Making Archetypal Analysis Practical

Christian Thurau // christian.thurau@unbelievable-machine.com
February 11, 2015

Table of Contents

k-Means Clustering

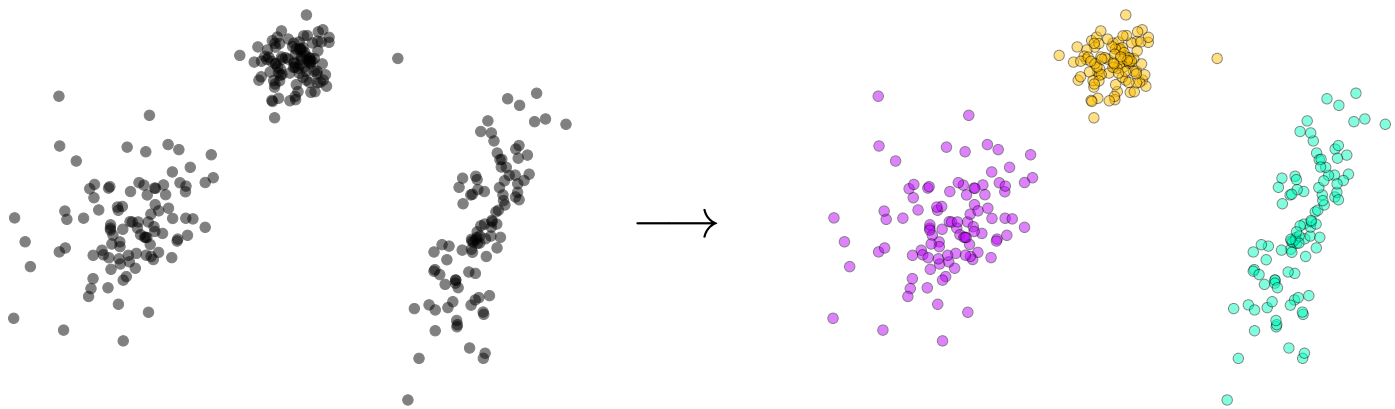
The average human has one breast and one testicle.

Des MacHale

k-Means Clustering

setting:

- assume a data set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^m$
- determine k clusters $C_i \subset X$ such that
  - elements assigned to C_i are *similar*
  - $C_i \cap C_j = \emptyset$ for $i \neq j$



k-Means Clustering

approach:

- define *similarity* and *clusters* in terms of centroids $\boldsymbol{\mu}_i \in \mathbb{R}^m$

$$C_i = \left\{ \boldsymbol{x}_j \in X \middle| \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2 \leq \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_l \right\|^2 \forall\, l \neq i \right\}$$

- determine suitable centroids
- consider objective function

$$E(k) = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in C_i} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2$$

k-Means Clustering

algorithm:

at $t = 0$, initialize $\boldsymbol{\mu}_1^{(t)}, \ldots, \boldsymbol{\mu}_k^{(t)}$

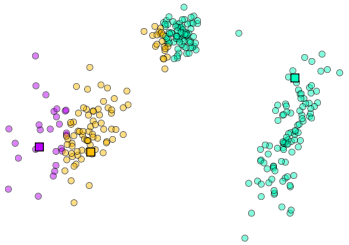
repeat until convergence

for each $\boldsymbol{\mu}_i^{(t)}$ compute $C_i^{(t)}$, $n_i = |C_i|$

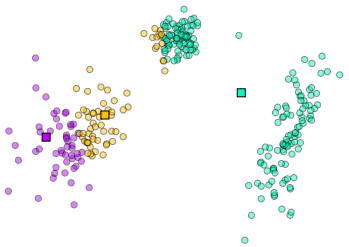
for each $C_i^{(t)}$ update the centroid as

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{n_i} \sum_{\boldsymbol{x}_j \in C_i^{(t)}} \boldsymbol{x}_j$$

$t \leftarrow t + 1$

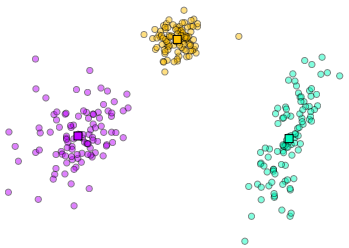


$t = 0$



$t = 1$

$\vdots$



$t = 5$

k-Means Clustering

note:

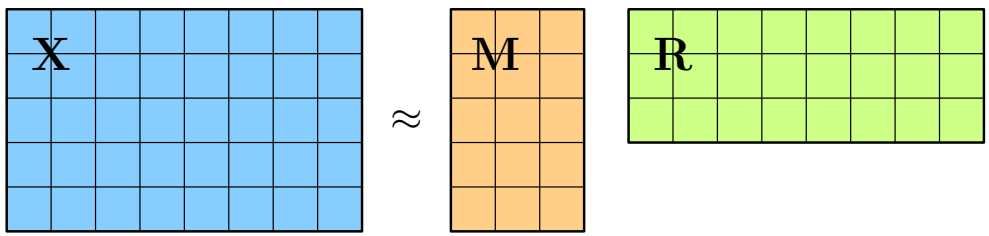
- k-means clustering is a matrix factorization problem

k-Means Clustering

note:

- k-means clustering is a matrix factorization problem
- upon convergence of the k-means algorithm, we have

$$\boldsymbol{X} \approx \boldsymbol{MR}$$
$$\Leftrightarrow \quad \boldsymbol{x}_j \approx \boldsymbol{M}\boldsymbol{r}_j$$



where

$\boldsymbol{X} \in \mathbb{R}^{m \times n} \Leftrightarrow$ data matrix

$\boldsymbol{M} \in \mathbb{R}^{m \times k} \Leftrightarrow$ centroid matrix

$\boldsymbol{R} \in \mathbb{R}^{k \times n} \Leftrightarrow$ indicator matrix

k-Means Clustering

note:

- k-means clustering is solving

$$\min_{\boldsymbol{M},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{MR} \right\|^2$$

under several *implicit constraints*

k-Means Clustering

note:

- k-means clustering is solving

$$\min_{\boldsymbol{M},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{M}\boldsymbol{R} \right\|^2$$

under several *implicit constraints*

- every centroid is a convex combination of data points

$$\boldsymbol{\mu}_i = \boldsymbol{X}\boldsymbol{s}_i \quad \Leftrightarrow \quad \boldsymbol{M} = \boldsymbol{X}\boldsymbol{S}$$

- $\boldsymbol{S}$ is column stochastic ($\boldsymbol{s}_i$ contains 0s and n_i values of $\frac{1}{n_i}$)
- $\boldsymbol{R}$ is column stochastic and binary ($\boldsymbol{r}_j$ contains a single 1)

k-Means Clustering

constraints on S:

$$s_{ji} \geq 0 \;\Leftrightarrow\; \boldsymbol{S} \succeq \mathbf{0}$$

$$\sum_{j=1}^{n} s_{ji} = 1 \;\Leftrightarrow\; \mathbf{1}^T \boldsymbol{s}_i = 1$$

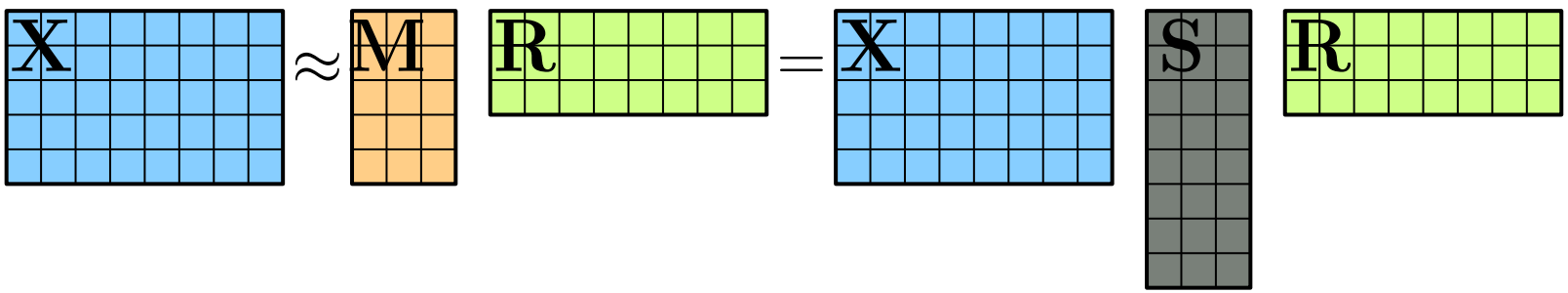
$$s_{ji} \in \left\{0, \frac{1}{n_i}\right\} \;\Leftrightarrow H(\boldsymbol{s}_i) = -\sum_{j=1}^{n} s_{ji} \log s_{ji} \gg 0$$

constraints on R:

$$r_{ij} \geq 0 \;\Leftrightarrow\; \boldsymbol{R} \succeq \mathbf{0}$$

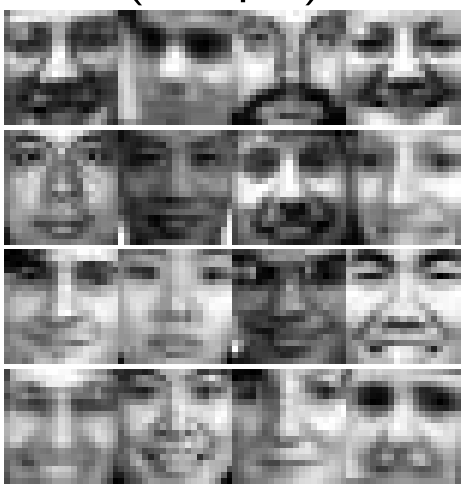
$$\sum_{i=1}^{k} r_{ij} = 1 \;\Leftrightarrow\; \mathbf{1}^T \boldsymbol{r}_j = 1$$

$$r_{ij} \in \{0, 1\} \;\Leftrightarrow H(\boldsymbol{r}_j) = -\sum_{i=1}^{k} r_{ij} \log r_{ij} = 0$$



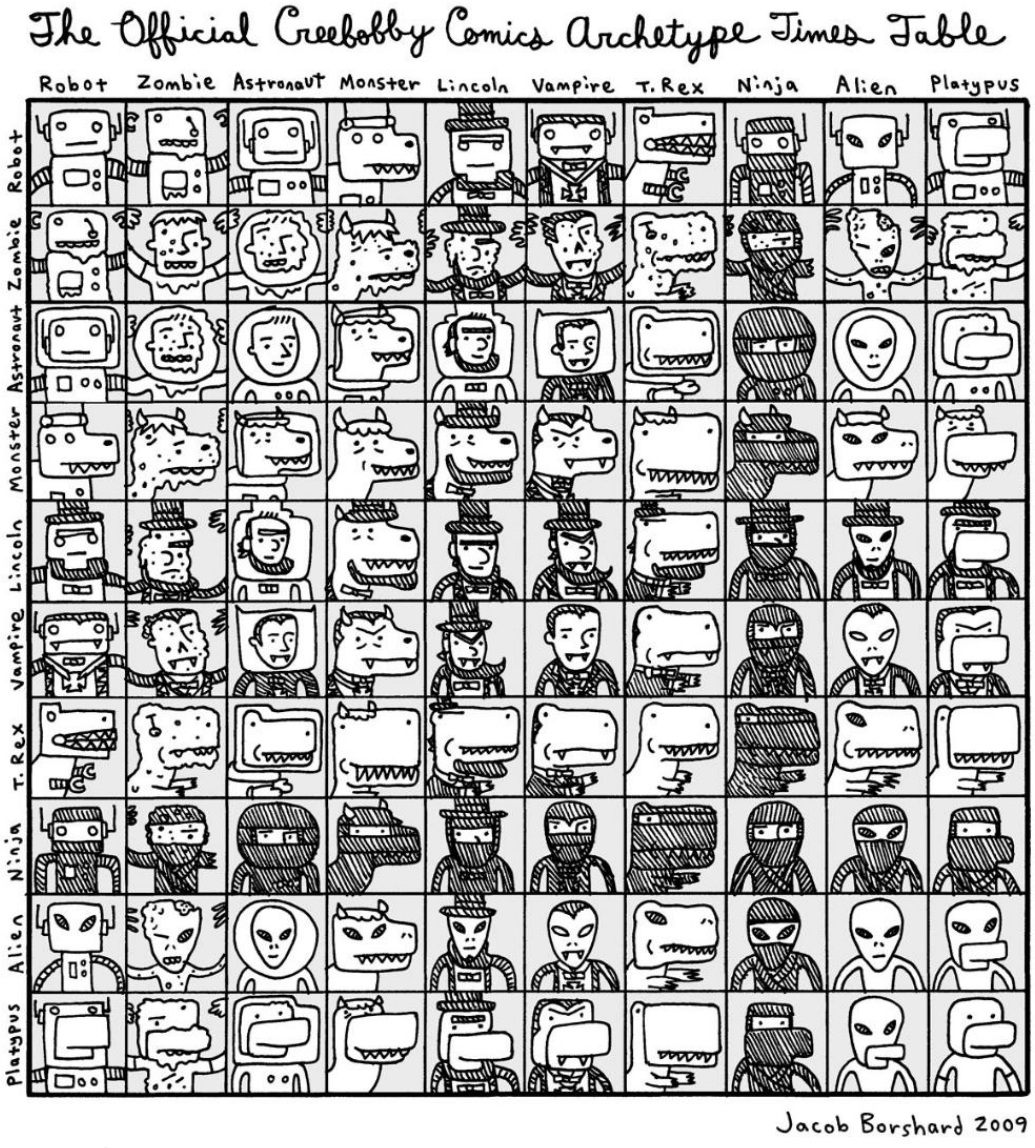
k-Means Clustering vs. archetypal analysis

data (sample)



k-means

AA



Archetypes

- **Plato**:

  ideals and/or pure forms that embody fundamental characteristics of a thing rather than its specific peculiarities

- **C.G. Jung**:

  innate, universal forms (the *hero*, the *great mother*, the *wise old man*, ...) that channel experiences and emotions, resulting in recognizable and typical behaviors with certain probable outcomes

- **A. Cutler** and **L. Breiman** (in Technometrics 36(4), 1994):

  archetypal analysis $\Leftrightarrow$ new way of data analysis for multivariate data

Archetypal Analysis what?

setting:

- assume a data matrix

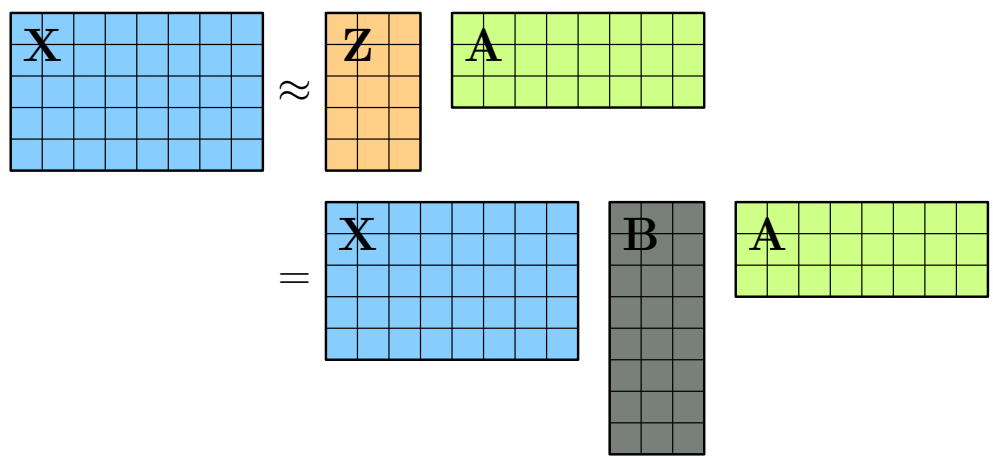
$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

- choose an integer $k \leq \min\{m, n\}$
- determine two column stochastic factor matrices

$$\boldsymbol{B} \in \mathbb{R}^{n \times k}$$
$$\boldsymbol{A} \in \mathbb{R}^{k \times n}$$

such that

$$\boldsymbol{X} \approx \boldsymbol{ZA}$$
$$= \boldsymbol{XBA}$$

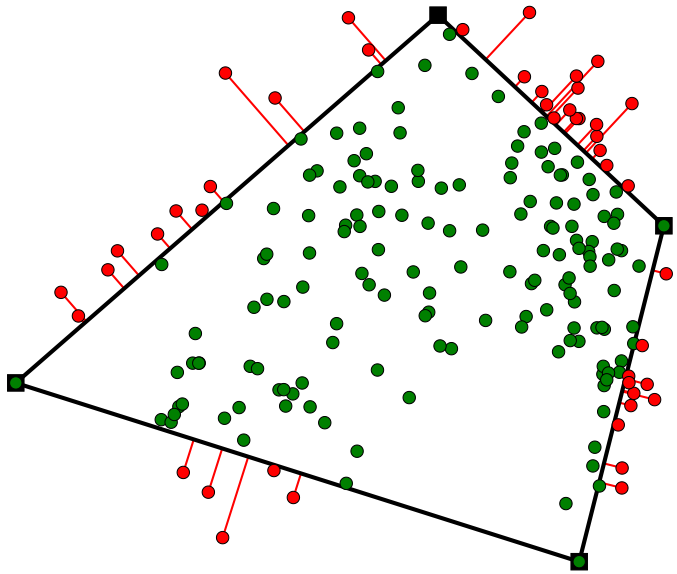


Archetypal Analysis what?

approach:

- solve

$$\min_{\boldsymbol{B},\boldsymbol{A}} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A} \right\|^2$$
$$\text{s.t.} \quad \boldsymbol{B} \succeq 0,\ \mathbf{1}^T \boldsymbol{b}_i = 1$$
$$\boldsymbol{A} \succeq 0,\ \mathbf{1}^T \boldsymbol{a}_j = 1$$



observe:

- substituting $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} \in \mathbb{R}^{m \times k}$ yields

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \boldsymbol{x}_j \, b_{ji} \qquad \text{and} \qquad \boldsymbol{x}_j \approx \sum_{i=1}^{k} \boldsymbol{z}_i \, a_{ij}$$

Archetypal Analysis
why?

therefore:

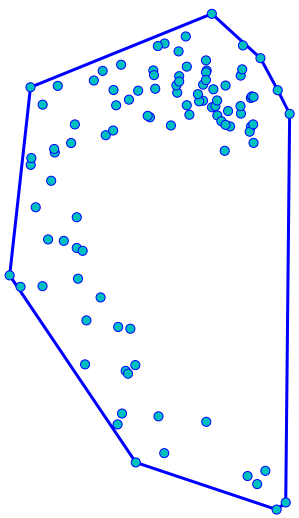
- the archetypes z_i are sparse convex combinations of the data x_j
- the data points x_j are sparse convex combinations of the archetypes z_i

The four basic personality types

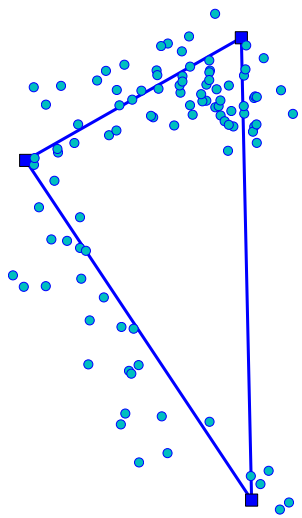
Archetypal Analysis
why?

properties:

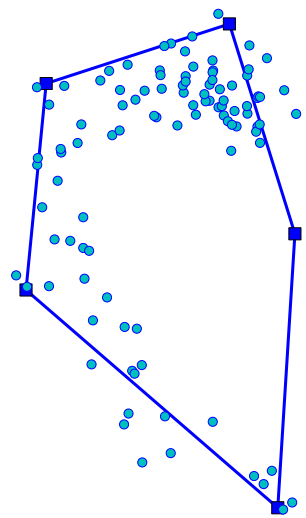
- for $k > 1$, archetypes reside on the data convex hull
- for increasing k, the archetypal hull approximates the data hull



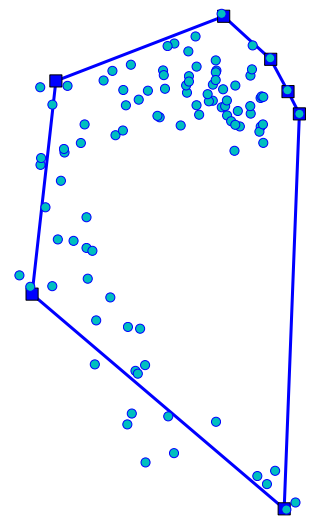
convex hull



$k = 3$



$k = 5$



$k = 7$

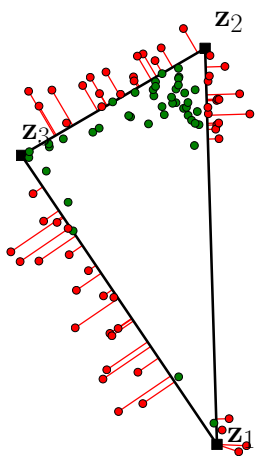
Archetypal Analysis
why?

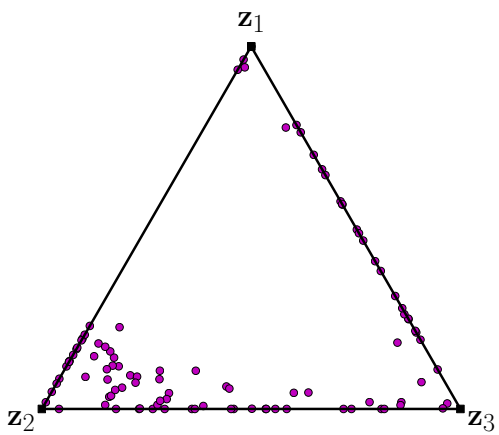
properties:

- the coefficient vectors $\boldsymbol{a}_j$ in $\boldsymbol{x}_j = \boldsymbol{Z}\,\boldsymbol{a}_j$ are *stochastic*

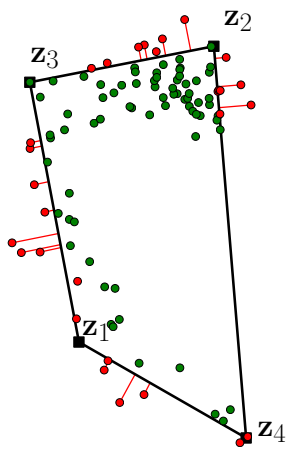
$\Rightarrow$ the coefficients a_{ij} can be thought of as $p(\boldsymbol{x}_j \mid \boldsymbol{z}_i)$

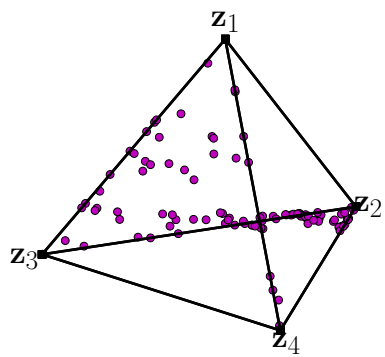
$\Rightarrow$ (soft)clustering, classification, ranking, ...





$k = 3$



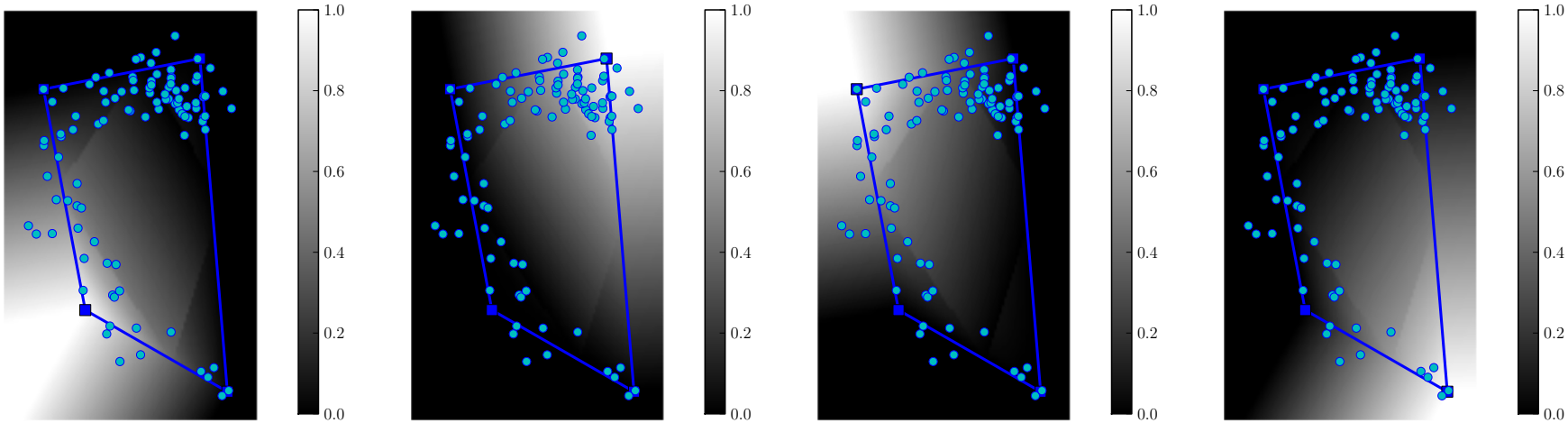


$k = 4$

Archetypal Analysis
why?

properties:

- archetypal analysis bridges geometry and probability



$p(\boldsymbol{x} \mid \boldsymbol{z}_1)$ $p(\boldsymbol{x} \mid \boldsymbol{z}_2)$ $p(\boldsymbol{x} \mid \boldsymbol{z}_3)$ $p(\boldsymbol{x} \mid \boldsymbol{z}_4)$

Archetypal Analysis how?

alternating least squares:

randomly initialize $\boldsymbol{B}$

repeat until convergence

compute $\boldsymbol{Z} = \boldsymbol{XB}$

solve under constraints
$\boldsymbol{A} = \text{argmin}_{\boldsymbol{A}} \left\| \boldsymbol{X} - \boldsymbol{ZA} \right\|^2$

compute $\hat{Z} = \boldsymbol{XA}^{\dagger}$

solve under constraints
$\boldsymbol{B} = \text{argmin}_{\boldsymbol{B}} \left\| \hat{\boldsymbol{Z}} - \boldsymbol{XB} \right\|^2$

projected gradients:

$$\left\| \boldsymbol{X} - \boldsymbol{XBA} \right\|^2 = \text{tr} \left[\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{X}^T\boldsymbol{XBA} + \boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{XBA} \right]$$

$$\frac{\partial E}{\partial \boldsymbol{A}} = 2 \left[\boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{XBA} - \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{X} \right]$$

$$\frac{\partial E}{\partial \boldsymbol{B}} = 2 \left[\boldsymbol{X}^T\boldsymbol{XBAA}^T - \boldsymbol{X}^T\boldsymbol{XA}^T \right]$$

randomly initialize $\boldsymbol{A}$ and $\boldsymbol{B}$

repeat until convergence

$\boldsymbol{A} \leftarrow \boldsymbol{A} - \eta_A \frac{\partial E}{\partial \boldsymbol{A}}$, project $\boldsymbol{a}_j$ onto Δ^{k-1}

$\boldsymbol{B} \leftarrow \boldsymbol{B} - \eta_B \frac{\partial E}{\partial \boldsymbol{B}}$, project $\boldsymbol{b}_i$ onto Δ^{n-1}

Archetypal Analysis
how?

note:

- archetypal analysis makes no implicit density assumption
- archetypal analysis is NP-hard
- both algorithms are heuristics
- both (implicitly) depend on $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{n \times n}$

  $\Rightarrow$ they scale with at least $O(n^2)$

$\Rightarrow$ both do not apply to BIG DATA

Archetypal Analysis
how?

three key observations:

1. $\boldsymbol{Z} = \boldsymbol{XB}$ comprises information as to *extreme* data points
2. extreme data points are surprisingly easy to determine
3. if $\boldsymbol{Z}$ were known, then

$$\left\|\boldsymbol{X} - \boldsymbol{ZA}\right\|^2 = \left\|\boldsymbol{x}_1 - \boldsymbol{Z}\boldsymbol{a}_1\right\|^2 + \left\|\boldsymbol{x}_2 - \boldsymbol{Z}\boldsymbol{a}_2\right\|^2 + \dots$$

Archetypal Analysis
how?

three key observations:

1. $\boldsymbol{Z} = \boldsymbol{XB}$ comprises information as to *extreme* data points
2. extreme data points are surprisingly easy to determine
3. if $\boldsymbol{Z}$ were known, then

$$\left\|\boldsymbol{X} - \boldsymbol{ZA}\right\|^2 = \left\|\boldsymbol{x}_1 - \boldsymbol{Z}\boldsymbol{a}_1\right\|^2 + \left\|\boldsymbol{x}_2 - \boldsymbol{Z}\boldsymbol{a}_2\right\|^2 + \dots$$

two key ideas:

1. perform AA on an appropriate subset
2. decouple the computation of $\boldsymbol{Z}$ and $\boldsymbol{A}$

Making Archetypal Analysis practical how?

next: brief summary of

- C. Bauckhage and C. Thurau: Making Archetypal Analysis Practical, DAGM, 2009.
- C. Thurau, K. Kersting, and C. Bauckhage: Yes we can: simplex volume maximization for descriptive web-scale matrix factorization. ACM CIKM, 2010.
- C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage: Convex non-negative matrix factorization for massive datasets. Knowledge and Information Systems, 29(2):457-478, 2011.
- K. Kersting, C. Bauckhage, C. Thurau, and M. Wahabzada: Matrix Factorization as Search. ECML/PKDD, 2012.

Making Archetypal Analysis practical
how?

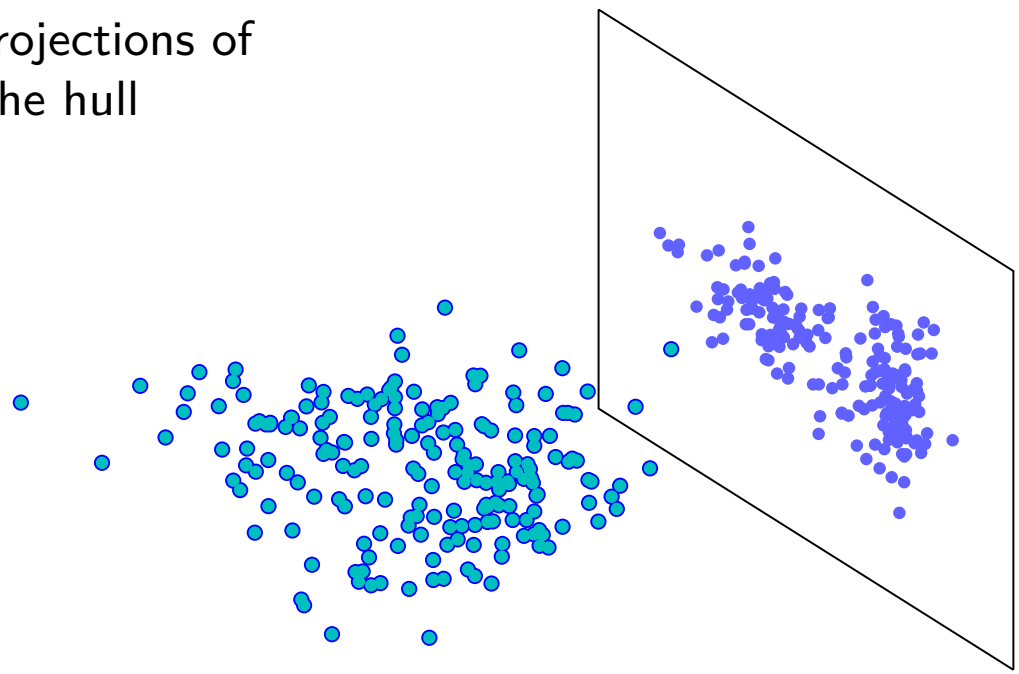
sampling the convex hull:

- archetypes are mixtures of points on the data convex hull

$\Rightarrow$ restrict algorithm to $X^H \subseteq X$

- in $\mathbb{R}^m$, convex hull computation is "expensive" $(O(n^{\lfloor m/2 \rfloor + 1}))$

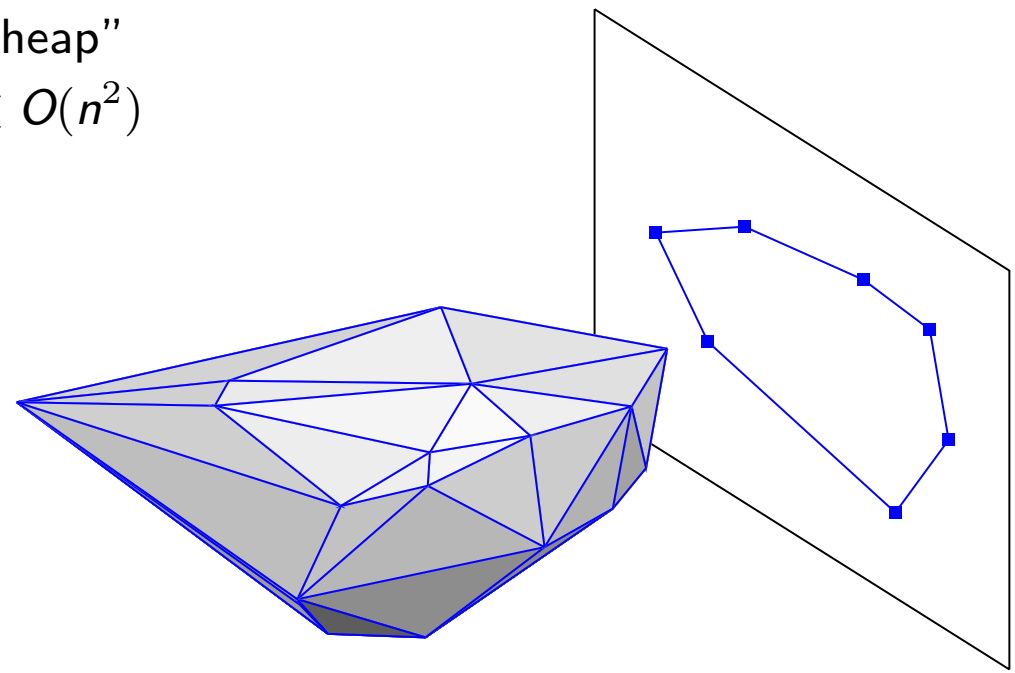
$\Rightarrow$ consider (many) 2D projections of the data and sample the hull



Making Archetypal Analysis practical how?

sampling the convex hull:

- if P is a polytope and $\pi : \boldsymbol{x} \rightarrow \boldsymbol{M}\boldsymbol{x} + \boldsymbol{t}$ is an affine map, then $\pi(P)$ is a polytope
- every vertex of $\pi(P)$ corresponds to a vertex of P
- sampling the hull is "cheap"
- effort is then $O(\eta^2) \ll O(n^2)$
- η is $\Omega(\sqrt{\log n})$



Making Archetypal Analysis practical how?

using distance geometry:

- given the distances d_{ij} between k vertices of an $k-1$ simplex S, then

$$V(S)^2 = \frac{-1^k}{2^{k-1}\left((k-1)!\right)^2} \det(\boldsymbol{A})$$

where $\det(\boldsymbol{A})$ is the CM determinant

Cayley-Menger determinant

$$\begin{vmatrix} 0 & 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & d_{11}^2 & d_{12}^2 & \dots & d_{1k}^2 \\ 1 & d_{11}^2 & 0 & d_{22}^2 & \dots & d_{2k}^2 \\ 1 & d_{12}^2 & d_{22}^2 & 0 & \dots & d_{3k}^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{1k}^2 & d_{2k}^2 & d_{3k}^2 & \dots & 0 \end{vmatrix}$$

Making Archetypal Analysis practical
how?

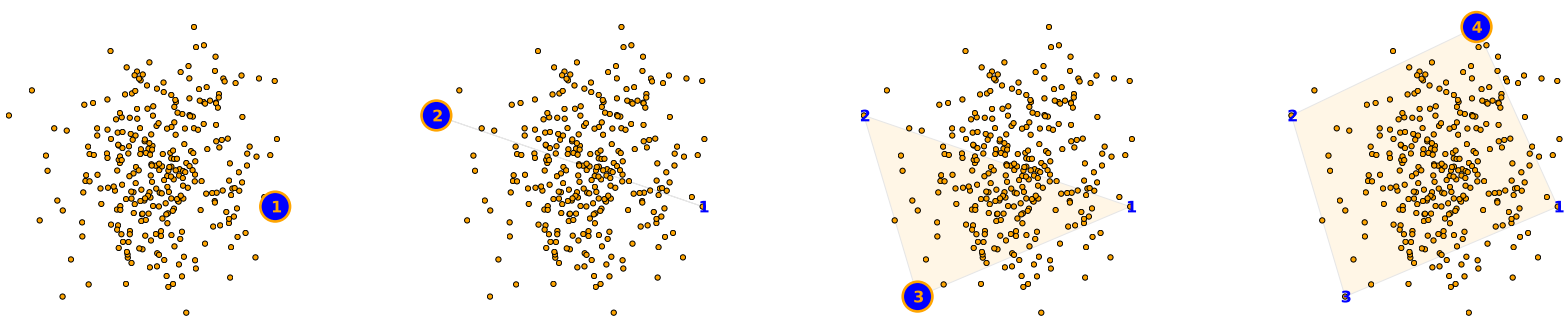
simplex volume maximization (SiVM):

- determine $\boldsymbol{Z}$ by fitting a simplex of maximal volume into the data

$\Rightarrow$ fast, greedy, iterative algorithm

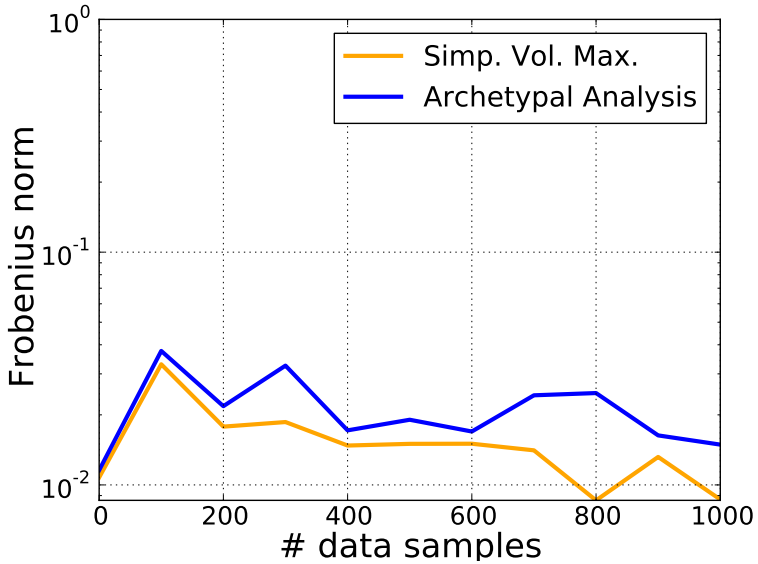
- given S with $k-1$ vertices, determine next vertex $\boldsymbol{z}_k \in \boldsymbol{X}$ s.t.

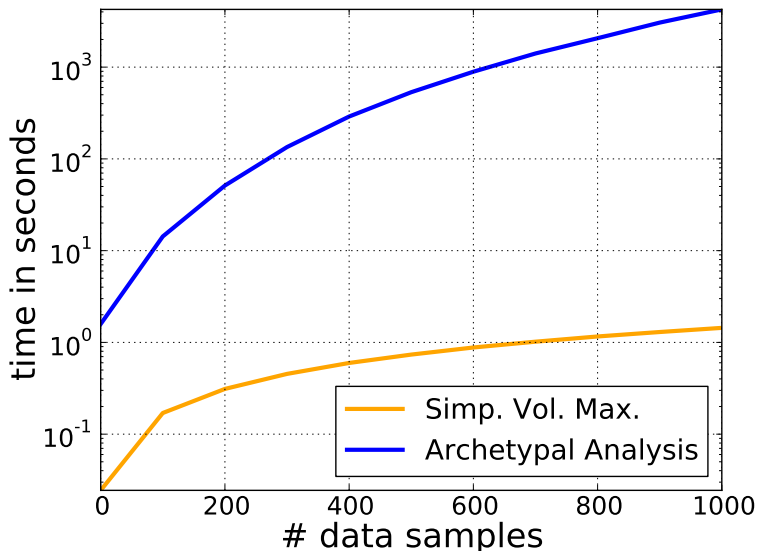
$$\boldsymbol{z}_k = \underset{j}{\text{argmax}}\ V(S \cup \boldsymbol{x}_j)^2$$



Making Archetypal Analysis practical how?

simplex volume maximization (SiVM):





Applications

Applications
mining World of Warcraft

idea / motivation:

- automatically characterize MMORPG teams / guilds
- understand group behavior in MMORPGs
- simplify game design process

Applications
mining World of Warcraft

data:

- 150 million in-game observations of
  - 18 million characters
  - in 1.4 million guilds
- observations reveal a player's experience level (10-80)
- for each guild, determine a 70 dim. experience histogram

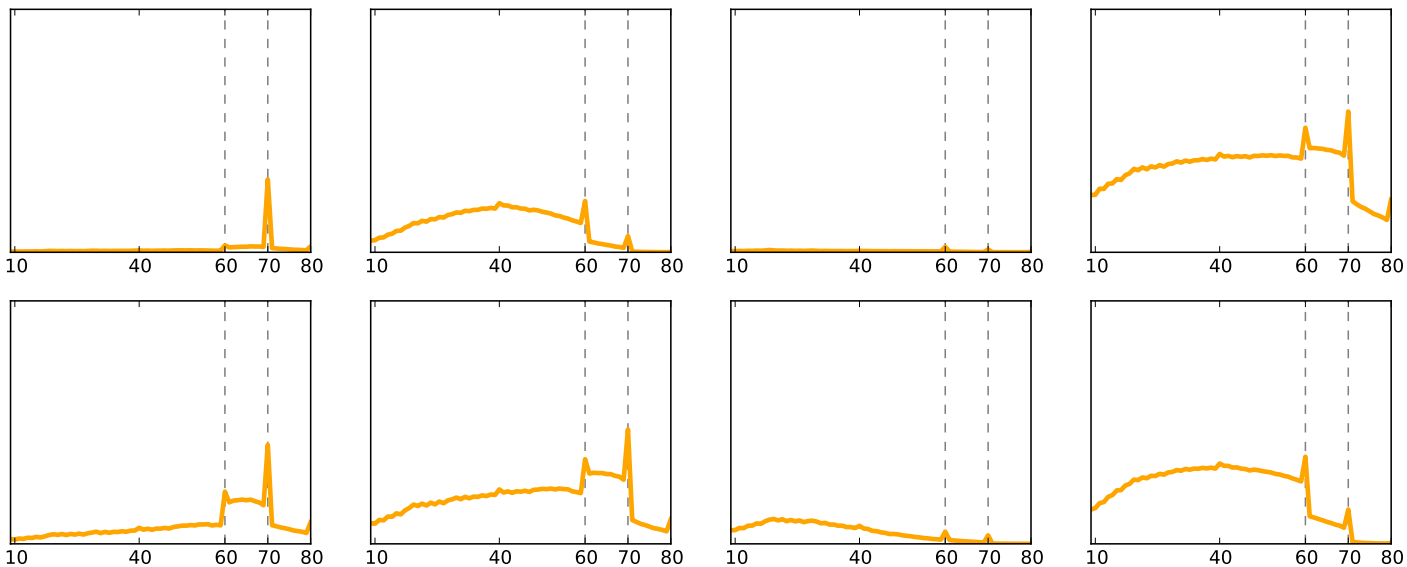
$\Rightarrow$ data matrix of 98 million entries

details in:

- C. Thurau and C. Bauckhage: Analyzing the evolution of social groups in World of Warcraft. IEEE CIG, 2010.

Applications
mining World of Warcraft

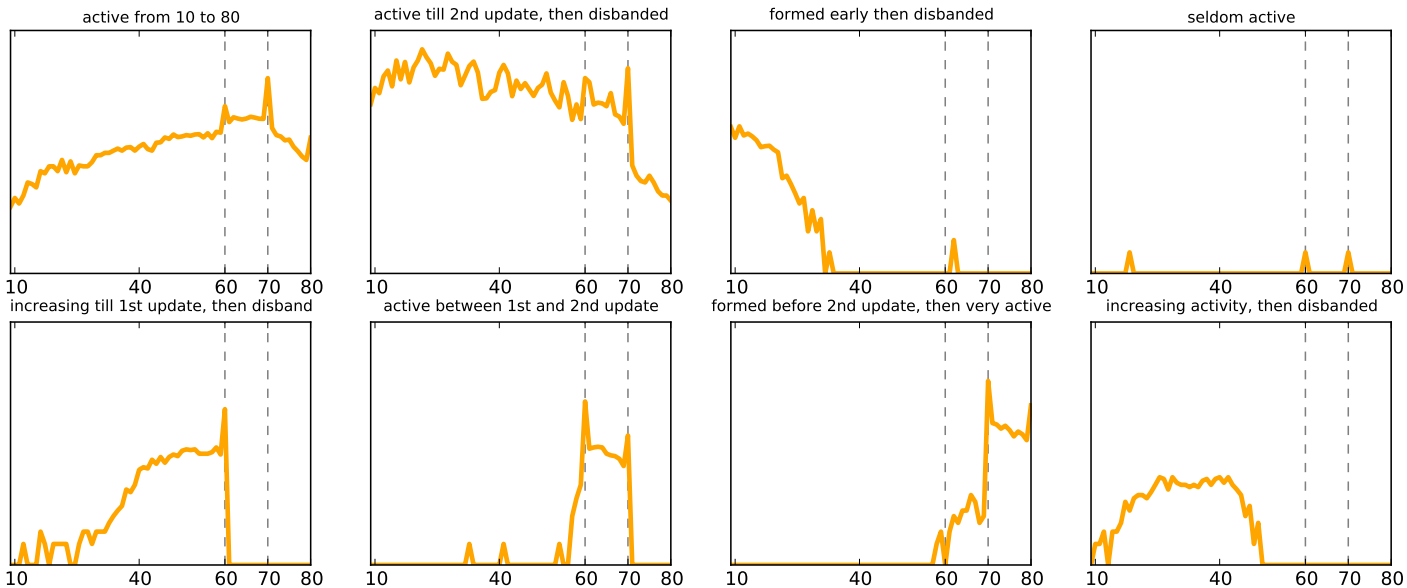
basis vectors obtained from k-means:



- rather similar entities
- difficult to interpret

Applications
mining World of Warcraft

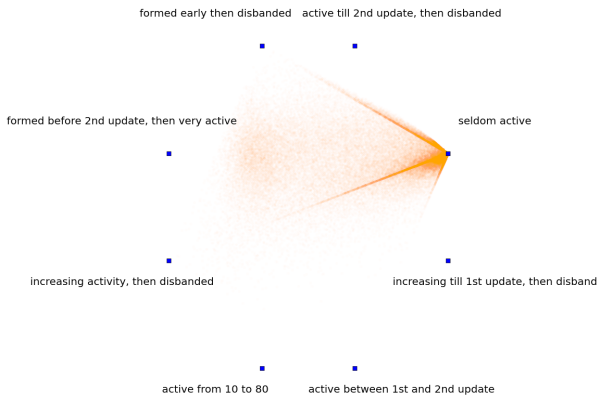
basis vectors resulting from SiVM:



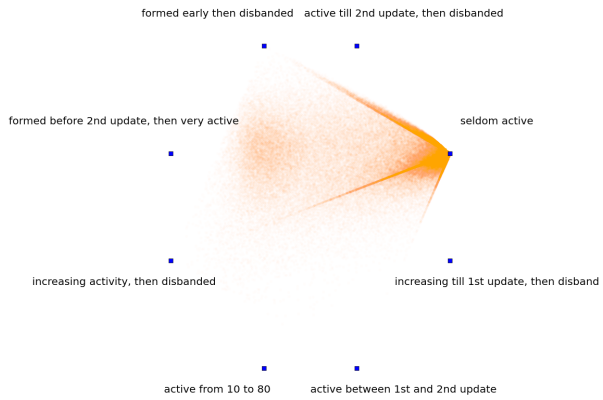
- more meaningful (since more *extreme*) entities
- correspond to existing guilds

Applications
mining World of Warcraft

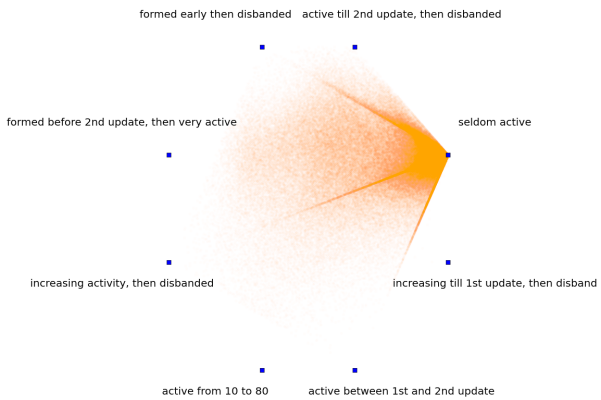
temporal evolution:



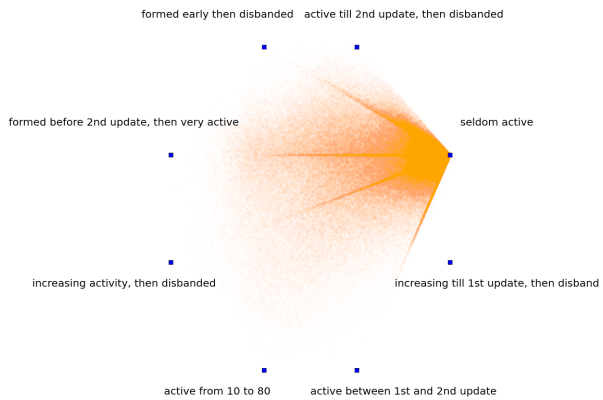
1 year



2 years



3 years



4 years

Applications
authorship analysis of twitter tweets

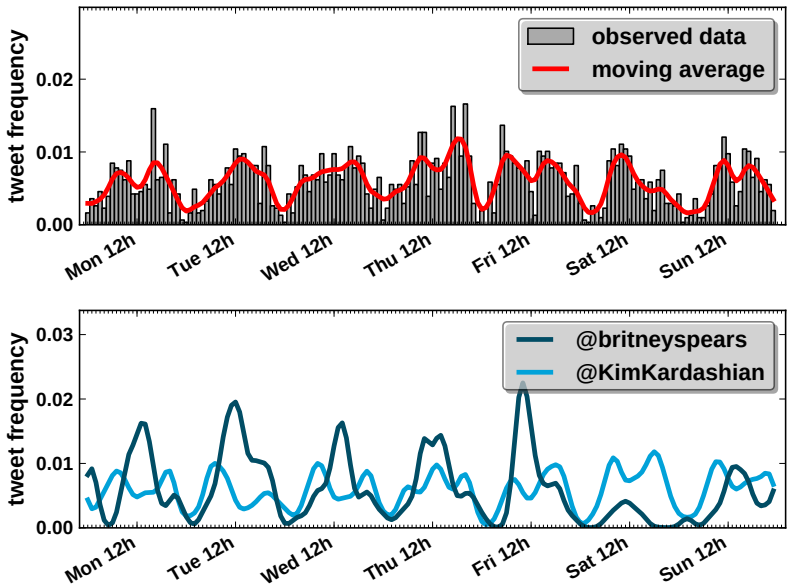
idea / motivation:

- do celebrities really tweet themselves?

data / features:

- 300 celebrity twitteres
- 1.5 million tweets
- 30 dim. feature vectors

$\Rightarrow$ data matrix of 45 million entries



details in:

- C. Thurau, K. Kersting, and C. Bauckhage: Yes we can: simplex volume maximization for descriptive web-scale matrix factorization. ACM CIKM, 2010.

Applications
authorship analysis of twitter tweets

@aplusk

z_1	I often ponder as to how similar the name Adam and the word "atom" are to one another	0.871
z_2	@thesettlers Its R	0.051
z_3	new blog... http://tinyurl.com/cggp43	0.024
z_4	http://www.ustream.tv/ashton	0.020
z_5	@toojiggy what?	0.011
z_6	@DillonPrat amazing!	0.009
z_7	HOlY GOOD GOATS MILK WE'VE GOT A QB!!!!!!!!!! THANK YOU BABY JESUS BUDAH ALLAH KABBALAH DARWIN WHOEVER!!!!!	0.007
z_8	Victory is ours!!!!!!!!	0.007

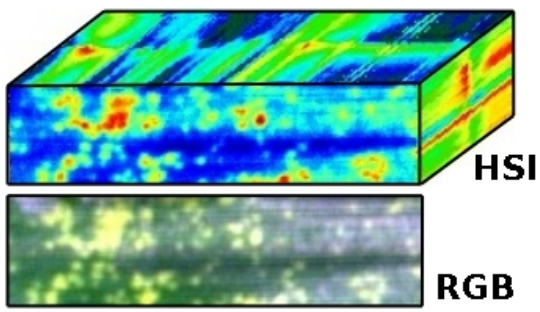
@britneyspears

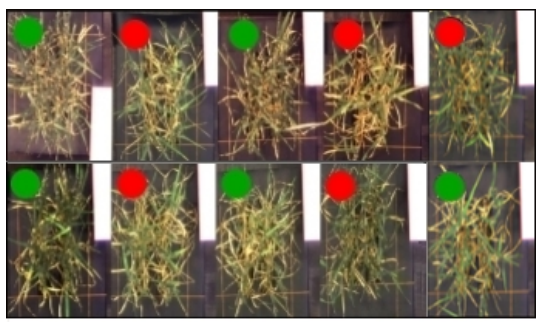
z_1	Brit's up for a People's Choice Award for Fave Scene Stealing Guest Star from How I Met Your Mother. Please vote: http://tinyurl.com/5ns9n3	0.317
z_2	Over a million hits in 24 hours on the new http://www.britneyspears.com. You guys really are the best fans in the world!	0.301
z_3	New leg, new song, new outfits, new do.... -Brit	0.301
z_4	Have you joined the Britney social network yet? Connect with other Britney fans at http://www.circusvip.com	0.032
z_5	NYC....Here I come!!! -Brit	0.022
z_6	HAPPY BIRTHDAY BRETT!!!!! XOXO -Britney	0.011
z_7	Happy Holidays from Britney!! http://tinyurl.com/7uyrr4	0.011
z_8	Thanks @britneylive @filmmaker311 @singlemomindebt @youbetheanchor @perezhilton! Were you there? Share @ http://britneyspears.com/tour	0.005

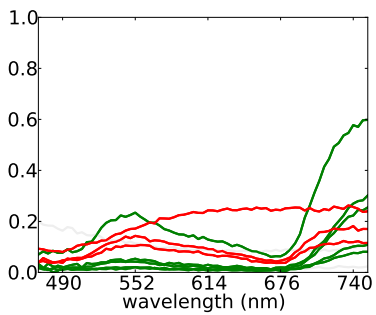
Applications
drought stress recognition in plants

idea / motivation:

- drought / abiotic stress depreciate crop yields by up to 70%
- understanding the mechanisms of stress resistance is of vital importance; phenomic approaches are called for
- hyperspectral imaging provides an auspicious modality







Applications
drought stress recognition in plants

data / approach:

- 205 daily measurements of healthy and stressed pants
- recorded with Surface Optics Corp. SOC-700
- each image is a $640 \times 640 \times 69$ data cube

$\Rightarrow$ data matrix of 5.8 billion entries

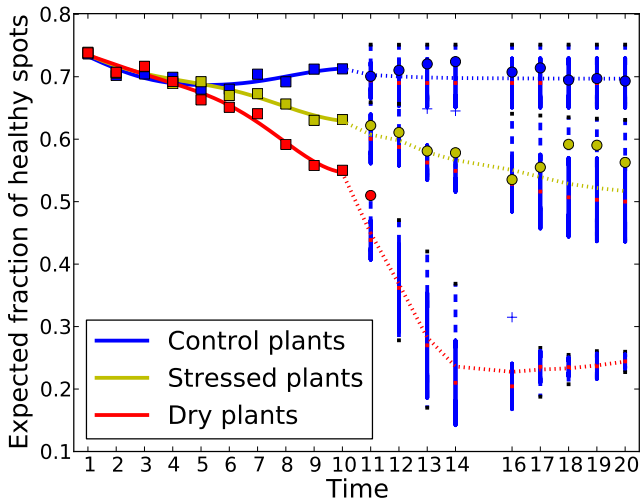
- extract extreme spectra
- fit Dirichlet distributions
- perform Bayesian regression

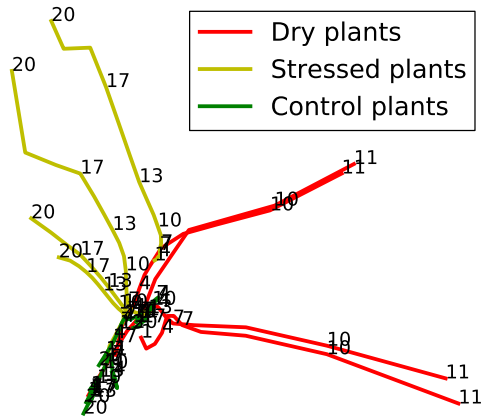
details in:

- C. Römer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Leon, C. Thurau, C. Bauckhage, K. Kersting, U. Rascher, and L. Plümer: Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis. Functional Plant Biology 39(11), 2012.
- C. Bauckhage, K. Kersting: Data Mining and Pattern Recognition in Agriculture. KI 27(4), 2013.

Applications
drought stress recognition in plants

results:

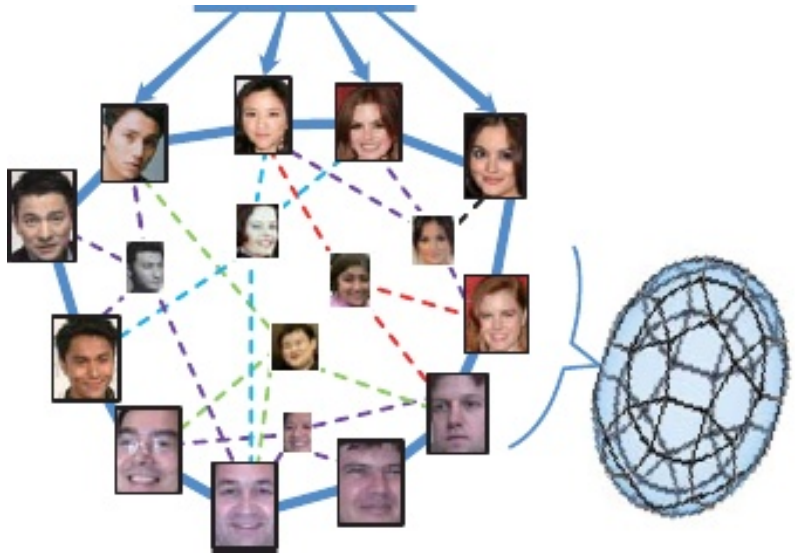




Applications
classification via archetype hulls

results:

- convex archetype regions as class model
- classify against the nearest archetype hull
- straight forward geometric classifier
- well suited for high dimensional data



details in:

- C. Thurau: Nearest archetype hull methods for large-scale data classification. ICPR 2010.
- Yuanjun Xiong, Wei Liu, Deli Zhao, and Xiaoou Tang: Face Recognition via Archetype Hull Ranking. ICCV 2013

Take home message

- k-means is a matrix factorization problem
- it is not always a good choice and lacks interpretability
- archetypal analysis is data analysis with *extremes*
- it yields intuitive and interpretable results
- it implicitly depends on distances
- it thus applies to many scenarios
- it allows for efficient computation

Credits

- Prof. Christian Bauckhage
  University Bonn and Fraunhofer IAIS
- Prof. Kristian Kersting
  TU Dortmund
- Dr. Mirwaes Wahabzada
  Fraunhofer IAIS



Thank you for your attention!

christian.thurau@unbelievable-machine.com

Appendix

note:

- k-means clustering implicitly fits a Gaussian mixture model
- k-means clustering is NP hard (Aloise et al., ML, 75(2), 2009)
- introducing indicator variables

$$r_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{x}_j \in C_i \\ 0, & \text{otherwise} \end{cases}$$

its objective can be cast as

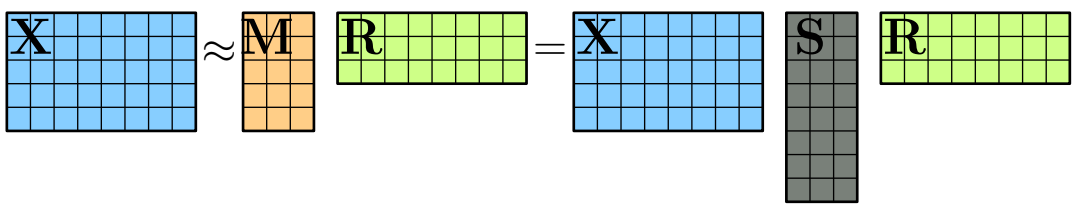
$$E(k) = \sum_{i=1}^{k} \sum_{j=1}^{n} r_{ij} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2$$

Appendix

hard k-means clustering

$$\min_{\boldsymbol{S},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{XSR} \right\|^2 - \sum_{i=1}^{k} H(\boldsymbol{s}_i)$$

$$\begin{aligned} \boldsymbol{S} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{s}_i &= 1 \\ \text{s.t.} \qquad \boldsymbol{R} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{r}_j &= 1 \\ H(\boldsymbol{r}_j) &= 0 \end{aligned}$$

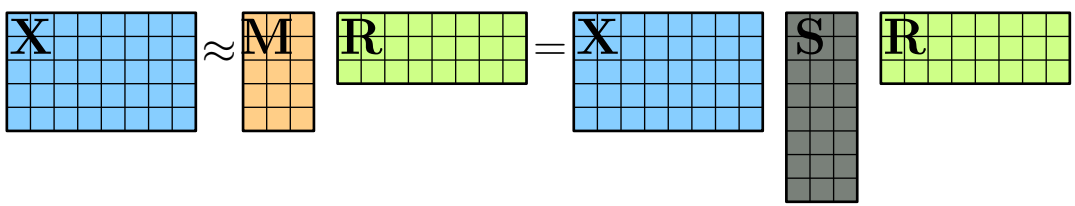


Appendix

soft k-means clustering

$$\min_{\boldsymbol{S},\boldsymbol{R}} \left\| \boldsymbol{X} - \boldsymbol{XSR} \right\|^2 - \sum_{i=1}^{k} H(\boldsymbol{s}_i)$$

$$\begin{aligned} \boldsymbol{S} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{s}_i &= 1 \\ \text{s.t.} \quad \boldsymbol{R} &\succeq \mathbf{0} \\ \mathbf{1}^T \boldsymbol{r}_j &= 1 \\ H(\boldsymbol{r}_j) &= 0 \end{aligned}$$

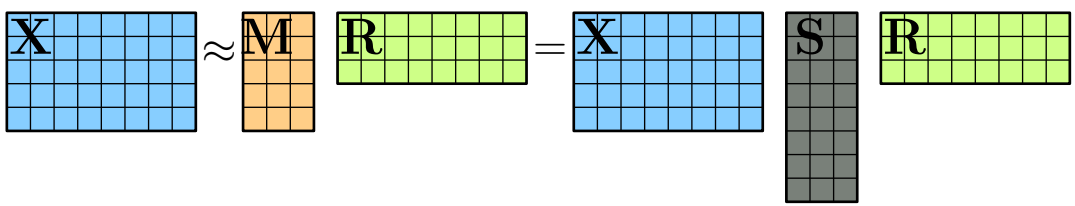


Appendix

archetypal analysis

$$\min_{S,R} \left\| X - XSR \right\|^2 - \sum_{i=1}^{k} H(s_i)$$

$$\begin{aligned} S &\succeq 0 \\ 1^T s_i &= 1 \\ \text{s.t.} \quad R &\succeq 0 \\ 1^T r_j &= 1 \\ H(r_j) &= 0 \end{aligned}$$

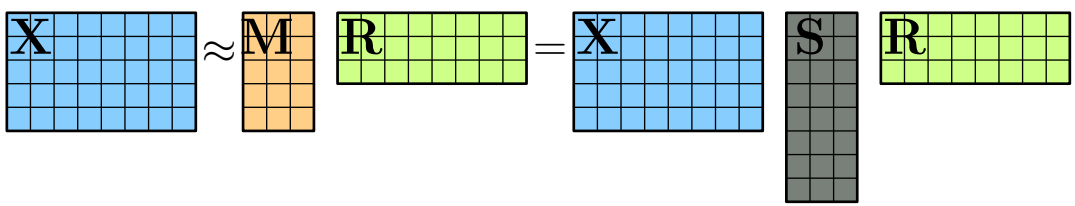


Appendix

non-negative matrix factorization

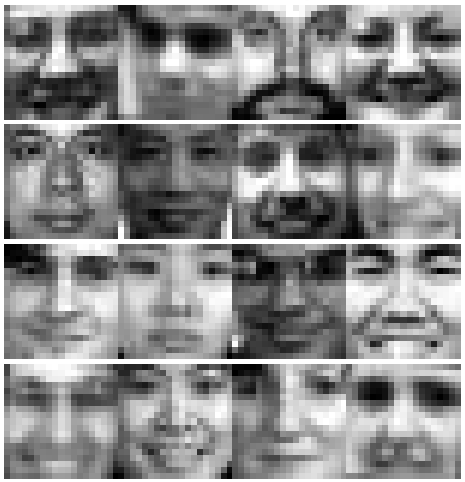
$$\min_{S,R} \left\| X - XSR \right\|^2 - \sum_{i=1}^{k} H(s_i)$$

$$\text{s.t.} \quad \begin{aligned} S &\succeq 0 \\ 1^T s_i &= 1 \\ R &\succeq 0 \\ 1^T r_j &= 1 \\ H(r_j) &= 0 \end{aligned}$$



Appendix

data (sample)



SVD

NMF

k-means

AA