians-Universität München, Germany

Benchmark Experiments – A Tool for Analyzing Statistical Learning Algorithms

Archetypal analysis

The Merriam-Webster Online Dictionary (2008) defines an archetype as

> *"the original pattern or model of which all things of the same type are representations or copies."*

The aim of archetypal analysis is to find these "pure types"—the archetypes.

Archetypes in psychology: "Ancient or archaic images that derive from the collective unconscious".

Realizations, for examples, in fairy tales:

Hero: the bravest and the cleverest of all

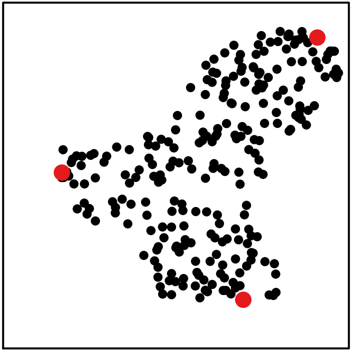
Monster: the scariest and most evil of all

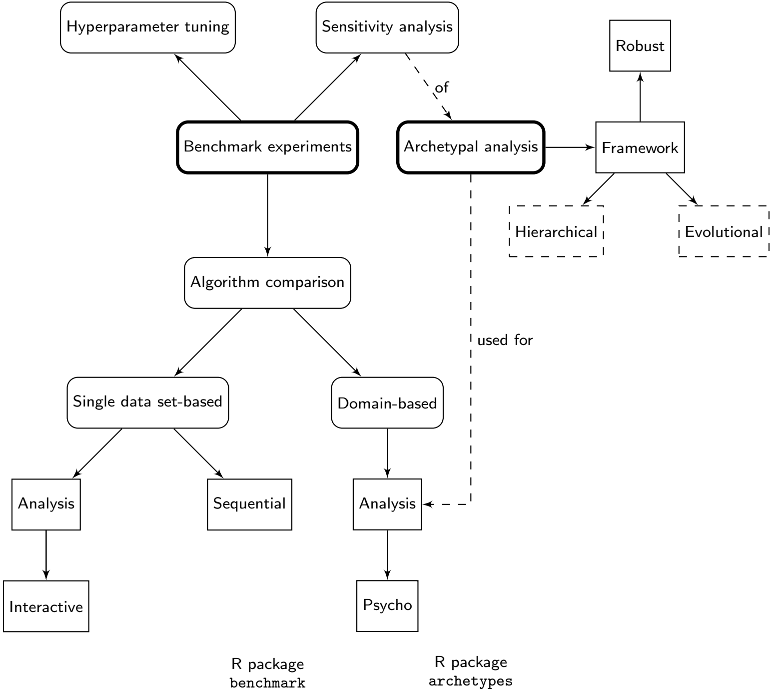
etc.: princess, stepmother, ...

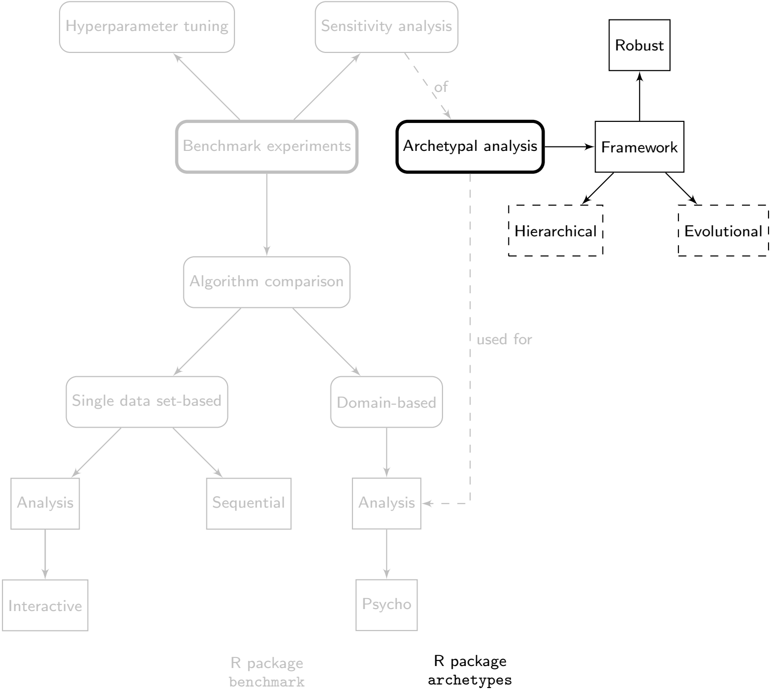
Archetypes in everyday language: Archetypal symbolism often arises in our common usage.

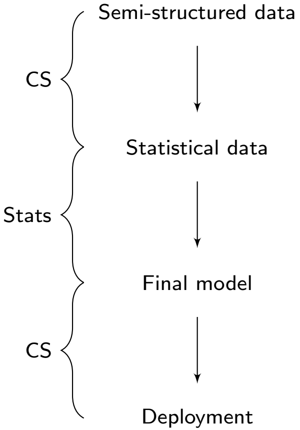
- "There is only one *Lionel Messi*."
- "A little bit of *Chuck Norris* is in everyone."

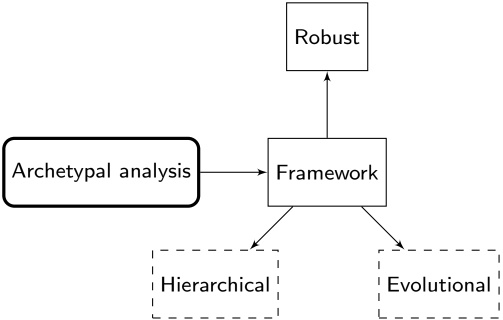
Archetypes in data analysis: Extremal observations in a (multivariate) data set.




Hyperparameter tuning
Sensitivity analysis
Robust
of
Benchmark experiments
Archetypal analysis
Framework
Hierarchical
Evolutional
Algorithm comparison
used for
Single data set-based
Domain-based
Analysis
Sequential
Analysis
Interactive
Psycho
R package
benchmark
R package
archetypes


Hyperparameter tuning
Sensitivity analysis
Robust
of
Benchmark experiments
Archetypal analysis
Framework
Hierarchical
Evolutional
Algorithm comparison
used for
Single data set-based
Domain-based
Analysis
Sequential
Analysis
Interactive
Psycho
R package
benchmark
R package
archetypes





R package
`archetypes`

Archetypal analysis for data analysis

Archetypal analysis (Cutler and Breiman, 1994) has the aim to find a few, not necessarily observed, extremal observations (the archetypes) in a multivariate data set such that

1. all the observations are approximated by convex combinations of the archetypes, and
2. all the archetypes are convex combinations of the observations.

Optimization problem:

Consider an $n \times m$ matrix X. For a 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

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

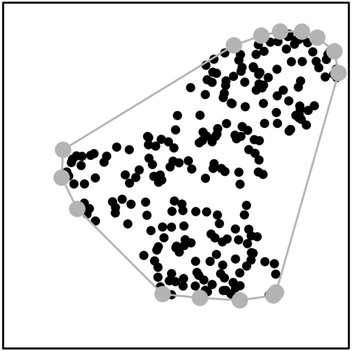
subject to the constraints

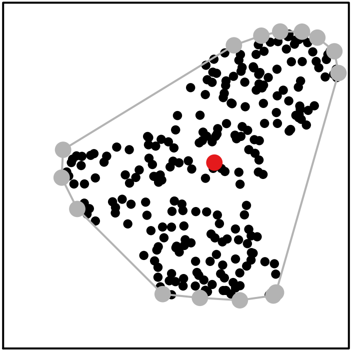
$$\sum_{j=1}^{k} \alpha_{ij} = 1 \text{ with } \alpha_{ij} \geq 0 \text{ and } i = 1, \dots, n,$$
$$\sum_{i=1}^{n} \beta_{ji} = 1 \text{ with } \beta_{ji} \geq 0 \text{ and } j = 1, \dots, k.$$

Approximation of the convex hull:

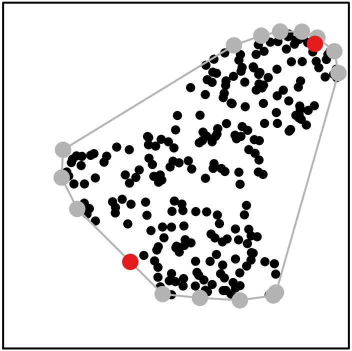
The archetypes lie on the boundary of the convex hull of the data; let N be the number of data points which define the boundary of the convex hull, then if ...

- $1 < k < N$, there are k archetypes on the boundary which minimize RSS;
- $k = N$, exactly the data points which define the convex hull are the archetypes with $\text{RSS} = 0$;
- $k = 1$, the sample mean minimizes the RSS.

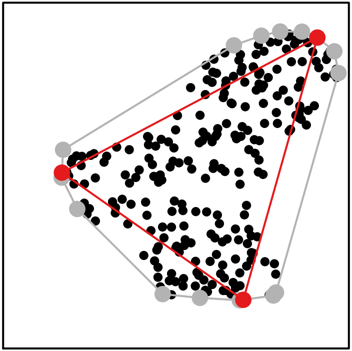




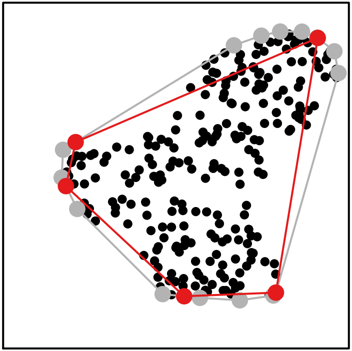
$k = 1$



$k = 2$



$k = 3$

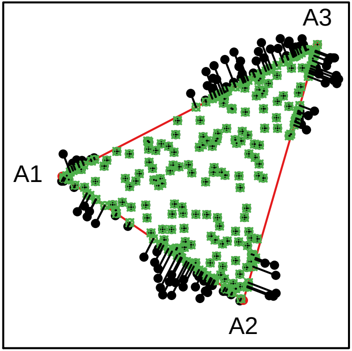


$k = 5$

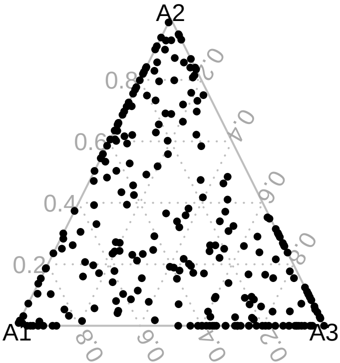
Approximation of the data:

Every observation can either be exactly represented or approximated by a convex combination of the archetypes.

Since $\alpha_{ij} \geq 0$ and $\sum_{j=1}^{k} \alpha_{ij} = 1$, the coefficients α_{ij} of an observation x_i can be interpreted as probabilities $p(x_i|z_j)$ indicating membership to classes represented by the archetypes.



Data approximation, $k = 3$



α coefficients, $k = 3$

Archetypal algorithm

Cutler and Breiman (1994) present an 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.

Archetypal algorithm

Cutler and Breiman (1994) present an 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.

- In each iteration $n + k$ convex least squares problems are solved (each of $O(n^2)$).

Archetypal algorithm

Cutler and Breiman (1994) present an 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.

- In each iteration $n + k$ convex least squares problems are solved (each of $O(n^2)$).
- The algorithm converges, but not necessarily against the global minimum; use different initializations.

Archetypal algorithm

Cutler and Breiman (1994) present an 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.

- In each iteration $n + k$ convex least squares problems are solved (each of $O(n^2)$).
- The algorithm converges, but not necessarily against the global minimum; use different initializations.
- There is no rule for the correct number of archetypes k; use the "elbow criterion" to find an appropriate value k.

Given the number of archetypes k:

1. Data preparation and initialization of archetypes Z, and coefficients α, β
2. Loop until RSS reduction is sufficiently small or the number of maximum iterations is reached:

   2.1 Find best α for the given set of archetypes Z: solve n convex least squares problems ($i = 1, \dots, n$)

   $$\min_{\alpha_i} \frac{1}{2}\|X_i - Z\alpha_i\|_2 \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_{j=1}^{k} \alpha_{ij} = 1.$$

   2.2 Recalculate archetypes $\tilde{Z}$: solve system of linear equations $X = \alpha\tilde{Z}^\top$.

   2.3 Find best β for the given set of archetypes $\tilde{Z}$: solve k convex least squares problems ($j = 1, \dots, k$)

   $$\min_{\beta_j} \frac{1}{2}\|\tilde{Z}_j - X\beta_j\|_2 \text{ subject to } \beta_j \geq 0 \text{ and } \sum_{i=1}^{n} \beta_{ji} = 1.$$

   2.4 Recalculate archetypes Z: $Z = X\beta$.

   2.5 Calculate residual sum of squares RSS.

3. Post-processing

Given the number of archetypes k:

1. Data preparation and initialization of archetypes Z, and coefficients α, β
2. Loop until RSS reduction is sufficiently small or the number of maximum iterations is reached:

   2.1 Find best α for the given set of archetypes Z: solve n convex least squares problems ($i = 1, \dots, n$)

   $$\min_{\alpha_i} \frac{1}{2} \|X_i - Z\alpha_i\|_2 \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_{j=1}^{k} \alpha_{ij} = 1.$$

   2.2 Recalculate archetypes $\tilde{Z}$: solve system of linear equations $X = \alpha \tilde{Z}^\top$.

   2.3 Find best β for the given set of archetypes $\tilde{Z}$: solve k convex least squares problems ($j = 1, \dots, k$)

   $$\min_{\beta_j} \frac{1}{2} \|\tilde{Z}_j - X\beta_j\|_2 \text{ subject to } \beta_j \geq 0 \text{ and } \sum_{i=1}^{n} \beta_{ji} = 1.$$

   2.4 Recalculate archetypes Z: $Z = X\beta$.

   2.5 Calculate residual sum of squares RSS.
3. Post-processing

Given the number of archetypes k:

1. Data preparation and initialization of archetypes Z, and coefficients α, β
2. Loop until RSS reduction is sufficiently small or the number of maximum iterations is reached:

   2.1 Reweight data: $X = w(\cdot)X$

   2.2 Find best α for the given set of archetypes Z: solve n convex least squares problems ($i = 1, \dots, n$)

   $$\min_{\alpha_i} \frac{1}{2}\|X_i - Z\alpha_i\|_2 \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_{j=1}^{k} \alpha_{ij} = 1.$$

   2.3 Recalculate archetypes $\tilde{Z}$: solve system of linear equations $X = \alpha\tilde{Z}^\top$.

   2.4 Find best β for the given set of archetypes $\tilde{Z}$: solve k convex least squares problems ($j = 1, \dots, k$)

   $$\min_{\beta_j} \frac{1}{2}\|\tilde{Z}_j - X\beta_j\|_2 \text{ subject to } \beta_j \geq 0 \text{ and } \sum_{i=1}^{n} \beta_{ji} = 1.$$

   2.5 Recalculate archetypes Z: $Z = X\beta$.

   2.6 Calculate residual sum of squares RSS.

3. Post-processing

Given the number of archetypes k:

1. Data preparation and initialization of archetypes Z, and coefficients α, β
2. Loop until RSS reduction is sufficiently small or the number of maximum iterations is reached:

   2.1 Reweight data: $X = w(\cdot)X$

   2.2 Find best α for the given set of archetypes Z: solve n convex least squares problems ($i = 1, \dots, n$)

   $$\min_{\alpha_i} \frac{1}{2} \|X_i - Z\alpha_i\|_2 \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_{j=1}^{k} \alpha_{ij} = 1.$$

   2.3 Recalculate archetypes $\tilde{Z}$: solve system of linear equations $X = \alpha \tilde{Z}^\top$.

   2.4 Find best β for the given set of archetypes $\tilde{Z}$: solve k convex least squares problems ($j = 1, \dots, k$)

   $$\min_{\beta_j} \frac{1}{2} \|\tilde{Z}_j - X\beta_j\|_2 \text{ subject to } \beta_j \geq 0 \text{ and } \sum_{i=1}^{n} \beta_{ji} = 1.$$

   2.5 Recalculate archetypes Z: $Z = X\beta$.

   2.6 Calculate residual sum of squares RSS.

3. Post-processing

Given the number of archetypes k:

1. **Data preparation** and **initialization** of archetypes Z, and coefficients α, β
2. Loop until **RSS** reduction is sufficiently small or the number of maximum iterations is reached:

   2.1 Reweight data: $X = w(\cdot)X$

   2.2 Find best α for the given set of archetypes Z: solve n convex least squares problems ($i = 1, \ldots, n$)

   $$\min_{\alpha_i} \frac{1}{2}\|X_i - Z\alpha_i\|_2 \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_{j=1}^{k} \alpha_{ij} = 1.$$

   2.3 Recalculate archetypes $\tilde{Z}$: solve system of line

   $X = \alpha\tilde{Z}^\top$.

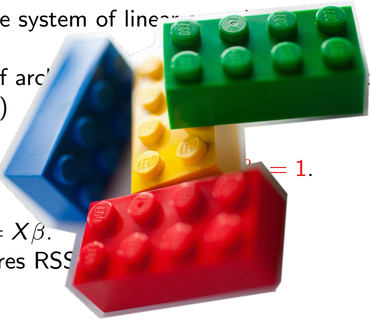
   2.4 Find best β for the given set of arc

   squares problems ($j = 1, \ldots, k$)

   $$\min_{\beta_j} \frac{1}{2}\|\tilde{Z}_j - X\beta_j\|_2 \text{ subje} \quad = 1.$$

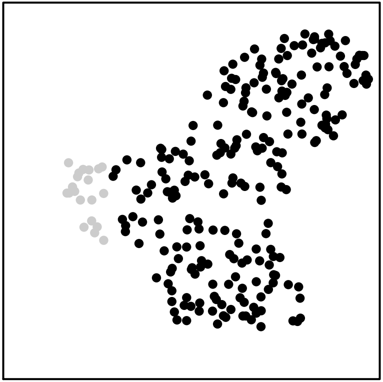
   2.5 Recalculate archetypes Z: $Z = X\beta$.

   2.6 Calculate residual sum of squares RSS



3. Post-processing

Weighted archetypes



Weighted archetypal problem:

Let W be a $n \times n$ diagonal matrix of weights for the observations. The weighted archetypal problem is then

$$\mathrm{RSS} = \|W(X - \alpha Z^\top)\|_2 \quad \rightarrow \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

Weighted archetypal problem:

Let W be a $n \times n$ diagonal matrix of weights for the observations. The weighted archetypal problem is then

$$\mathrm{RSS} = \|W(X - \alpha Z^\top)\|_2 \quad \rightarrow \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

Weighting the residuals is equivalent to weighting the data set:

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

Weighted archetypal problem:

Let W be a $n \times n$ diagonal matrix of weights for the observations. The weighted archetypal problem is then

$$\mathrm{RSS} = \|W(X - \alpha Z^\top)\|_2 \quad \to \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

Weighting the residuals is equivalent to weighting the data set:

$$W(X - \alpha Z^\top) = W(X - \alpha\beta^\top X) = \\ WX - (W\alpha)(\beta^\top W^{-1})(WX) = \tilde{X} - \tilde{\alpha}\tilde{Z}^\top$$

Therefore the problem can be reformulated as minimizing

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

Weighted archetypal problem:

Let W be a $n \times n$ diagonal matrix of weights for the observations. The weighted archetypal problem is then

$$\mathrm{RSS} = \| W(X - \alpha Z^\top) \|_2 \quad \to \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

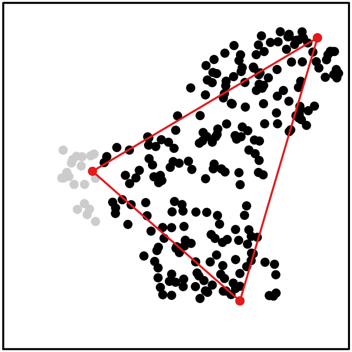
Weighting the residuals is equivalent to weighting the data set:

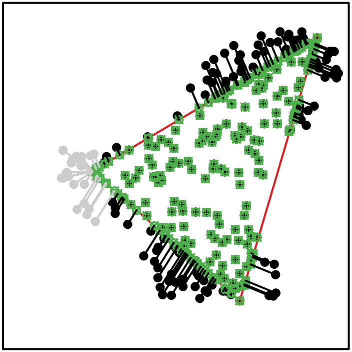
$$W(X - \alpha Z^\top) = W(X - \alpha \beta^\top X) = \\ WX - (W\alpha)(\beta^\top W^{-1})(WX) = \tilde{X} - \tilde{\alpha} \tilde{Z}^\top$$

Therefore the problem can be reformulated as minimizing

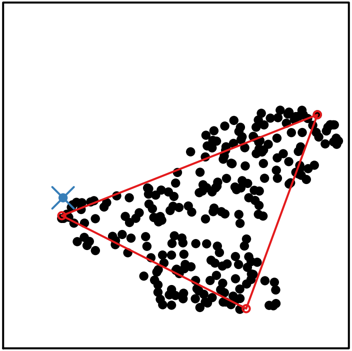
$$\mathrm{RSS} = \| \tilde{X} - \tilde{\alpha} \tilde{Z} \|_2 \text{ with } \tilde{Z} = \tilde{\beta} \tilde{X} \text{ and } \tilde{X} = WX.$$

**:** change data preparation (Step 1) and post-processing (Step 3)



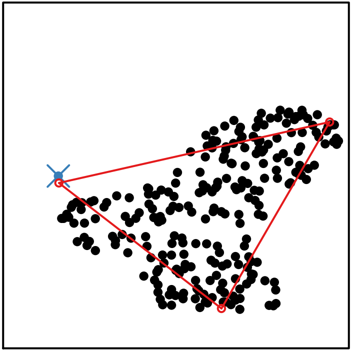


Robust archetypes



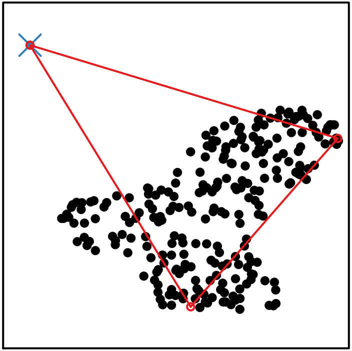
Breakdown point of 0%

Robust archetypes



Breakdown point of 0%

Robust archetypes



Breakdown point of 0%

M-estimator:

Let $R = (X - \alpha Z^\top)$ be the matrix of the residuals. In the standard archetypal problem,

$$\text{RSS} = \|R\|_2 \quad \rightarrow \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

have large residuals large effects, which privileges outliers.

M-estimator:

Let $R = (X - \alpha Z^\top)$ be the matrix of the residuals. In the standard archetypal problem,

$$\text{RSS} = \|R\|_2 \quad \rightarrow \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

have large residuals large effects, which privileges outliers.

M-estimators try to reduce the effect of outliers by replacing the squared residuals by another function $\rho(\cdot)$,

$$\rho(R) \quad \rightarrow \quad \min_{\alpha,\beta}, \qquad Z = X^\top \beta$$

which is less increasing than the square.

Iterated reweighted least squares problem:

The problem can be reformulated as an iterated reweighted least squares one: in the i-th iteration

$$\mathrm{R}^{i} = \|w(R^{(i-1)})R\|_2$$

is solved with $w(\cdot)$ a weight function depending on $\rho(\cdot)$ and the residuals of the $(i-1)$-th iteration.

Iterated reweighted least squares problem:

The problem can be reformulated as an iterated reweighted least squares one: in the i-th iteration

$$\mathrm{R}^i = \|w(R^{(i-1)})R\|_2$$

is solved with $w(\cdot)$ a weight function depending on $\rho(\cdot)$ and the residuals of the $(i-1)$-th iteration.

We use the generalized *Bisquare* function: Let R_j be the j-th row of R, and $w(R) = \mathrm{diag}(\tilde{w}(R_i))$

$$\tilde{w}(R_j) = \begin{cases} (1 - \|R_j/t\|_2^2)^2, & \|R_i\|_1 < t \\ 0, & \|R_i\|_1 \geq t \end{cases}$$

with tuning parameter t. Following Cleveland we use $t = 6s$ with s the median of the residuals unequal to zero.

Iterated reweighted least squares problem:

The problem can be reformulated as an iterated reweighted least squares one: in the i-th iteration

$$\mathrm{R}^i = \|w(R^{(i-1)})R\|_2$$

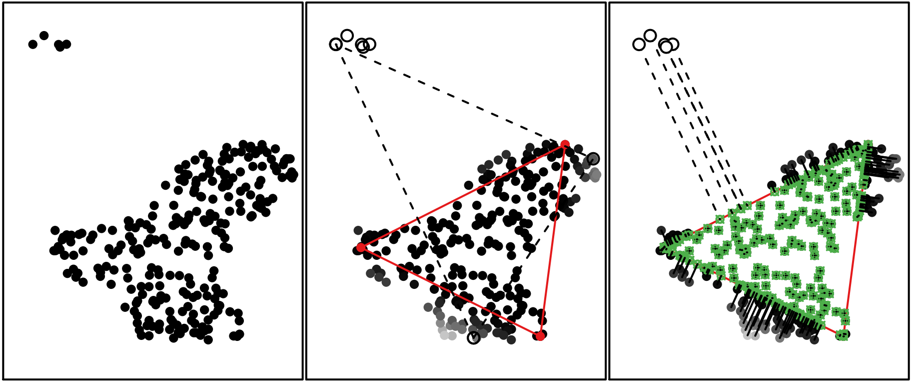
is solved with $w(\cdot)$ a weight function depending on $\rho(\cdot)$ and the residuals of the $(i-1)$-th iteration.

We use the generalized *Bisquare* function: Let R_j be the j-th row of R, and $w(R) = \mathrm{diag}(\tilde{w}(R_i))$

$$\tilde{w}(R_j) = \begin{cases} (1 - \|R_j/t\|_2^2)^2, & \|R_i\|_1 < t \\ 0, & \|R_i\|_1 \geq t \end{cases}$$

with tuning parameter t. Following Cleveland we use $t = 6s$ with s the median of the residuals unequal to zero.

**:** change data reweighting (Step 2.1) and post-processing (Step 3)



Implementation

```
> library("archetypes")
```

```
> archetypesFamily("original")  # or "weighted", "robust", etc.
...
 $ undummyfn   :function (x, zs, ...)
 $ initfn      :function (x, p, ...)
 $ alphasfn    :function (coefs, C, d, ...)
 $ betasfn     :function (coefs, C, d, ...)
 $ zalphasfn   :function (alphas, x, ...)
 $ globweightfn:function (x, weights)
 $ weightfn    :function (x, weights)
...
```

```
> archetypes(data, k, family = archetypesFamily("original"), ...)
> stepArchetypes(data, ..., k, nrep)
```

Application example

Sports analytics is "*the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision makers and enable them to help their organizations in gaining a competitive advantage on the field of play.*" (Alamar and Mehrotra, 2011)

In sports analytics, performance analysis, talent analytics, etc., usually the extreme performances are the most interesting aspect.

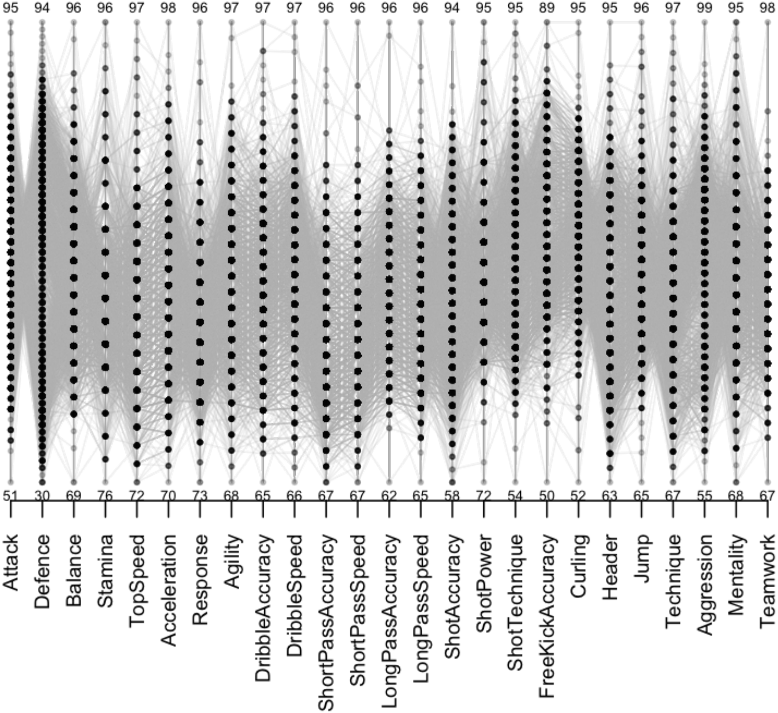
Archetypal soccer players:

Determine the archetypal soccer players and performance profiles of soccer player playing in four European top leagues.

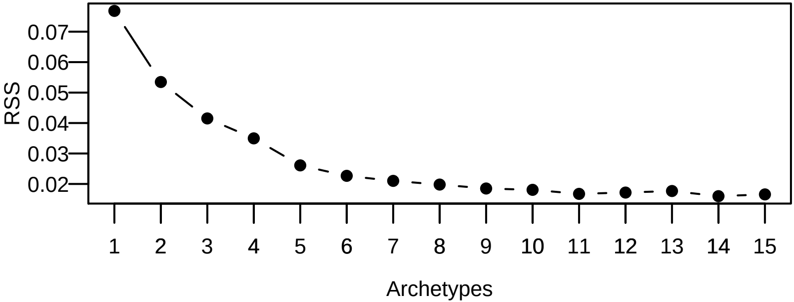
The skill ratings are from the PES Stats Database, a community-based approach to create a database with accurate statistics and skill ratings for soccer players (originally for the video game "Pro Evolution Soccer" by Konami).

The extracted data set consists of 25 skills of 1658 players (all positions—Defender, Midfielder, Forward—except Goalkeepers) from the German Bundesliga, the English Premier League, the Italian Serie A, and the Spanish La Liga.


PSD – PES STATS DATABASE
PSD – PES STATS DATABASE
pesstatsdatabase.com/PSD/Player_old.php?Id=602&Club=25
Leo MESSI
Club: FC Barcelona
Name: Leo MESSI
Shirt Name: MESSI
Number: 10
Positions: SS*, AMF
Nationality: Argentinian
Age: 24 (24/06/1987)
Height: 169 cm
Weight: 67 kg
Injury Tolerance: A
Foot: L
Side: B
Attack: 98
Defence: 44
Balance: 83
Stamina: 82
Top Speed: 82
ATT
SPD
TEC
POW
STA
DEF
PLAYER INDEX CARDS:
P05 – Mazing Run
P10 – Incisive Run
P18 – Talisman
S01 – Marauding
S02 – Passer
S03 – 1-on-1 Finish
zotero


95 94 96 96 97 98 96 97 97 97 96 96 96 96 94 95 95 89 95 95 96 97 99 95 98
51 30 69 76 72 70 73 68 65 66 67 67 62 65 58 72 54 50 52 63 65 67 55 68 67
Attack
Defence
Balance
Stamina
TopSpeed
Acceleration
Response
Agility
DribbleAccuracy
DribbleSpeed
ShortPassAccuracy
ShortPassSpeed
LongPassAccuracy
LongPassSpeed
ShotAccuracy
ShotPower
ShotTechnique
FreeKickAccuracy
Curling
Header
Jump
Technique
Aggression
Mentality
Teamwork

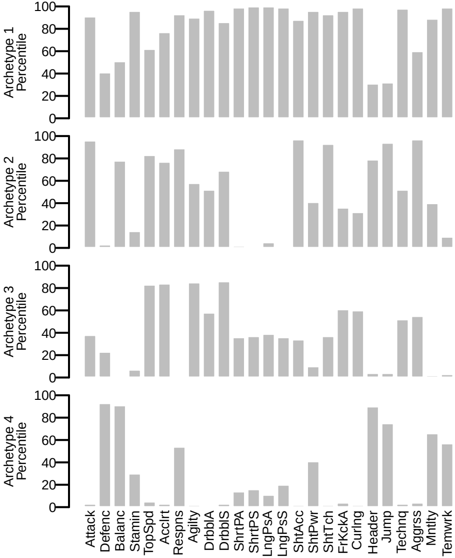
Model fitting and selection:



Model diagnostics:

Outliers, residuals, etc.

Model interpretation:



Archetypes:

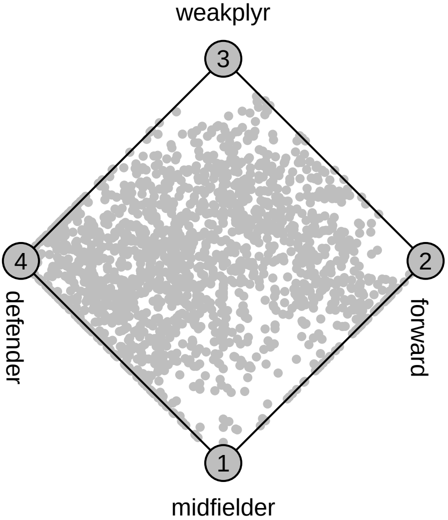
A1 is the archetypal midfielder with all skills high except the defense, balance, header, and jump.

A2 is the archetypal center forward with high skills in attack, shot, acceleration, header and jump, and low passing skills.

A3 is the archetypal weak player with high skills in running, but low skills in most ball related skills.

A4 is the archetypal defender with high skills in defense, balance, header, and jump.

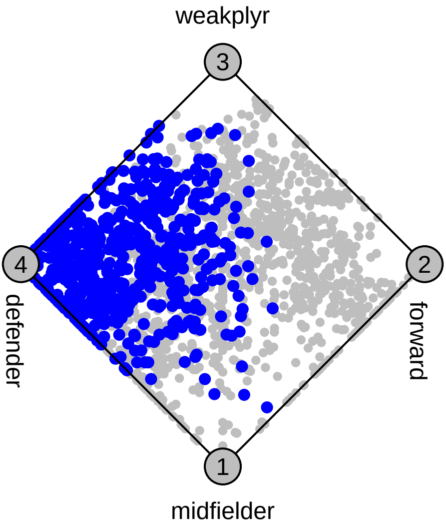
Model interpretation:



Archetypal soccer players:

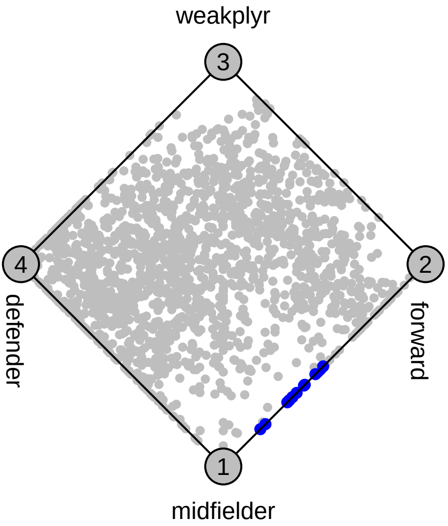
A1: Xavi Hernandez
A2: Filippo Inzaghi
A3: Florin Matei
A4: Massimo Paci

Model interpretation:



Talent scouting:
Find a new defender with offensive qualities.

Model interpretation:



Best soccer players:
A combination of Archetype 1 and Archetype 2 with Archetype 1 contributing more than Archetype 2.

1. Wayne Rooney
2. Leo Messi
3. Cristiano Ronaldo
4. Antonio Di Natale

...

Model deployment:

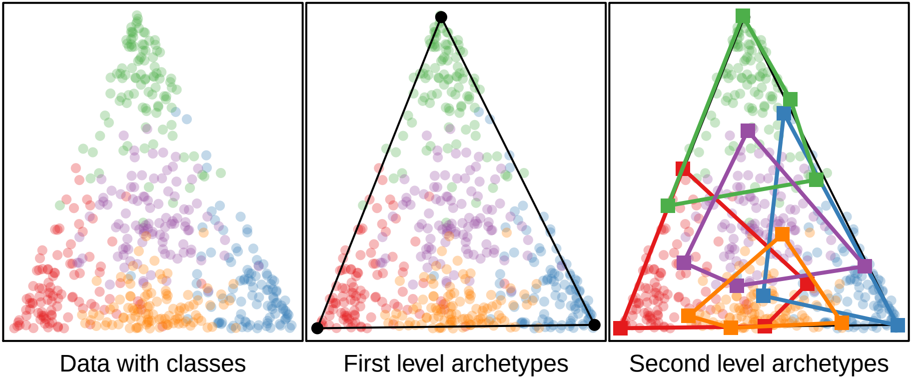
Enable end users to use the model without R; e.g., export the prediction model as standalone software component (as compiled C code).

```
$ sccratypes -i new_players.csv
     A1;      A2;      A3;      A4
0.42517; 0.57495; 0.00000; 0.00000
0.39257; 0.13046; 0.44863; 0.02833
0.08466; 0.01386; 0.04016; 0.86132
0.28616; 0.39076; 0.32318; 0.00000
0.00000; 0.30894; 0.62645; 0.06464
```

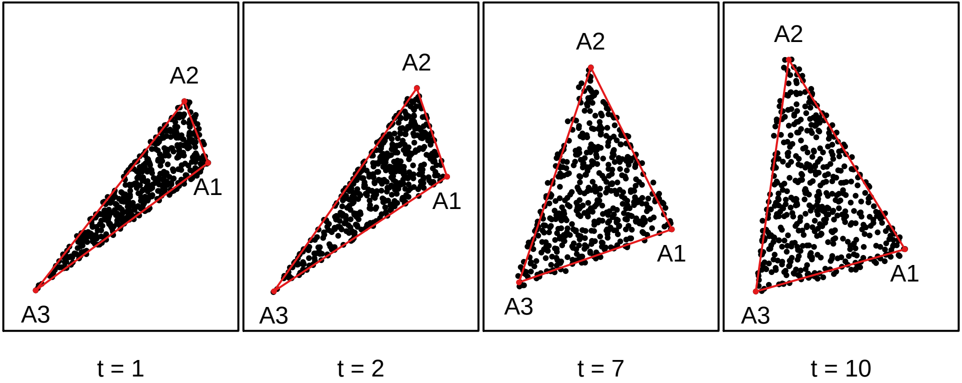
Outlook

Archetypal analysis is "*a promising unsupervised learning tool for many machine learning problems and as the representation is unique in general we believe the method holds particularly great promise for data mining application.*" (Mørup and Hansen, 2012)

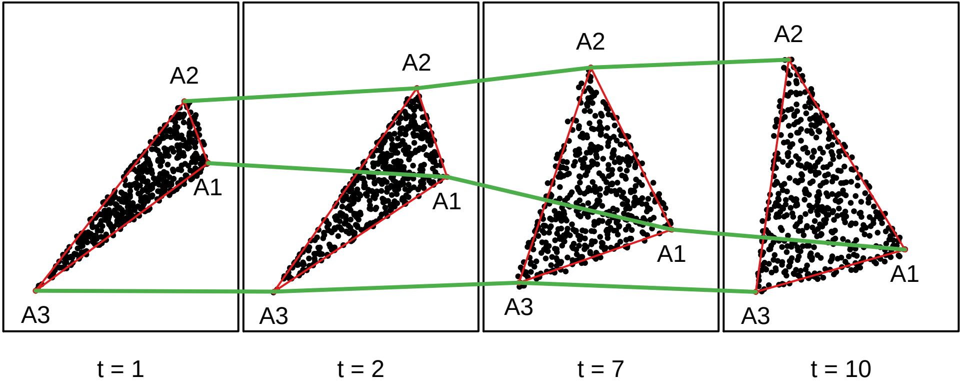
Hierarchical or multilevel archetypal analysis:



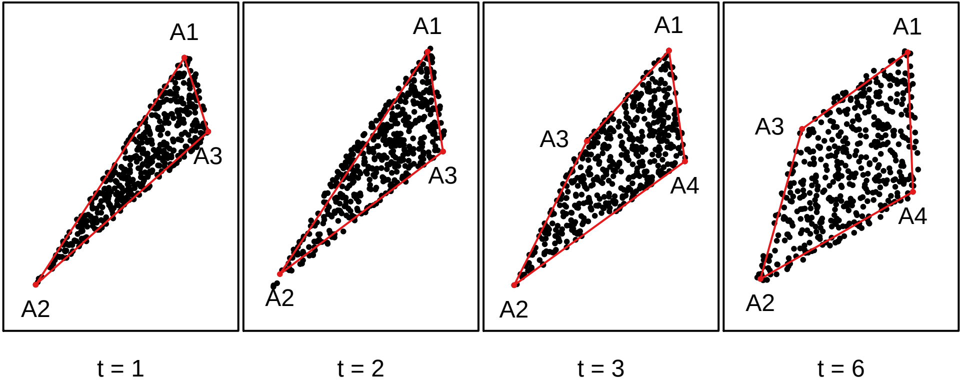
Evolutionary archetypal analysis:



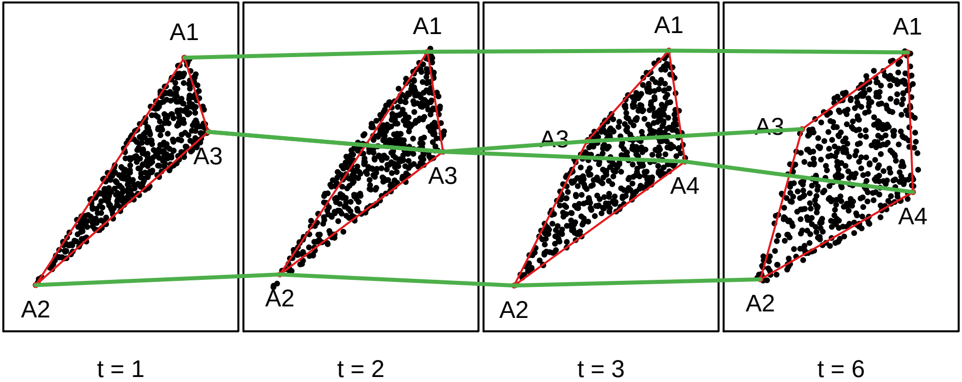
Evolutionary archetypal analysis:



Evolutionary archetypal analysis:



Evolutionary archetypal analysis:



Summary

- Archetypal analysis estimates data-driven (multivariate) extreme values which are easily interpretable by human experts.
- Application in a variety of areas, e.g., in performance analysis (economics), gene analysis (biology), and collaborative filtering (pattern recognition).
- Expertise from different areas, i.e., statistics, machine learning, and computer science.

`http://archetypes.r-forge.r-project.org/`

References

Benjamin Alamar and Vijay Mehrotra. Beyond 'moneyball': The rapidly evolving world of sports analytics, part I, September/October 2011. URL http://www.analytics-magazine.org/special-articles/391-beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i.

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

Manuel J. A. Eugster. *Benchmark Experiments – A Tool for Analyzing Statistical Learning Algorithms*. Dr. Hut-Verlag, 2011a. ISBN 9783868539103. URL http://edoc.ub.uni-muenchen.de/12990/. PhD thesis, Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany.

Manuel J. A. Eugster. *SportsAnalysis: Infrastructure for Sports Analysis*, 2011b. URL http://soccer.r-forge.r-project.org. R package.

Manuel J. A. Eugster. *archetypes: Archetypal Analysis*, 2011c. URL http://archetypes.r-forge.r-project.org. R package; denoted as the reference implementation by one of the original authors (see A. Cutler, Remembering Leo Breiman, The Annals of Applied Statistics, 4(4), 2010).

Manuel J. A. Eugster. Performance profiles based on archetypal athletes. *International Journal of Performance Analysis in Sport*, 12(1):166–187, 2012. URL http://www.ingentaconnect.com/content/uwic/ujpa/2012/00000012/00000001/art00016.

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. URL http://www.jstatsoft.org/v30/i08.

Manuel J. A. Eugster and Friedrich Leisch. Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis*, 55(3):1215–1225, 2011. doi: 10.1016/j.csda.2010.10.017.

Merriam-Webster Online Dictionary. Archetype. In *Merriam-Webster Online Dictionary*. 2008. Retrieved October 6, 2008, from http://www.merriam-webster.com/dictionary/archetype.

Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012. doi: 10.1016/j.neucom.2011.06.033. URL http://dx.doi.org/10.1016/j.neucom.2011.06.033.

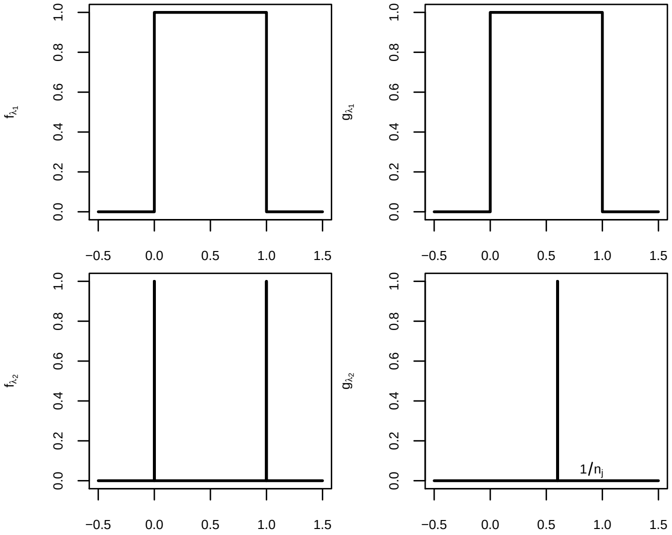
Prototypes problem I

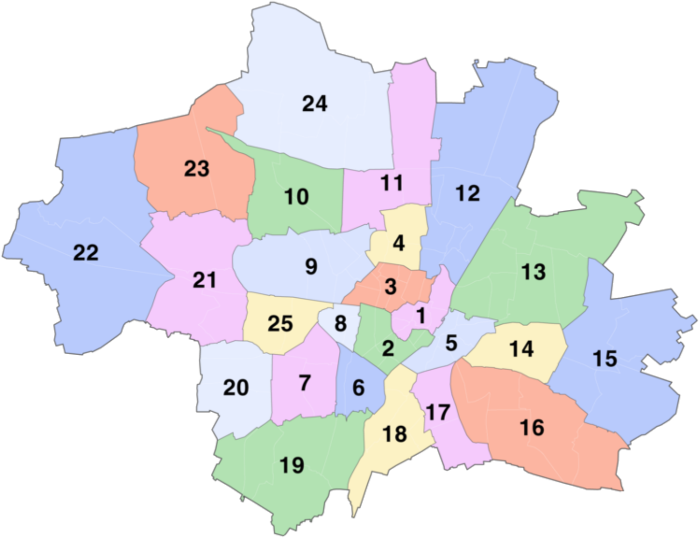
Given is an $n \times m$ matrix X representing a multivariate data set with n observations and m attributes. For a given k, the matrix Z of k m-dimensional prototypes are calculated according to:

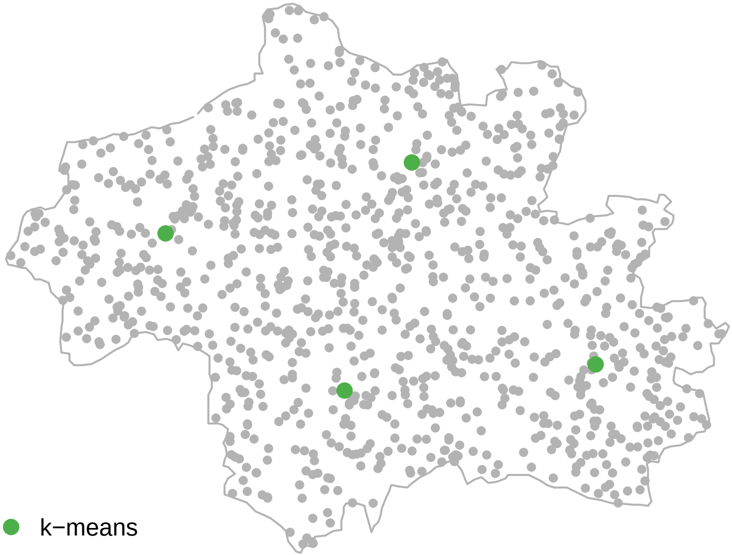
$$
\begin{aligned}
& w\|X - \alpha Z^T\| \rightarrow \min \\
& \text{with } \sum_{j=1}^{k} \alpha_{ij} = 1 \text{ and } f_\lambda(\alpha_{ij}) > 0 \\
& Z = X^T \beta \\
& \text{with } \sum_{i=1}^{n} \beta_{ij} = 1 \text{ and } g_\lambda(\beta_{ij}) > 0
\end{aligned}
$$

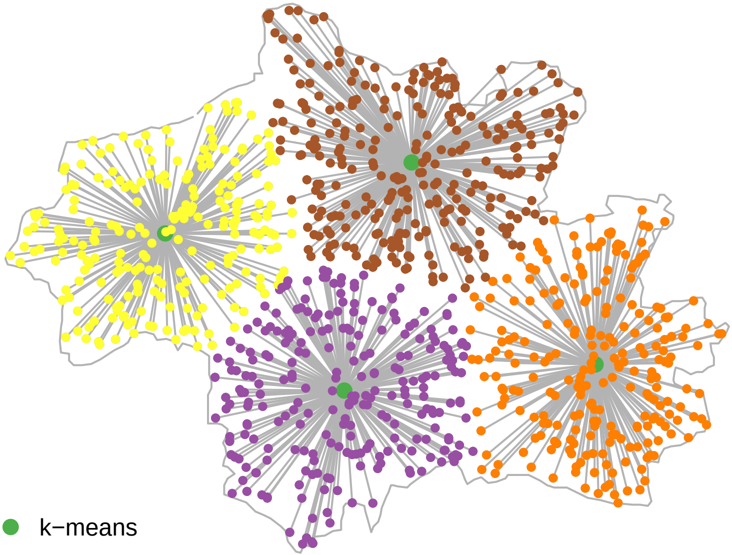
whereas $i = 1, \ldots, n$, $j = 1, \ldots, k$; α, the coefficients of the prototypes, is an $n \times k$ matrix; β, the coefficients of the observations, is an $n \times k$ matrix; w, the weighting vector, is an n vector; $\| \cdot \|$ is an appropriate matrix norm; f and g are evaluation functions of the coefficients for a specific constraint controlled by the parameter λ.

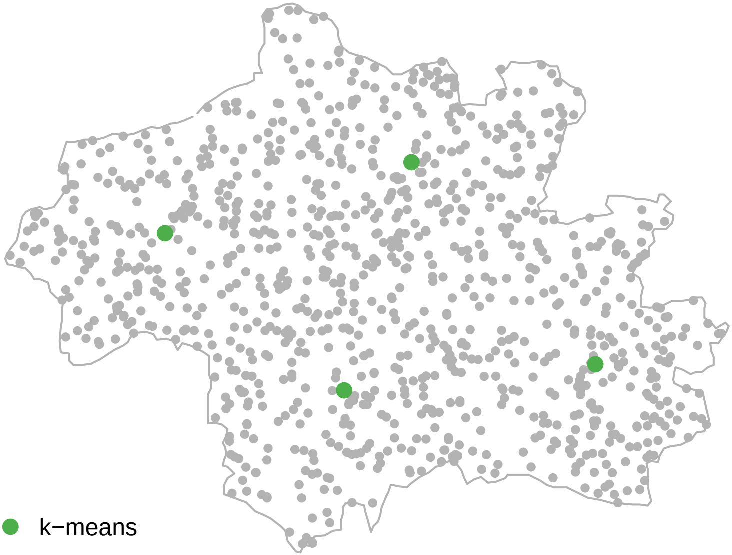
Prototypes problem II

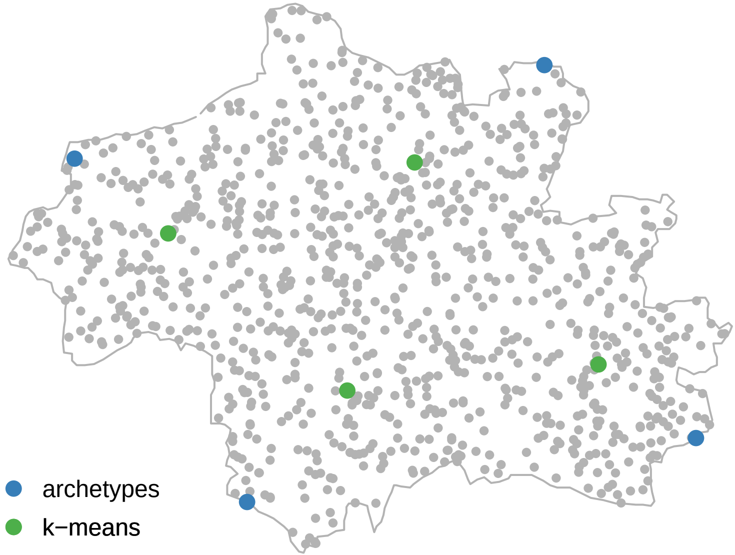


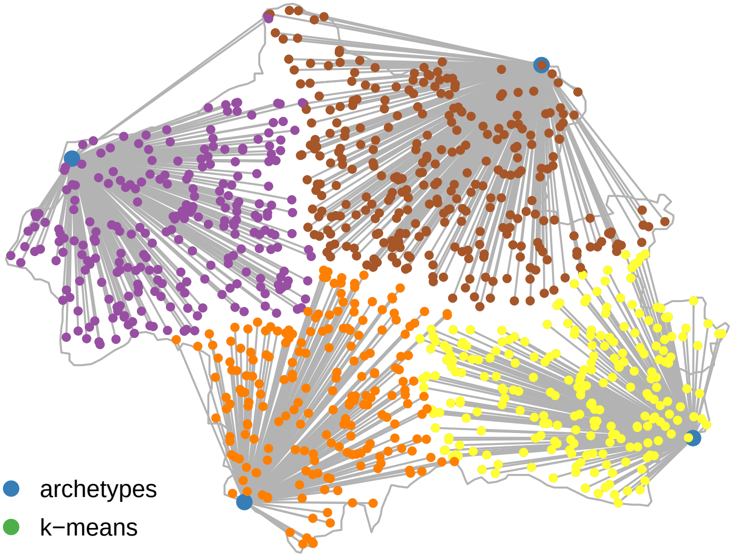

24
23
22
10
11
12
4
9
21
3
13
1
25
8
2
5
14
15
20
7
6
17
16
18
19


k−means


k−means


k−means


archetypes
k−means


archetypes
k−means