Consistency of Archetypal Analysis

Braxton Osting

University of Utah

April 20, 2021

Based on joint work with 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.

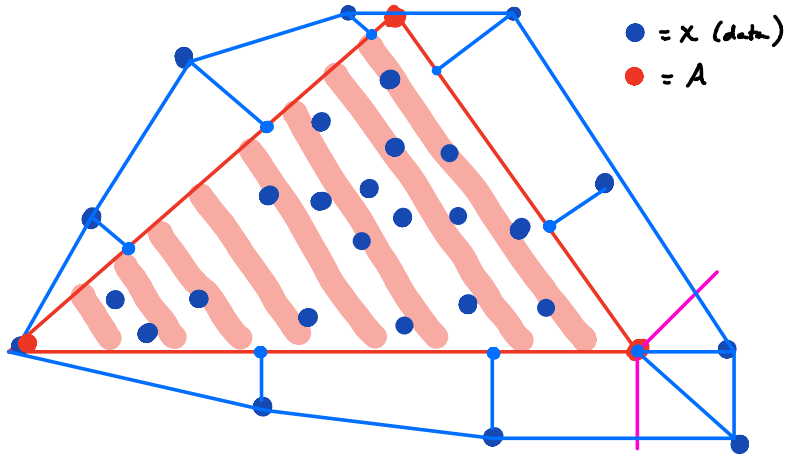
Given $k \in \mathbb{N}$ and data $X_N = \{x_i\}_{i \in [N]} \subset \mathbb{R}^d$.

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 efficiently solved using an alternating minimization algorithm.
- 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]:

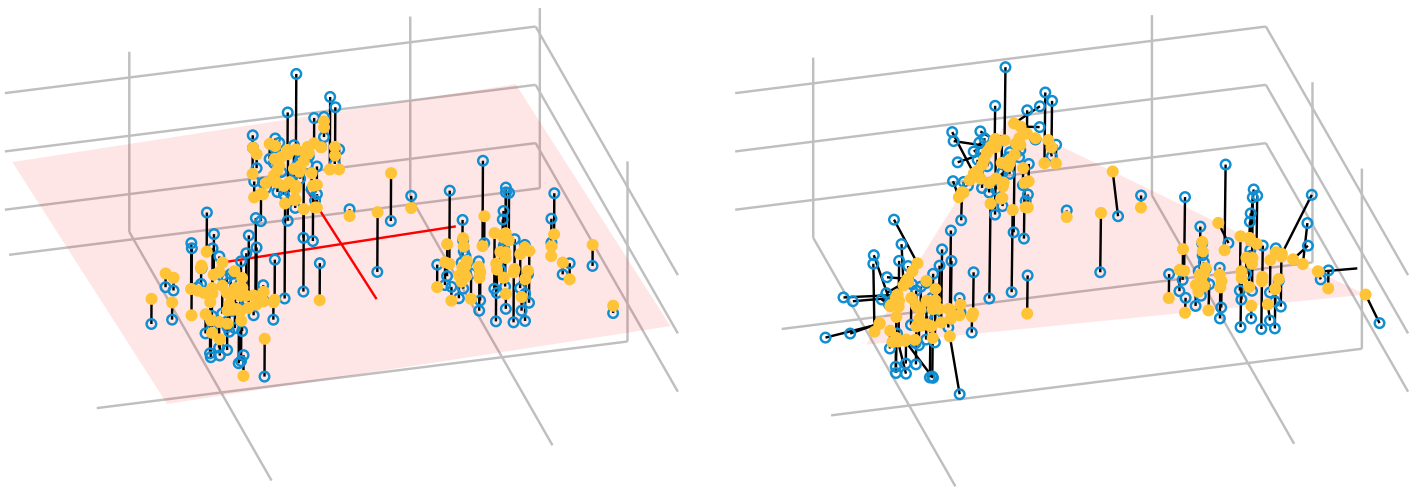
$$\min_{A\in\{\mathrm{co}(X_N)\}^k} \frac{1}{N}\sum_{i\in[N]} d^2(x_i, \mathrm{co}(A)) \quad \Longleftrightarrow \quad \min_{\substack{\mathcal{A}\in\mathbb{R}^{N\times k},\ \mathcal{B}\in\mathbb{R}^{k\times N} \\ \mathcal{A},\mathcal{B}\geq 0,\ \mathcal{A}^T 1=1,\ \mathcal{B}^T 1=1}} \frac{1}{N}\|\mathbf{X}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

- 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]:

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



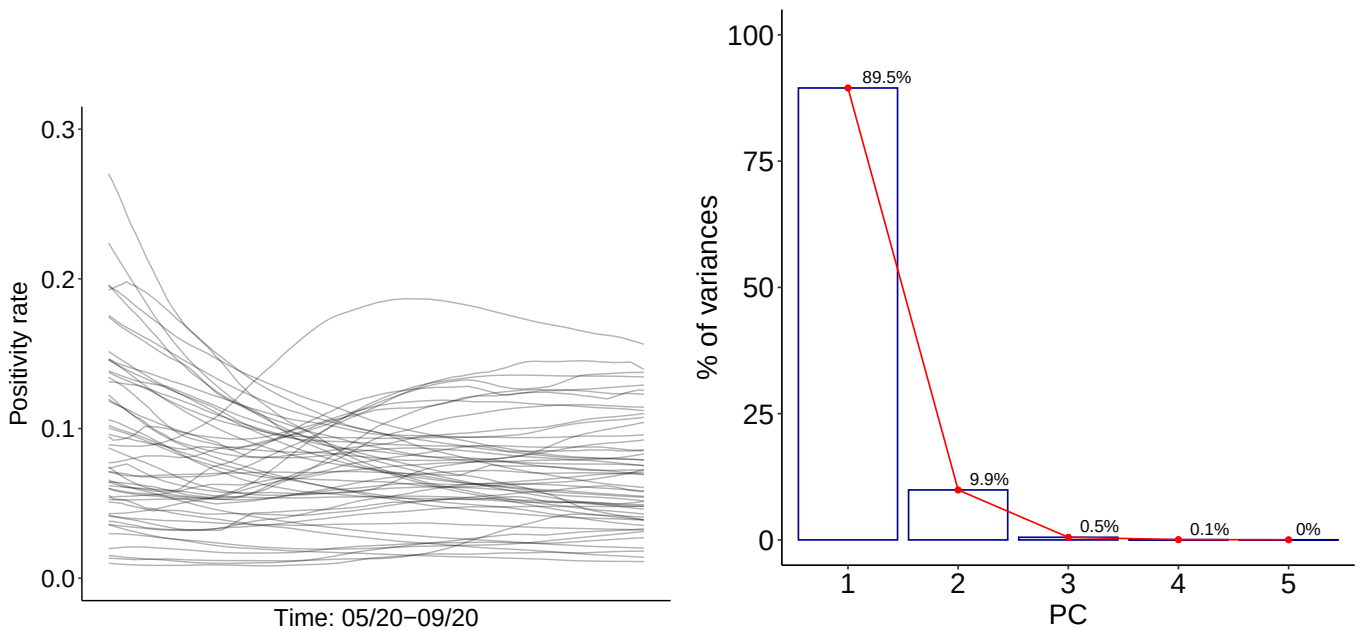
Further comparison to other matrix factorization and clustering methods can be found in [Mørup and Hansen, Neurocomputing, 2012].

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:

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

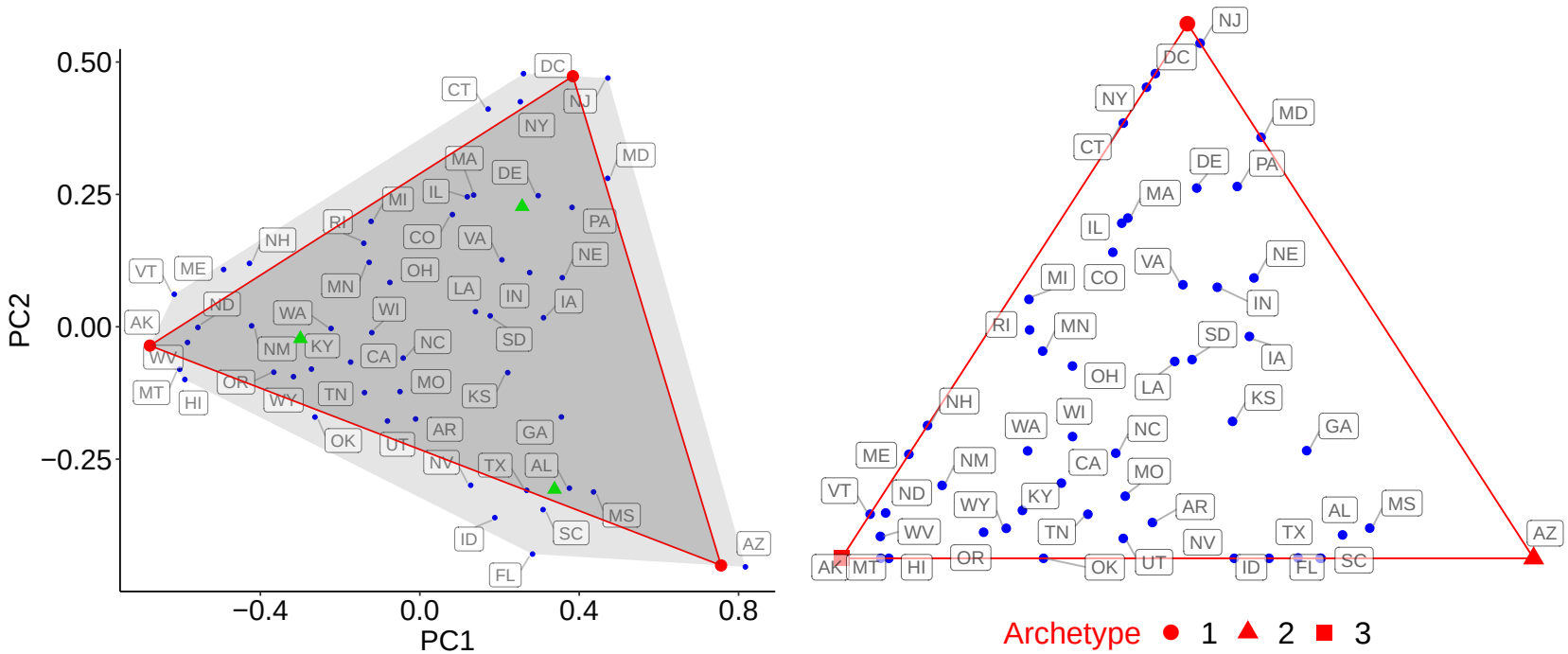
The average positivity rate is taken as the 7-day moving average of positivity rates. The time range is between May 20 and Sep 20, 2020.



(**Left**) Visualization 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.

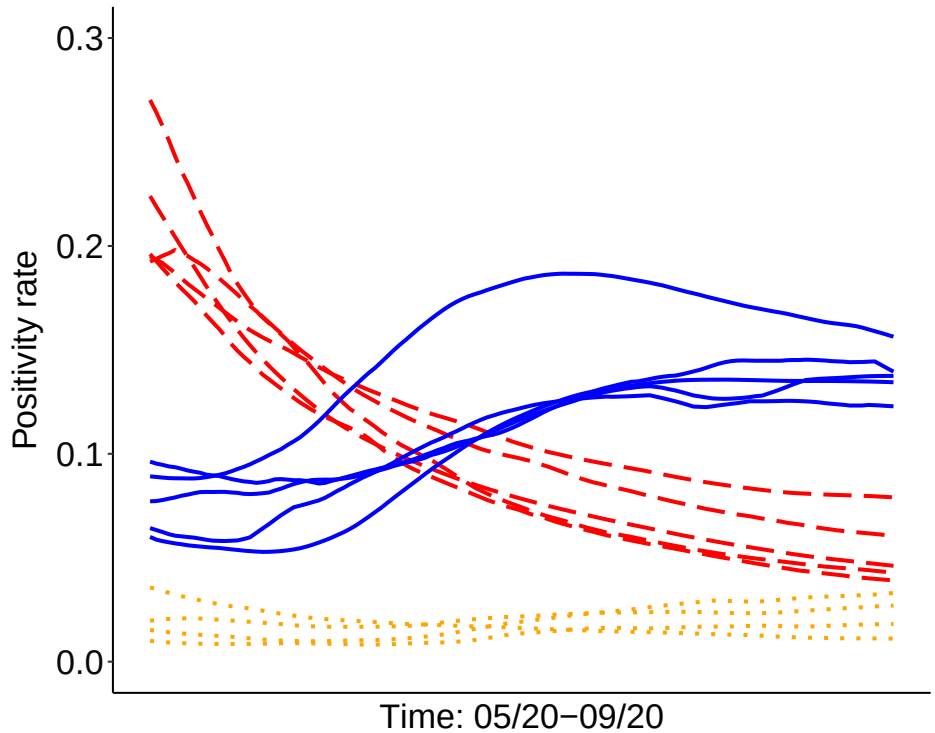
[1] `https://covidtracking.com/data/api`.

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 AA coefficients of the projected reduced data points with respect to three archetypes.

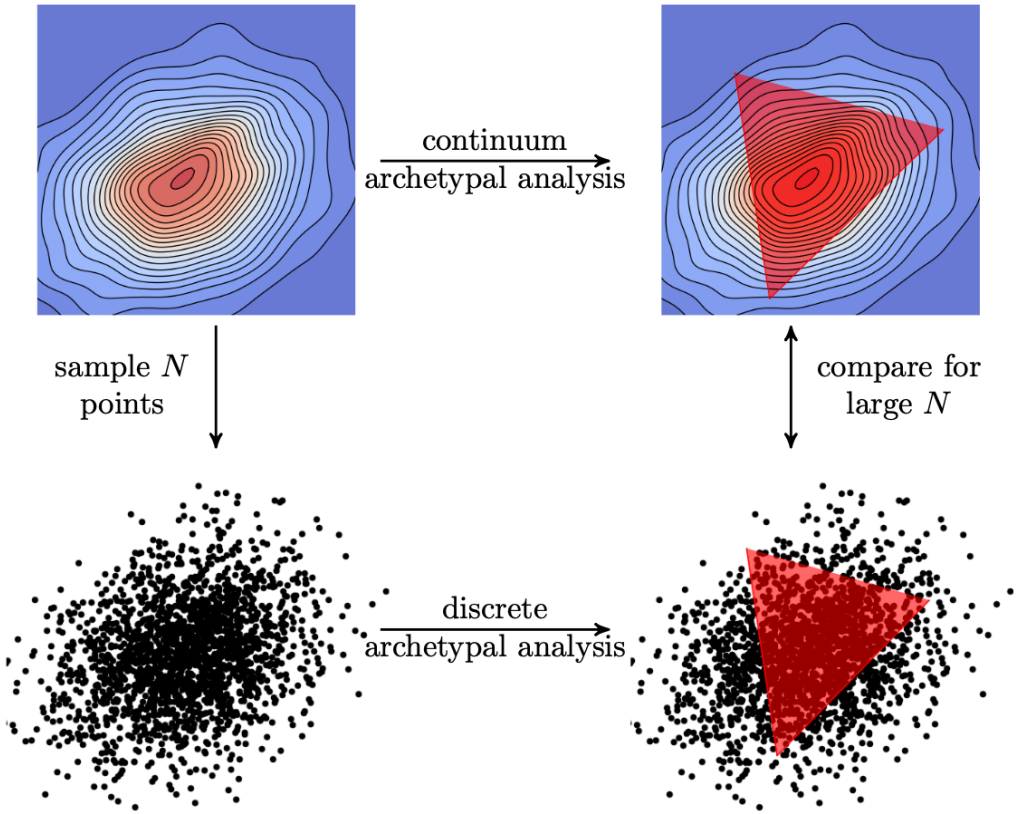
Example: Covid-19 pandemic in the US



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.

Related Results

A selection of consistency results concerning 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].
- **Spectral Clustering**: Diffusion dynamics [Coifman and Lafon, ACHA, 2005; Berry and Sauer, ACHA, 2015]. Partitioning [Luxburg et al., AOS, 2008; Garcia Trillos et al., JMLR, 2016; Garcia Trillos et al., ACHA, 2018, Osting and Reeb, SIMA, 2017].
- **PageRank**: [Yuan, Calder and Osting, EJAM, 2021].

Consistency of Archetypal Analysis

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

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

where $\mu_N(x) = \frac{1}{N}\sum_{i\in[N]} \delta_{x_i}(x)$ is the emperical 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

$$\min_{A\in\{\mathrm{co}(\mathrm{supp}(\mu))\}^k} F_\mu(A)$$

where $F_\mu(A)^2 = \int_{\mathbb{R}^d} d^2(x, \mathrm{co}(A))\, d\mu(x)$.

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*

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

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

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

Consistency of AA: Bounded Support

Proof Sketch

- *Continuity + Compactness $\Rightarrow$ Existence of minimizers.*
- *Compactness + Triangle Inequality $\Rightarrow$ Consistency.*
- *Random Geometry + Dudley's Inequality $\Rightarrow$ Convergence rate.*

Remarks

- *Compactness is necessary for the problem to make sense.*
- *Minimizers are not unique in general.*

  For example, 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 convexity assumption can be relaxed [Brunel, Bernoulli, 2019].*

Consistency of AA: Unbounded Support

We now consider the consistency problem when the probability measure, μ, has non-compact support. Here we have that $\text{co}(X_1) \subseteq \text{co}(X_2) \subseteq \cdots$ and $\text{co}(X_N)$ is a.s. unbounded as $N \to \infty$.

In this case, it is clear that there can be no limiting problem for the archetype pointset A_N as $N \to \infty$; the problem, as stated, is inconsistent.

Consequently, we must modify AA to obtain a consistency result.

Modification: Introduce a variance regularization term to prevent dispersion of the archetypes:

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

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

We refer to this modified problem as *variance-regularized* Archetypal Analysis.

Consistency of AA: Unbounded Support

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

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

- *For each N, the variance-regularized AA problem has at least one solution $A_N^{(\alpha)}$.*
- *Moreover, suppose that μ-a.s., for any $r > 0$, there exists some $N_r \in \mathbb{N}$ such that $B(r) \subset co(X_N)$ for $N > N_r$.*
  Then $A_N^{(\alpha)} \to A_\star^{(\alpha)}$ (up to a subsequence) in the Hausdorff distance, where

$$A_\star^{(\alpha)} \in \underset{A\in\{co(supp(\mu))\}^k}{\arg\min} \left[F_\mu(A)^2 + \frac{\alpha}{k} \sum_{\ell\in[k]} \|a_\ell - \bar{a}\|_2^2 \right]^{1/2}, \quad \bar{a} = \frac{1}{k} \sum_{\ell\in[k]} a_\ell.$$

- *For large α,*

$$\max_{a\in A_\star^{(\alpha)}} \|a - \bar{x}\|_2 \lesssim \alpha^{-1/4}, \quad \bar{x} = \int_{\mathbb{R}^d} x \, d\mu(x).$$

Computing approximate solutions to AA

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

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

which can be approximately solved via alternating minimization:

1: Initialize $\mathcal{A}$ and set $\mathcal{Z} = \mathbf{X}\mathcal{A}$
2: **while** Not Converged **do**
3:

$$\mathcal{B} \leftarrow \underset{\mathcal{B} \in \mathbb{R}^{k \times N}}{\arg\min} \ \frac{1}{N} \|\mathbf{X} - \mathcal{Z}\mathcal{B}\|_F^2, \quad \text{s.t. } \mathcal{B} \geq 0, \ \mathcal{B}^T 1 = 1.$$

$$\mathcal{A} \leftarrow \underset{\mathcal{A} \in \mathbb{R}^{N \times k}}{\arg\min} \ \frac{1}{N} \|\mathbf{X} - \mathbf{X}\mathcal{A}\mathcal{B}\|_F^2, \quad \text{s.t. } \mathcal{A} \geq 0, \ \mathcal{A}^T 1 = 1.$$

4: **end while**

After applying a Gauss-Seidel strategy and reformulating the second problem, both subproblems can be seen as constrained least squares problems where the (convex) constraint set is the unit simplex. These are efficiently solved using a projected gradient descent method.

Numerical Simulations: Consistency

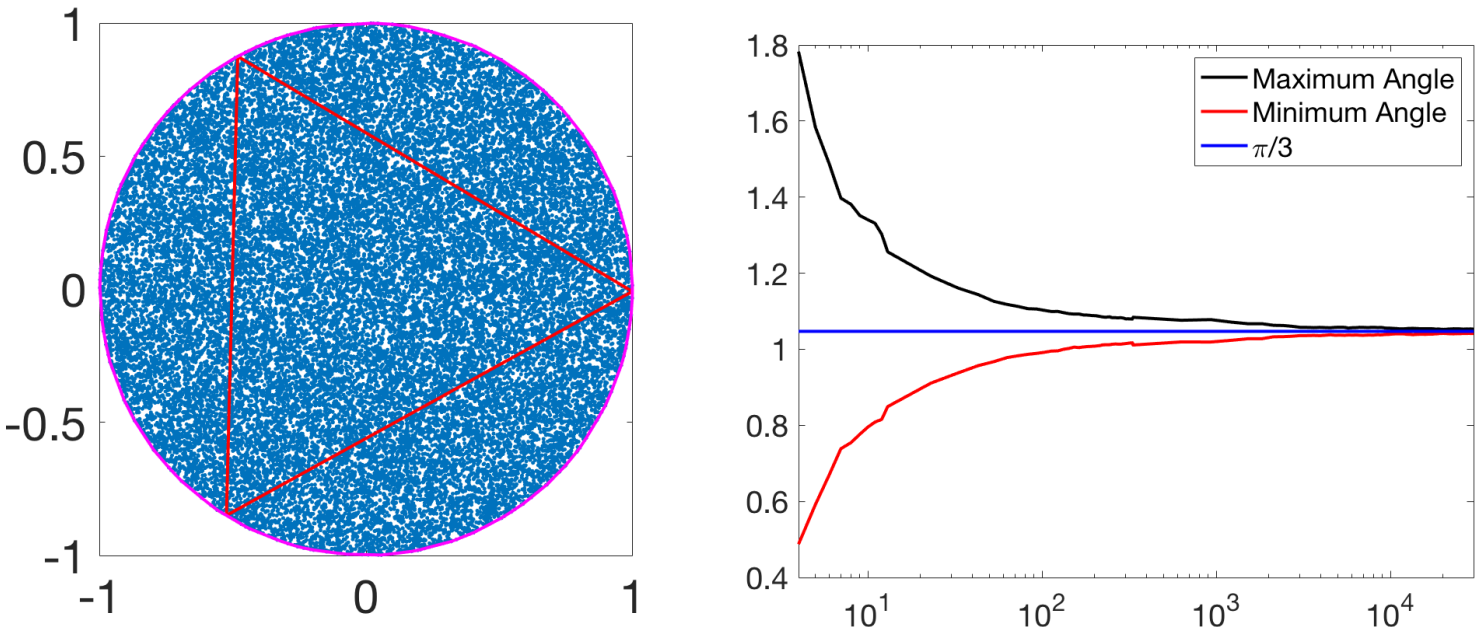


Figure: AA applied to a dataset independently sampled from a uniform distribution on the unit disk in $\mathbb{R}^2$, as opposed to the theoretic solutions (inscribed equilateral triangles).

Numerical Simulations: Dependence on regularization parameter, α

Variance-regularized Archetypal Analysis:

$$\text{Objective function:} \quad \frac{1}{N}\sum_{i\in[N]} d^2(x_i, \text{co}(A)) + \frac{\alpha}{k}\sum_{\ell\in[k]} \|a_\ell - \bar{a}\|_2^2,$$

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

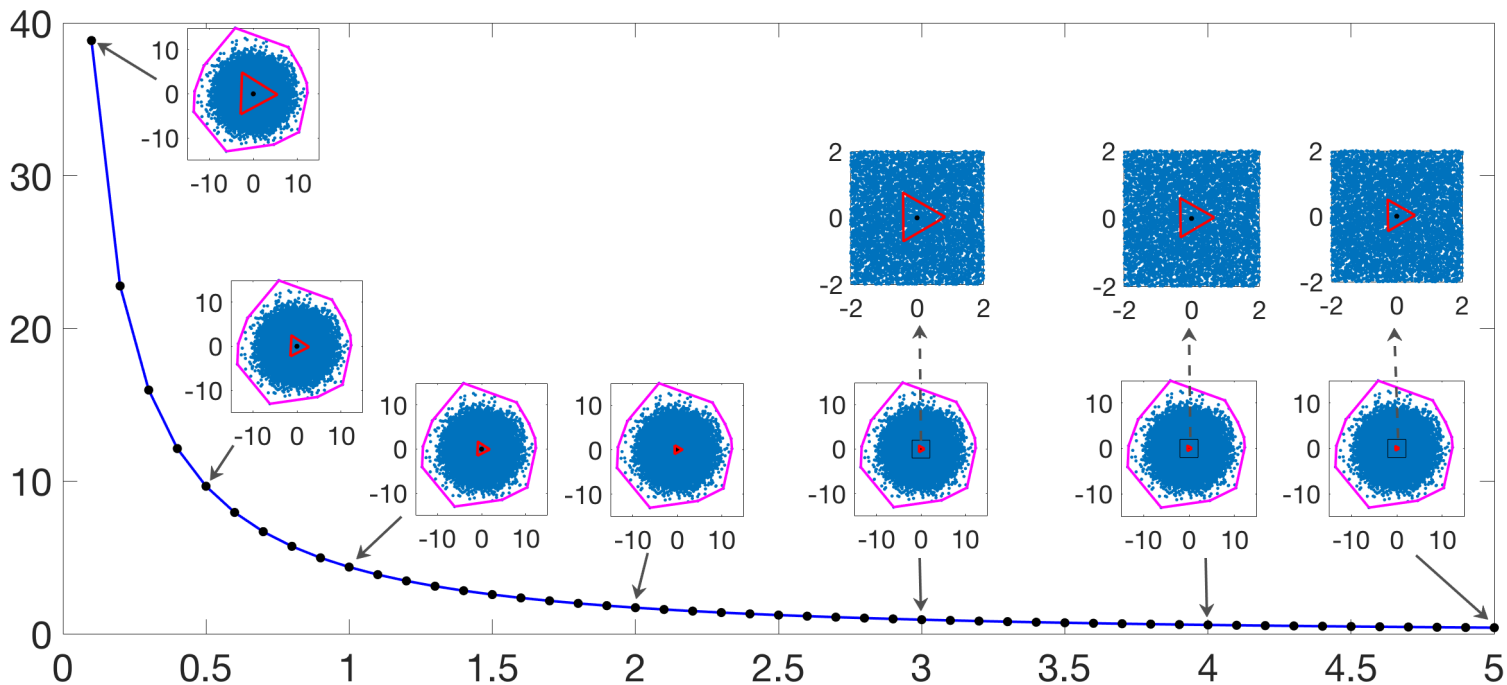


Figure: Variance-regularized AA applied to a dataset with increasing parameter α. The x-axis is the parameter α, and the y-axis is the area of the convex hull of the archetypes.

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.

Future directions:

- Consider the modified Archetypal Analysis by replacing the objective function by $F(A) = W_2(\mu_N, |A|^{-1}\mathbb{I}_A dx)$, where W_2 is the 2-Wasserstein distance.
- Investigate deep archetypal analysis [van Dijk et. al, CoRR, 2019; Keller et al., Pattern Recognition, 2019; Keller et al., Arxiv., 2020].

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


A. Yuan, J. Calder, and B. Osting, A continuum limit for the pagerank algorithm, to appear in *European Journal of Applied Mathematics* (2021). http://arxiv.org/abs/2001.08973


T. H. Reeb and B. Osting, Consistency of Dirichlet Partitions, *SIAM Journal on Mathematical Analysis* (2017). https://arxiv.org/abs/1708.05472

Thanks to support by NSF DMS 17-52202.

Future direction: Wasserstein metric-based Archetypal Analysis — joint work with Katy Craig, Dong Wang, and Yiming Xu

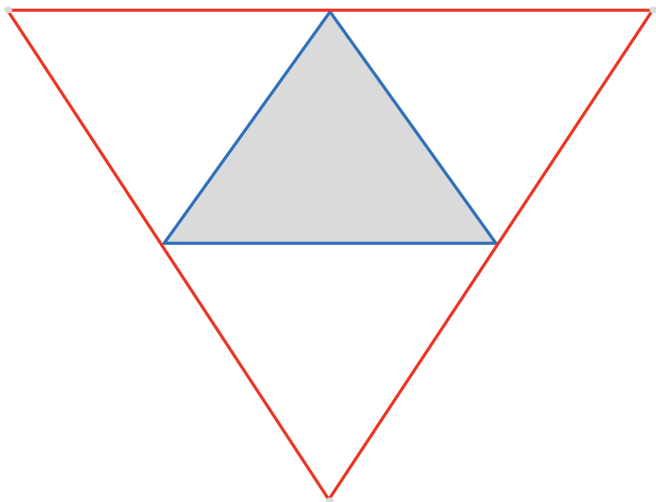
Given a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ and an integer $k \geq d+1$, we consider

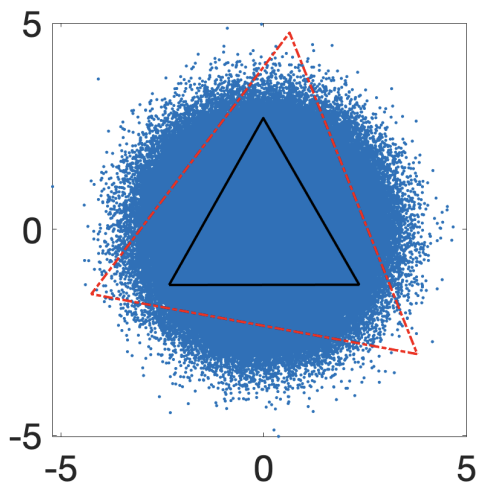
$$\Omega_* = \min_{\Omega \in S_k} W_2\left(\mu, |\Omega|^{-1}\mathbb{I}_\Omega dx\right),$$

where W_2 denotes the 2-Wasserstein metric, $\frac{1}{|\Omega|}1_\Omega$ denotes the uniform distribution on Ω, and

$$S_k = \{\Omega \subset \mathbb{R}^d : \Omega \text{ is a convex polytope with at most } k \text{ vertices and nonempty interior}\}.$$

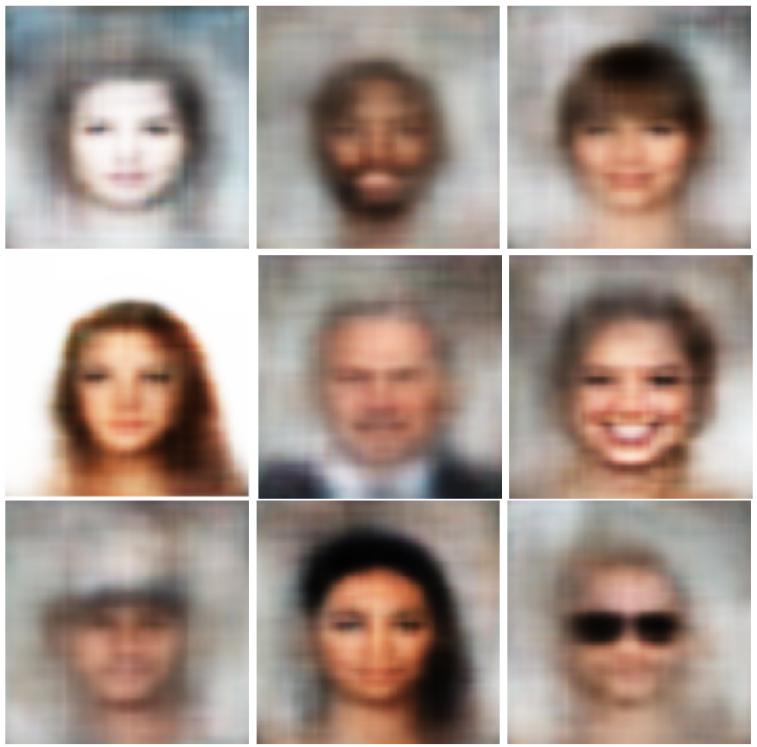
- Note: it is no longer necessary to require that $\Omega \subset \mathrm{co}(\mathrm{supp}(\mu))$
- This problem can be solved in small dimensions by reformulating it as a semi-discrete optimal transport problem
- We expect (and empirically observe) that this problem is less sensitive to outliers. In the examples below, the archetypal triangle for the Wasserstein metric (blue and black) are less sensitive to outliers than the archetypal triangle for the Euclidean metric (red).





Future direction: deep archetypal analysis — joint work with Lam Nguyen, Darshan Shimpi, Tanner Sims, Grace Siu, RK Yoon, Rich Medina

- Here, an autoencoder is designed so that the encoded points (in the latent space) lie within a (fixed) archetypal polytope.
- This reduces the dimension of the data while provides more interpretability of the encoding.
- SRI undergraduate student project
- Beginning to use these tools to create interpretable latent spaces for automatically studying online hate speech.



Example: A selection of archetype points from a 53 dimensional latent space for the Celeb A dataset.