Archetypal Analysis

Braxton Osting

University of Utah

October, 2021

Parts of this talk are based on joint work with Ruijian Han (CUHK), Yiming Xu (UU), Dong Wang (CUHK), and Dominique Zosso (MSU)

Archetypal Analysis

Archetypal Analysis is an unsupervised learning method that uses a convex polytope to summarize multivariate data.

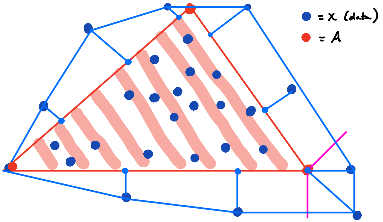
Given $k \in \mathbb{N}$ and data $X_N = \{x_i\}_{i\in[N]} \subset \mathbb{R}^d$.

Find a cardinality k pointset $A = \{a_\ell\}_{\ell\in[k]} \subset \mathbb{R}^d$ that solves

$$\min_{A \subset \mathrm{co}(X_N)} F(A)$$

where $F(A)^2 = \frac{1}{N}\sum_{i=1}^{N} d^2(x_i, \mathrm{co}(A))$.

We refer to points in $A^\star$ as *archetype points* and $\mathrm{co}(A^\star)$ as the *archetype polytope*.



Archetypal analysis with $k = 3$ and $d = 2$. Data points (blue) are projected onto the convex hull (red).

- Archetypal analysis was proposed in [Cutler and Breiman, Technometrics, 1994], where they proved:
  (i) If $k = 1$, then the archetype point is the mean of the data, X_N.
  (ii) For $1 < k < N$, there exists an archetype pointset, $A = \{a_\ell\}_{\ell\in[k]}$ and furthermore, there exists an archetype pointset on the boundary of $\mathrm{co}(X_N)$.
  (iii) Finally for $k \geq N$, the archetype pointset is given by $A = X_N$, with value $F(A) = 0$.
- They demonstrated that archetypal analysis can be reformulated as a nonlinear least squares problem and solved using an alternating minimization algorithm (small d, N, k).
- Archetypal analysis is also sometimes referred to as *principal convex hull analysis*, although we don't use this language here.

Algebraic formulation of archetypal analysis

Given $k \in \mathbb{N}$ and data $X_N = \{x_i\}_{i\in[N]} \subset \mathbb{R}^d$.

Geometric formulation. Find a pointset $A \in \{\mathrm{co}(X_N)\}^k$ that solves

$$\min_{A\in\{\mathrm{co}(X_N)\}^k} \frac{1}{N}\sum_{i=1}^{N} d^2(x_i, \mathrm{co}(A))$$

Algebraic formulation. Write $\mathbf{X} = [x_1, \cdots, x_N] \in \mathbb{R}^{d\times N}$. We can rewrite AA as the *non-negative matrix factorization* problem,

$$\begin{aligned} \min_{\mathcal{A}\in\mathbb{R}^{N\times k}, \mathcal{B}\in\mathbb{R}^{k\times N}} \quad & \frac{1}{N}\|\mathbf{X} - \mathbf{X}\mathcal{A}\mathcal{B}\|_F^2 \\ \text{s.t.} \quad & \mathcal{A}, \mathcal{B} \geq 0,\ \mathcal{A}^T 1 = 1,\ \mathcal{B}^T 1 = 1, \end{aligned}$$

Here:

- the columns of $\mathbf{X}\mathcal{A} \in \mathbb{R}^{d\times k} \in$ are the k archetype points and
- the columns of $\mathbf{X}\mathcal{A}\mathcal{B} \in \mathbb{R}^{d\times N}$ are the projection of the data points onto $\mathrm{co}(A)$.

Comparison to other unsupervised learning methods

Given $k \in \mathbb{N}$ and $X_N = \{x_i\}_{i\in[N]} \subset \mathbb{R}^d$.

- Archetypal Analysis [extreme patterns]:

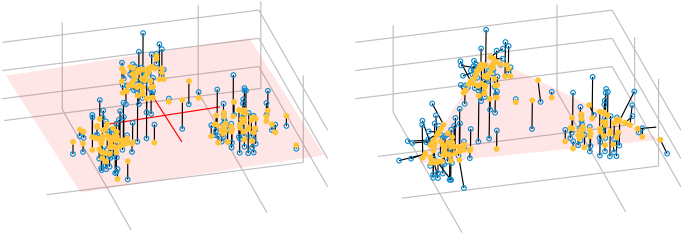
$$\min_{A\in\{\mathrm{co}(X_N)\}^k} \frac{1}{N}\sum_{i\in[N]} d^2(x_i, \mathrm{co}(A)) \iff \min_{\substack{\mathcal{A}\in\mathbb{R}^{N\times k},\ \mathcal{B}\in\mathbb{R}^{k\times N}\\ \mathcal{A},\mathcal{B}\geq 0,\ \mathcal{A}^T 1=1,\ \mathcal{B}^T 1=1}} \frac{1}{N}\|\mathbf{X}-\mathbf{X}\mathcal{A}\mathcal{B}\|_F^2$$

- K-Means [clustering]:

$$\min_{A\in\{\mathbb{R}^d\}^k} \frac{1}{N}\sum_{i\in[N]} d^2(x_i, A).$$

- Principal Component Analysis (PCA) [dimensionality reduction]:

$$\max_{V\in Gr(k,d)} \|\mathrm{Cov}(\mathrm{Proj}_V(X_N))\|_F^2 \iff \min_{\substack{U\in\mathbb{R}^{N\times k}\\ U^tU=I}} \|\mathbf{X}-\mathbf{X}UU^t\|_F^2$$



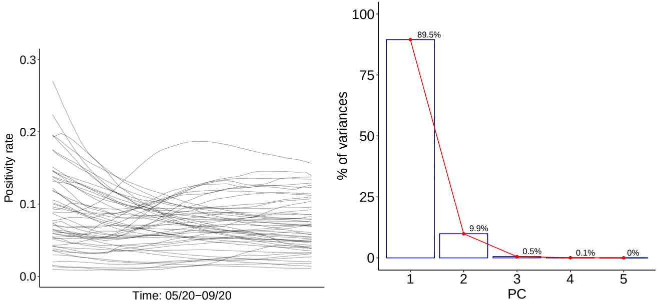
Further comparison to other matrix factorization and clustering methods can be found in [Mørup and Hansen, Neurocomputing, 2012].

Example: Covid-19 pandemic in the US

There are 51 data points[1] (50 states + D.C.), each corresponding to a time series of the (average) positivity rates. The positivity rate on a day is calculated using the following formula:

$$\text{Positivity rate} = \frac{\text{Total \# of positive cases by the day}}{\text{Total \# of tests by the day}} \times 100\%.$$

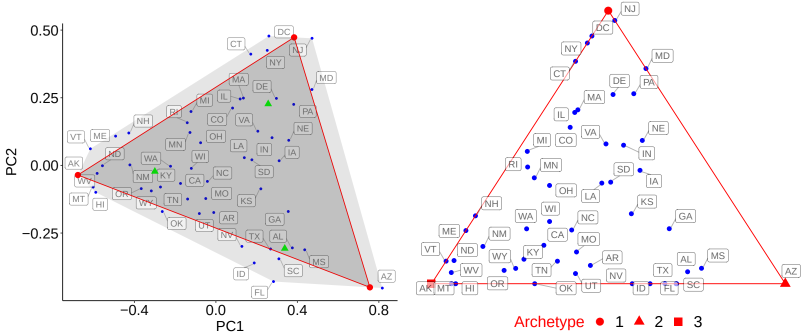
The average positivity rate is taken as the 7-day moving average of positivity rates. The time range is between May 20 and Sep 20, 2020.



(**Left**) Plot of average positivity rates in 50 states + D.C. from May 20 to Sep 20.
(**Right**) Variances explained by the first five PCs of the dataset.

[1] `https://covidtracking.com/data/api`.

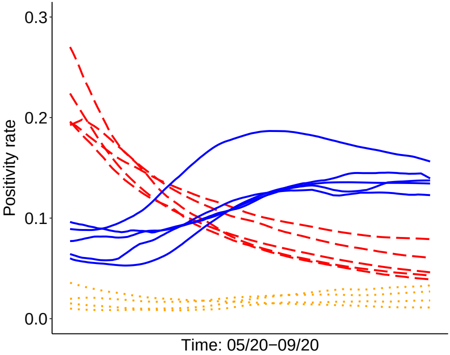
Example: Covid-19 pandemic in the US



(**Left**) Archetypal analysis ($k = 3$) applied to the reduced data representations under the first two PCs. The archetypes (red circles) are compared to the centers (green triangles) given by k-means.

(**Right**) Visualization of the data in AA coordinates.

Example: Covid-19 pandemic in the US



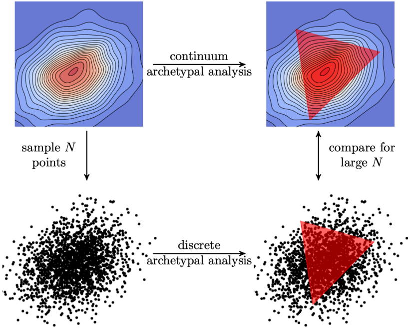
Positivity rate curves of the states near three archetypes:

1. red dashed curves (First outbreak, steadily declining),
2. blue solid curves (Second outbreak, growing and gradually stabilizing) and
3. orange dotted curves (Consistently low-positivity rates).

Consistency

Typically, a consistency result for an *estimate* has the following components:

- A statistical *assumption* on the generation of data.
- A mathematical *object* identified under the *assumption*.
- A statement of how the estimate converges to the *object* as the sample size tends to infinity, *i.e.*, a notion of convergence.
- If possible, an upper bound for the convergence rate.



Many consistency results for unsupervised learning:

- **K-Means Clustering**: [Pollard, AOS, 1981; Pollard, AOP, 1982; Sun et al., EJS, 2012].
- **PCA**: Small dimension/large sample [Girshick, AOS, 1939]. Large dimension/fixed sample [Jung and Marron, AOS, 2009]. Large dimension/large sample (under the random matrix setup) [Baik et al., AOP, 2005; Baik et al., J. Multivar. Anal, 2006].

Consistency of Archetypal Analysis
— joint work with Dong Wang, Yiming Xu, and Dominique Zosso

Suppose that $x_1, x_2, \ldots$ are independently sampled from the probability measure μ and denote the first N points by $X_N = \{x_i\}_{i\in[N]}$.

For each N, let A_N denote the optimal solution to the AA problem

$$\min_{A\in\{\mathrm{co}(X_N)\}^k} F(A).$$

Is there a set A (depending on μ), such that $A_N \to A$ as $N \to \infty$ in some sense?

To identify the limiting problem, it is useful to write

$$F(A)^2 = \frac{1}{N}\sum_{i=1}^{N} d^2(x_i, \mathrm{co}(A)) = \int_{\mathbb{R}^d} d^2(x, \mathrm{co}(A))\, d\mu_N(x).$$

where $\mu_N(x) = \frac{1}{N}\sum_{i\in[N]} \delta_{x_i}(x)$ is the empirical measure associated with the data X_N.

Since $\mu_N \rightharpoonup \mu$ as $N \to \infty$, It is natural to consider as a limiting problem

$$\min_{A\in\{\mathrm{co}(\mathrm{supp}(\mu))\}^k} F_\mu(A), \qquad \text{where} \quad F_\mu(A)^2 = \int_{\mathbb{R}^d} d^2(x, \mathrm{co}(A))\, d\mu(x).$$

Consistency of AA: Bounded Support

Theorem (O., Wang, Xu, Zosso, 2021)

Fix $k \in \mathbb{N}$. Let μ be a probability measure on $\mathbb{R}^d$ with compact support and a density. Suppose $X_N := \{x_i\}_{i\in[N]} \overset{iid}{\sim} \mu$. Then,

- *For each N, the AA problem has at least one solution A_N.*
- *$A_N \to A_\star$ (along a subsequence) in the Hausdorff distance, where*

$$A_\star \in \underset{A \in \{co(supp(\mu))\}^k}{\arg\min} F_\mu(A), \qquad F_\mu(A) = \left[\int_{\mathbb{R}^d} d^2(x,A)\, d\mu(x)\right]^{1/2}.$$

proof: Compactness + Triangle inequality

- *If $supp(\mu)$ is convex[2], then for large N, with probability at least $1 - N^{-2}$,*

$$F_\mu(A_N) - F_\mu(A_\star) \lesssim \left(\frac{\log N}{N}\right)^{1/d}.$$

proof: Random geometry + Dudley's inequality

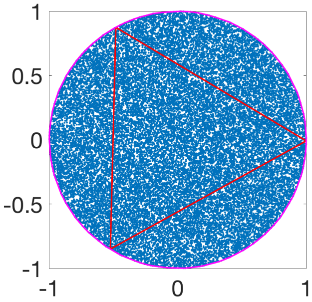
[2] The convexity assumption can be relaxed [Brunel, Bernoulli, 2019].

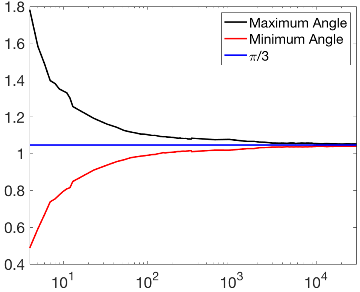
Example illustrating consistency

Theorem (O., Wang, Xu, Zosso, 2021)

When $d = 2, k \geq 3$, and μ is the uniform distribution on the unit disk, the solutions are the regular k-gons inscribed in the disk.

- The solution is non-unique.





(Left) AA applied to a dataset iid sampled from a uniform distribution on the unit disk.
(Right) The convergence of the solution to an equilateral triangle as $N \to \infty$.

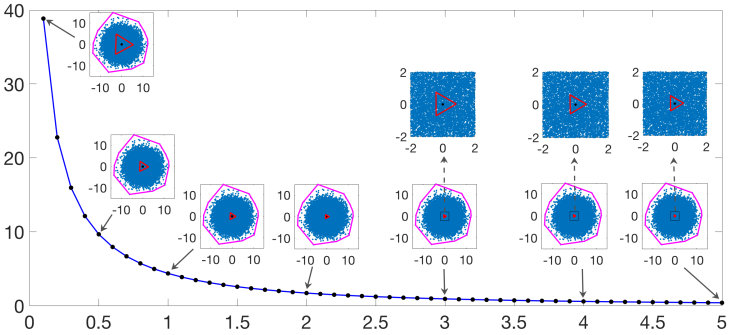
Probability Measures with Unbounded Support

Variance-regularized AA: To prevent dispersion of the archetypes, we introduce a variance regularization term

$$F_{\nu,\alpha}(A) = \frac{1}{N}\sum_{i\in[N]} d^2(x_i, \mathrm{co}(A)) + \frac{\alpha}{k}\sum_{\ell\in[k]} \|a_\ell - \bar{a}\|_2^2,$$

where $\bar{a}$ is the mean of $\{a_\ell\}_{\ell\in[k]}$ and $\alpha > 0$ is fixed.

- We prove a consistency result for this modified version.
- For large α, $\max_{a\in A_\star^{(\alpha)}} \|a - \bar{x}\|_2 \lesssim \alpha^{-1/4}$, where $\bar{x} = \int_{\mathbb{R}^d} x\, d\mu(x)$.



Variance-regularized AA applied to a dataset with increasing parameter α.

A practical challenge: computational complexity — joint work with Ruijian Han, Dong Wang, and Yiming Xu

Computational complexity limits the applicability of AA to large-scale data analysis, as it requires the solution to the following optimization problem

$$\min_{\substack{\boldsymbol{A}\in\mathbb{R}_{\mathrm{cs}}^{N\times k}\\ \boldsymbol{B}\in\mathbb{R}_{\mathrm{cs}}^{k\times N}}} \frac{1}{\sqrt{N}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}\boldsymbol{B}\|_F, \qquad \boldsymbol{X}=[x_1,\cdots,x_N]\in\mathbb{R}^{d\times N}.$$

An alternating minimization algorithm can be used to update $\boldsymbol{A}$ and $\boldsymbol{B}$ recursively:

Algorithm 1: Alternating Minimization (AM)

1: Initialize $\boldsymbol{XA}$
2: **while** not converged **do**
3: $\quad\boldsymbol{B}\leftarrow\arg\min_{\boldsymbol{B}'\in\mathbb{R}_{\mathrm{cs}}^{k\times N}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}\boldsymbol{B}'\|_F^2$ (N k-dimensional QPs)
4: $\quad\boldsymbol{A}\leftarrow\arg\min_{\boldsymbol{A}'\in\mathbb{R}_{\mathrm{cs}}^{N\times k}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}'\boldsymbol{B}\|_F^2$ (k N-dimensional QPs)
5: **end while**
6: final update for $\boldsymbol{B}$: $\boldsymbol{B}\leftarrow\arg\min_{\boldsymbol{B}'\in\mathbb{R}_{\mathrm{cs}}^{k\times N}}\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}\boldsymbol{B}'\|_F^2$
7: return $\boldsymbol{A},\boldsymbol{B}$

Complexity:

- Step 3 $\sim N\cdot\mathcal{C}(k)$
- Step 4 $\sim k\cdot\mathcal{C}(N)$,

where $\mathcal{C}(*)$ is the complexity for solving an $*$ dimensional QP.

Acceleration for AA

Previous work: Improved QP optimization methods

- Feasible optimization: projected gradients [Mørup and Hansen, Neurocomputing, 2012], active-subset [Chen et al., CVPR, 2014], Frank-Wolfe [Bauckhage et al., NCNC, 2015].
- Relaxation: decoupling [Mei et al., ECCV, 2018], sparse projection [Abrol and P. Sharma, ICML, 2020].

Previous work: Inherent complexity

- Sparse representation: random projections [Thurau et al., KAIS, 2011], NNLS [Mair et al., ICML, 2017] (acceleration for Step 4, empirical results only with no theoretical guarantee).
- Coreset [Mair and Brefeld, NeurIPS, 2019] (acceleration for Step 3, both empirical results and theoretical guarantee).

Our approach: We combine two approaches to reduce the inherent complexity:

1. *Reduce data dimensionality* via randomized low-rank approximation (data preprocessing)
2. *Reduce representation cardinality* via approximate convex hulls (acceleration for Step 4)

- Both approaches have theoretical guarantees
- Our approach can be further combined with both the improved QP optimization methods and the coreset method above

1. Data dimensionality reduction

Main idea: Find a low-dimensional representation $\widetilde{\boldsymbol{X}}$ for $\boldsymbol{X}$

First solution: A truncated SVD

Problem: Expensive when both d and N are large: $\mathcal{O}(dN\min\{d,N\})$.

Second solution: An *approximate* truncated SVD

Theorem (Han, O., Wang, Xu, 2021)

Denote the optimal objective value of AA as opt$(\boldsymbol{X})$. *Suppose* $\widetilde{\boldsymbol{X}}_p = \widetilde{\boldsymbol{U}}_p\widetilde{\boldsymbol{\Sigma}}_p\widetilde{\boldsymbol{V}}_p$ *is a 2-rank-p approximation*[3] *to* $\boldsymbol{X}$, *and denote* $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{\Sigma}}_p\widetilde{\boldsymbol{V}}_p$. *Let* $(\widetilde{\boldsymbol{A}},\widetilde{\boldsymbol{B}})$ *be a solution to the following AA for the approximate SVD representation for* $\boldsymbol{X}$*:*

$$\min_{\boldsymbol{A}\in\mathbb{R}_{cs}^{N\times k},\boldsymbol{B}\in\mathbb{R}_{cs}^{k\times N}} \frac{1}{\sqrt{N}}\left\|\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}\boldsymbol{A}\boldsymbol{B}\right\|_F .$$

Then,

$$\frac{1}{\sqrt{N}}\|\boldsymbol{X} - \boldsymbol{X}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{B}}\|_F \leq opt(\boldsymbol{X}) + 8\sigma_{p+1},$$

where σ_i *is the i-th largest singular value of* $\boldsymbol{X}$.

[3] $\mathrm{rank}(\widetilde{\boldsymbol{X}}_p) \leq p$ and $\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}_p\|_2 \leq 2\min_{\mathrm{rank}(\boldsymbol{X}_p)\leq p}\|\boldsymbol{X} - \boldsymbol{X}_p\|_2$.

Computation of $\widetilde{\boldsymbol{X}}_p$

$\widetilde{\boldsymbol{X}}_p$ can be found with high probability by applying a randomized block Krylov method [Musco and Musco, NIPS, 2015].

Implementation: Given $s, p \in \mathbb{N}$,

- generate p random initializations $\boldsymbol{S} \in \mathbb{R}^{N \times p}$, $\boldsymbol{S}_{ij} \sim \mathcal{N}(0, 1)$
- construct the Krylov subspace: $\boldsymbol{K} = [\boldsymbol{XS}, (\boldsymbol{XX}^T)\boldsymbol{XS}, \cdots, (\boldsymbol{XX}^T)^{s-1}\boldsymbol{XS}] \in \mathbb{R}^{d \times (sp)}$
- compute the QR decomposition of $\boldsymbol{K}$: $\boldsymbol{K} = \boldsymbol{QR}$
- compute the SVD of $\boldsymbol{X}^T\boldsymbol{Q}$: $\boldsymbol{X}^T\boldsymbol{Q} = \boldsymbol{U}_{\text{emd}}\boldsymbol{\Sigma}_{\text{emd}}\boldsymbol{V}_{\text{emd}}^T$
- compute $\widetilde{\boldsymbol{X}}_p$: $\widetilde{\boldsymbol{X}}_p = \boldsymbol{LL}^T\boldsymbol{X}$, with $\boldsymbol{L} = \boldsymbol{QV}_{\text{emd}}[:, 1:p]$

Lemma (Musco and Musco, NIPS, 2015)

For $\delta > 0$, if $p \gtrsim \log(1/\delta)$ and $s \gtrsim \log(N/\delta)$, then with probability at least $1 - \delta$,

$$\mathbb{P}\left[\widetilde{\boldsymbol{X}}_p \textit{ is a 2-rank-p approximation to } \boldsymbol{X}\right] \geq 1 - \delta.$$

Consequence: $\widetilde{\boldsymbol{X}}$ can be computed within time $\mathcal{O}(dN \log Np)$.

2. Representation cardinality reduction

<u>Main idea</u>: Use the extreme points of $\boldsymbol{X}$ as a dictionary

<u>Problem</u>: Expensive if $\boldsymbol{X}$ has a large number of extreme points

<u>Solution</u>: Select a few 'important' extreme points for representation

Theorem (Han, O., Wang, Xu, 2021)

Denote the optimal objective value of AA as $opt(\boldsymbol{X})$. *For* $T \subset [N]$, *suppose* $\boldsymbol{X}_T$ *satisfies*

$$d_H(co(\boldsymbol{X}_T), co(\boldsymbol{X})) \leq opt(\boldsymbol{X}) \cdot \epsilon,$$

where d_H *is the Hausdorff distance. Consider the following AA optimization problem constrained to* $co(\boldsymbol{X}_T)$:

$$\min_{\boldsymbol{A} \in \mathbb{R}_{cs}^{|T| \times k}, \boldsymbol{B} \in \mathbb{R}_{cs}^{k \times N}} \frac{1}{\sqrt{N}} \|\boldsymbol{X} - \boldsymbol{X}_T \boldsymbol{A} \boldsymbol{B}\|_F.$$

Then,

$$\min_{\boldsymbol{A} \in \mathbb{R}_{cs}^{|T| \times k}, \boldsymbol{B} \in \mathbb{R}_{cs}^{k \times N}} \frac{1}{\sqrt{N}} \|\boldsymbol{X} - \boldsymbol{X}_T \boldsymbol{A} \boldsymbol{B}\|_F \leq (1 + \epsilon) opt(\boldsymbol{X}).$$

Computation of $\boldsymbol{X}_T$

$\boldsymbol{X}_T$ can be found via random projections [Graham and Oberman, arXiv., 2017].
Idea: points that are more likely to be sampled are also more 'important'.

Implementation: Given $\eta > 0$ and $M \in \mathbb{N}$,

- Draw M iid (uniform) random vectors $\{v_i\}_{i\in[M]}$ on $\mathbb{S}^{d-1}$
- For v_i, find the column in $\boldsymbol{X}$ giving the largest v_i-projected value
- For $i \in [N]$, count the frequency f_i of $\boldsymbol{X}[:, i]$ being maximum, and rearrange f_i in decreasing order $f_{\tau_1} \geq \cdots \geq f_{\tau_N}$
- Choose $T = \{\tau_j\}_{j\in[L]}$, where $L = (d+1) \vee \min\{\ell : \sum_{j\leq\ell} f_{\tau_j} \geq 1 - \eta/3\}$

Theorem (Han, O., Wang, Xu, 2021)

For $i \in [N]$, denote κ_i the curvature of x_i: $\kappa_i := \sigma_{re}(\{v \in \mathbb{S}^{d-1} : v^T x_i > v^T x_j, j \neq i\})$. Denote q as the smallest integer such that $\sum_{i\in[q]} \kappa_i \geq 1 - \eta/18$, and the truncation gap $\Delta = \kappa_q - \kappa_{q+1}$. Under suitable conditions, if $\Delta > 0$ and

$$M \geq \max\left\{\frac{324q^2}{\eta^2}, \frac{4}{\Delta^2}\right\} \log\left(\frac{3N}{\sqrt{\delta}}\right),$$

then with probability at least $1-\delta$, $|T| \leq \max\{q, p+1\}$ and

$$d_H(co(\boldsymbol{X}_T), co(\boldsymbol{X})) \leq \min\left\{\sqrt{2}\pi\eta^{\frac{1}{d-1}}, 2\right\} \cdot \max_{i\in[N]} \|x_i\| \qquad \text{(curse of dimensionality)}$$

Approximate archetypal analysis (AAA))

Algorithm 2: Approximate Archetypal Analysis (AAA)

Input: $\{x_i\}_{i\in[N]}$: dataset, k: number of archetypes, p: approximation rank, s: Krylov subspace parameter, M: number of projections, η: approximation accuracy

Output: a solution to AA

1: generate p random initializations: $\boldsymbol{S} \in \mathbb{R}^{N\times p}$, $\boldsymbol{S}_{ij} \sim \mathcal{N}(0,1)$

2: construct the Krylov subspace: $\boldsymbol{K} = [\boldsymbol{XS}, (\boldsymbol{XX}^T)\boldsymbol{XS}, \cdots, (\boldsymbol{XX}^T)^{s-1}\boldsymbol{XS}] \in \mathbb{R}^{d\times(brown)}$

3: compute the QR decomposition of $\boldsymbol{K}$: $\boldsymbol{K} = \boldsymbol{QR}$

4: compute the SVD of $\boldsymbol{X}^T\boldsymbol{Q}$: $\boldsymbol{X}^T\boldsymbol{Q} = \boldsymbol{U}_{\text{emd}}\boldsymbol{\Sigma}_{\text{emd}}\boldsymbol{V}_{\text{emd}}^T$

5: form approximate SVD representation: $\widetilde{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\text{emd}}[1:p, 1:p](\boldsymbol{U}_{\text{emd}}[:, 1:p])^T$

6: apply random projections to $\widetilde{\boldsymbol{X}}$ with parameters (M, η) to find $\widetilde{\boldsymbol{X}}_T$

7: solve the reduced archetypal analysis problem:

$$(\widetilde{\boldsymbol{A}}_\star, \widetilde{\boldsymbol{B}}_\star) \in \arg \min_{\widetilde{\boldsymbol{A}} \in \mathbb{R}_{\text{cs}}^{|T|\times k}, \widetilde{\boldsymbol{B}} \in \mathbb{R}_{\text{cs}}^{k\times N}} \frac{1}{\sqrt{N}} \|\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}_T \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{B}}\|_F,$$

8: extend $\widetilde{\boldsymbol{A}}_\star$ to an $\mathbb{R}^{N\times k}$ matrix by first creating a zero matrix $\boldsymbol{A}_{null} \in \mathbb{R}^{N\times k}$, then $\boldsymbol{A}_{null}[T,:] \leftarrow \widetilde{\boldsymbol{A}}_\star$, and finally $\widetilde{\boldsymbol{A}}_\star \leftarrow \boldsymbol{A}_{null}$

9: return $(\widetilde{\boldsymbol{A}}_\star, \widetilde{\boldsymbol{B}}_\star)$

Theoretical guarantee for AAA

Theorem (Han, O., Wang, Xu, 2021)

Assuming $p \gtrsim \log(1/\delta)$, if

$$s \gtrsim \log\left(\frac{N}{\delta}\right) \qquad \eta = \left(\frac{opt(\boldsymbol{X})\epsilon}{\sqrt{2}\pi \max_{i\in[N]} \|x_i\|}\right)^{p-1}$$

$$M \gtrsim \max\left\{\frac{q^2}{\eta^2}, \frac{1}{\Delta^2}\right\}\log\left(\frac{N}{\delta}\right),$$

then with probability at least $1-2\delta$, $|T| \leq \max\{q, p+1\}$, and the approximate archetypes $\boldsymbol{X}\widetilde{\boldsymbol{A}}_\star$ as well as the coefficient matrix $\widetilde{\boldsymbol{B}}_\star$ returned by AAA satisfy

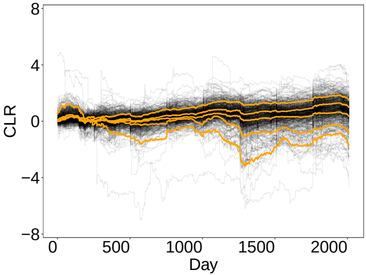
$$\frac{1}{\sqrt{N}}\|\boldsymbol{X} - \boldsymbol{X}\widetilde{\boldsymbol{A}}_\star\widetilde{\boldsymbol{B}}_\star\|_F \leq (1+\epsilon)opt(\boldsymbol{X}) + 8\sigma_{p+1},$$

where σ_i is the i-th largest singular value of $\boldsymbol{X}$.

Remark: Data preprocessing has complexity $\mathcal{O}(dN\log Np + \epsilon^{-2(p-1)}N\log^2 Npq)$. AM has complexity equal to solving an $p \times |T|$, $|T| \leq \max\{p+1, q\}$ size AA. The overall complexity for AAA is small if both p, q are small. In other words, $\boldsymbol{X}$ is approximately low-rank and has most of the curvature concentrated on a small subset of extreme points.

Numerical Example: S&P 500 stocks

572 S&P 500 stocks from 2011 to 2018 [4]. Each data point corresponds to the cumulative log-return (CLR) of the stock of a company from Jan 2011 to Dec 2018 (2012 days).

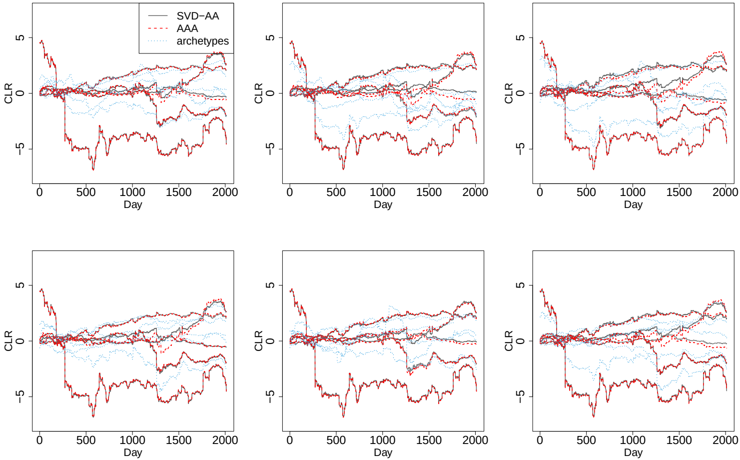


Cumulative log-return (CLR) of 572 S&P 500 stocks from January 2011 to December 2018. Orange curves are the centers of the K-means applied to $\boldsymbol{X}$ with $k = 5$.

Fix $k = 5$. Three different methods are applied to compute the archetypes: SVD-AA, AAA (with $p = 50, M = 10^4, \eta = 0.003$) and a package function `archetypes` in `R` for archetypal analysis. Each experiment is repeated 50 times.

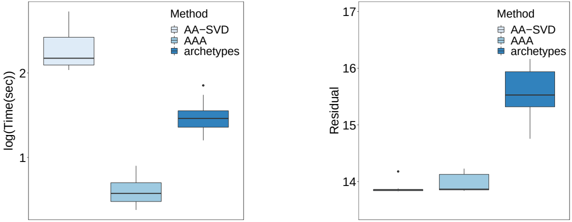
[4] This dataset is provided to us by Yu Zhu, a Ph.D. Student at the David Eccles Business School, University of Utah

Numerical Example: S&P 500 stocks

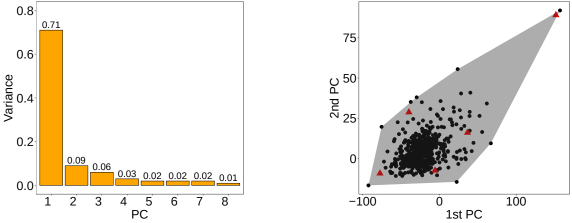


6 Instances of the computed archetypes by SVD-AA, AAA, and archetypes.

Numerical Example: S&P 500 stocks



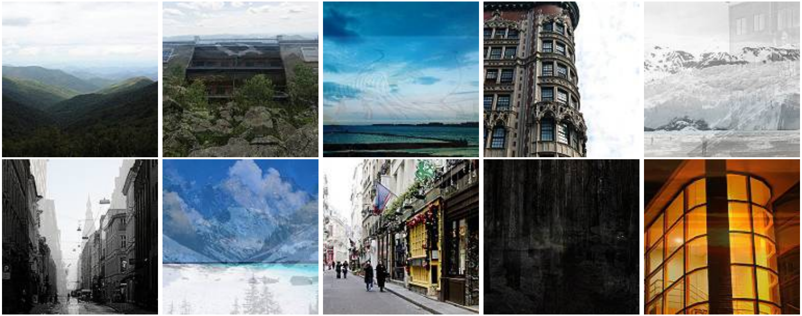
Boxplots of the running times **(Left)** and residuals **(Right)** of SVD-AA, AAA and archetypes for 50 experiments.



(Left) Variances explained by the first 8 principal components of $\boldsymbol{X}$.
(Right) Scatterplot of the reduced representation of $\boldsymbol{X}$ with respect to the first two PCs and its convex hull. The red triangles are the reduced representation of five archetypes

Numerical Example: Intel Image

Intel Image[5] is a public dataset consisting of 24000 images representing 6 different categories of scene: Buildings, Forest, Glacier, Mountain, Sea and Street. Each data point is a 150×150 pixel color image. We randomly select 3000 samples in Intel Image and apply AAA to extract $k = 10$ representative patterns. The input parameters for AAA are chosen as $p = 10$, $M = 10^5$ and $\eta = 0.003$.



Ten archetypes computed by AAA, which account for 44% of the total variance of the dataset. The computation time is 348.784s (85.012s for data dimensionality reduction, 4.715s for representation cardinality reduction and 259.057s for solving the reduced problem using AM).

[5] https://www.kaggle.com/puneet6060/intel-image-classification

Discussion

- For bounded distributions, we identified a continuum problem of archetypal analysis and established a consistency result including the convergence rate.
- For unbounded distributions, we introduced a variance-regularized problem and established a consistency result. We also investigated how the solutions depend on the regularization parameter.
- Devised an approximate algorithm for large-scale AA which enjoys theoretical guarantees

Thanks! Questions? Email: osting@math.utah.edu

B. Osting, D. Wang, Y. Xu, and D. Zosso, Consistency of archetypal analysis, *SIAM Journal on Mathematics of Data Science* (2021) https://arxiv.org/abs/2010.08148

R. Han, B. Osting, D. Wang, and Y. Xu, Probabilistic methods for approximate archetypal analysis, submitted (2021) http://arxiv.org/abs/2108.05767

Thanks to support by NSF DMS 17-52202.