Coresets for Archetypal Analysis

Sebastian Mair and Ulf Brefeld

Leuphana University of Lüneburg, Germany

Coresets

- coresets represent large data sets by weighted subsets
- on which models perform provably competitive compared to operations on all data

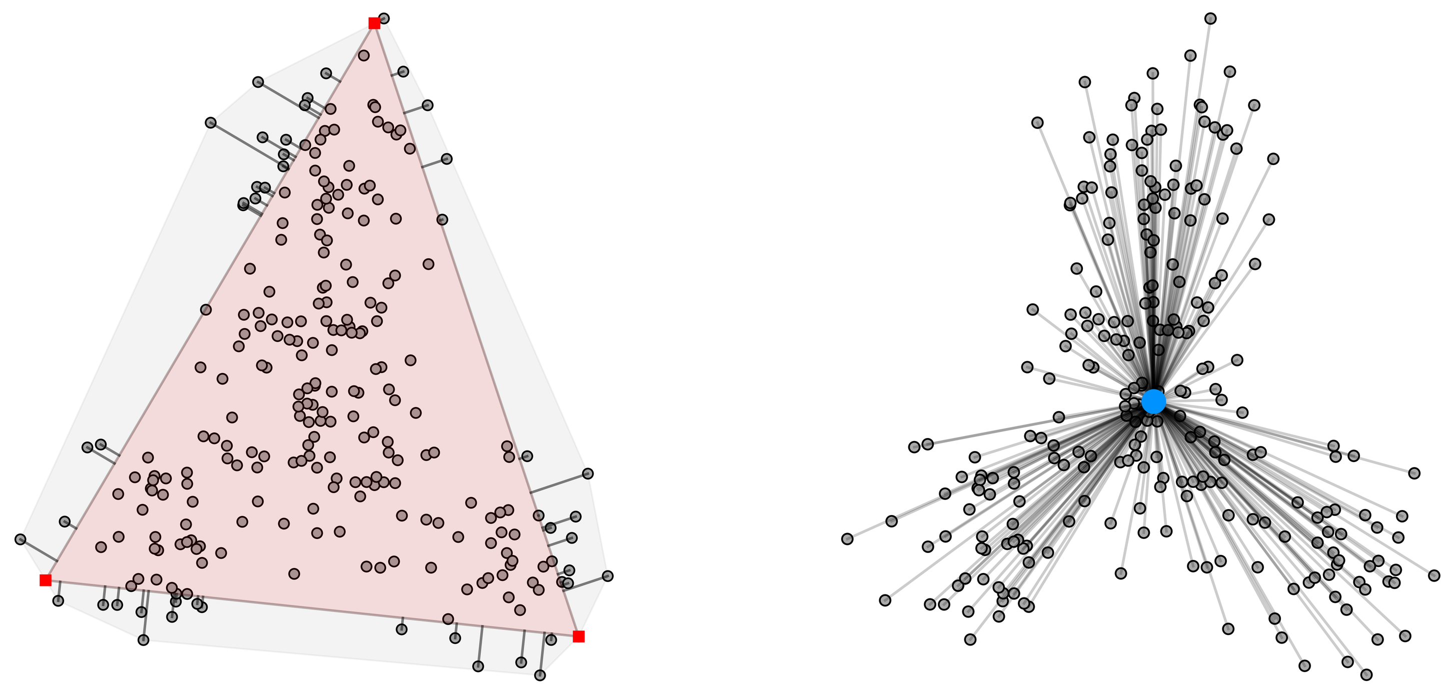
Archetypal Analysis (AA)

- AA is a variant of a interpretable matrix factorization
- factorize data $\mathbf{X}$ into convex weights $\mathbf{A}$ and archetypes $\mathbf{Z}$

$$\mathbf{X} = \mathbf{A}\mathbf{B}\mathbf{X} = \mathbf{A}\mathbf{Z}$$

- represent data points as a convex comb. of k archetypes
- represent archetypes as a convex combination of data
- as a result, the archetypes $\mathbf{Z}$ will be on the boundary of data
- find $\mathbf{A}$ and $\mathbf{B}$ by minimizing the residual sum of squares

$$\min \ \mathrm{RSS}(k) = \|\mathbf{X} - \mathbf{A}\mathbf{Z}\|_F^2 = \|\mathbf{X} - \mathbf{A}\mathbf{B}\mathbf{X}\|_F^2$$



Coresets for AA

- we propose to use the following sampling distribution

$$q(\mathbf{x}) = \frac{d(\mathbf{x},\mu)^2}{\sum_{i=1}^n d(\mathbf{x}_i,\mu)^2}$$

- sample a subset $\mathcal{C}$ of data of size at least

$$m \geq c\varepsilon^{-2}(dk \log k + \log \delta^{-1})$$

- assign the following weight to each sampled point

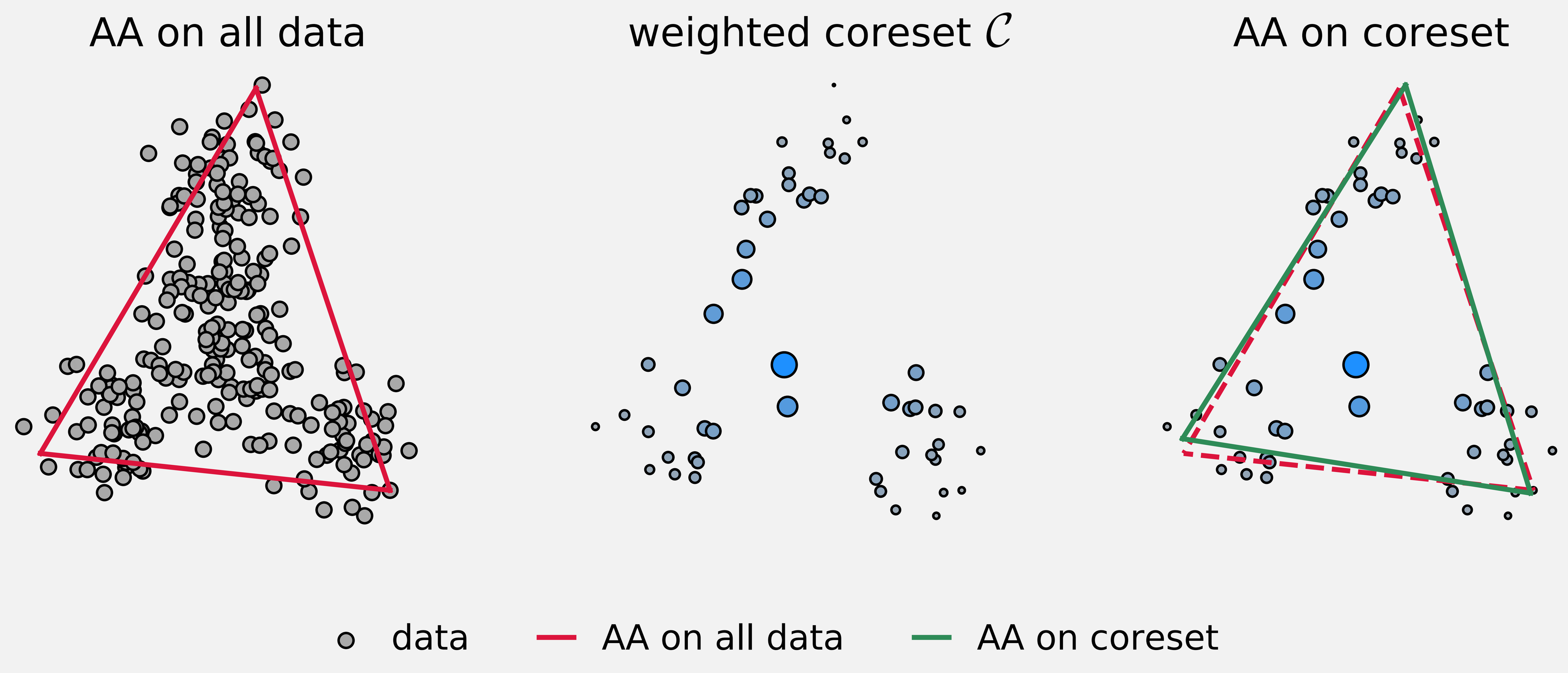
$$(m \cdot q(\mathbf{x}))^{-1}$$

- the following bound holds with probability of at least $1 - \delta$

$$|\phi_{\mathcal{X}}(Q) - \phi_{\mathcal{C}}(Q)| \leq \varepsilon \phi_{\mathcal{X}}(\{\mu\})$$

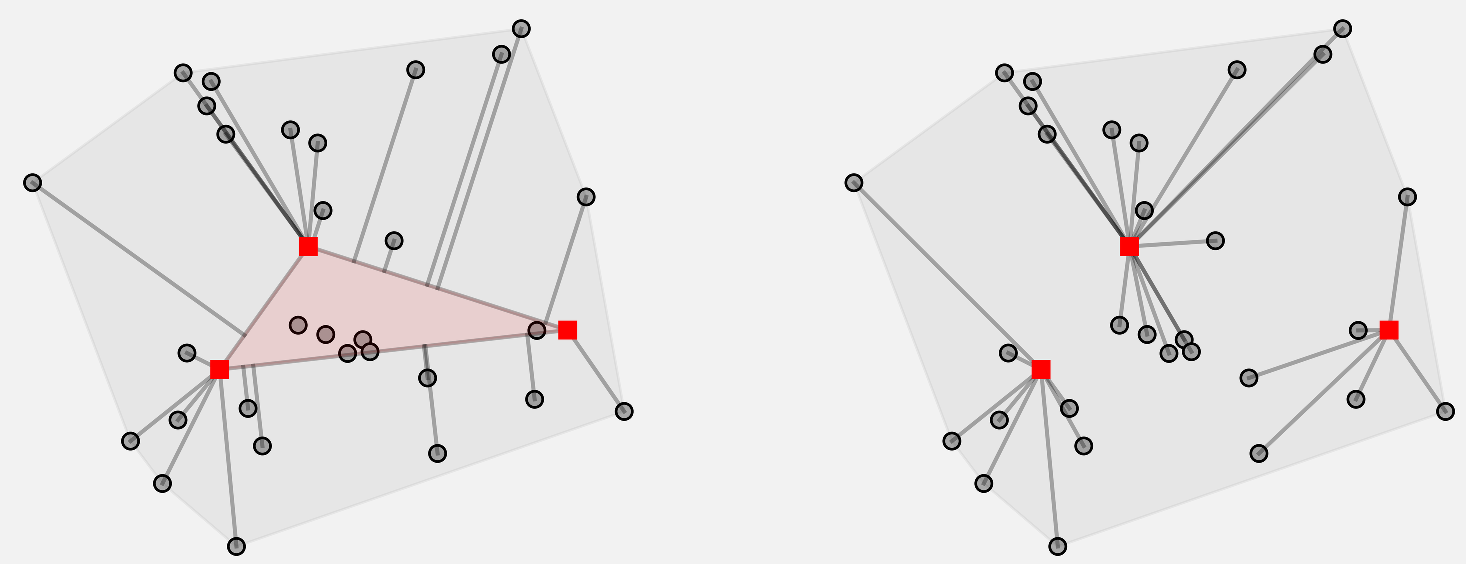
for any query $Q \subset \mathbb{R}^d$ of cardinality at most k satisfying

$$\mu \in \mathrm{conv}(Q)$$



Summary

- AA is an interpretable matrix factorization
- however, the interpretability comes with high computational cost
- we propose efficient coresets for scaling up archetypal analysis

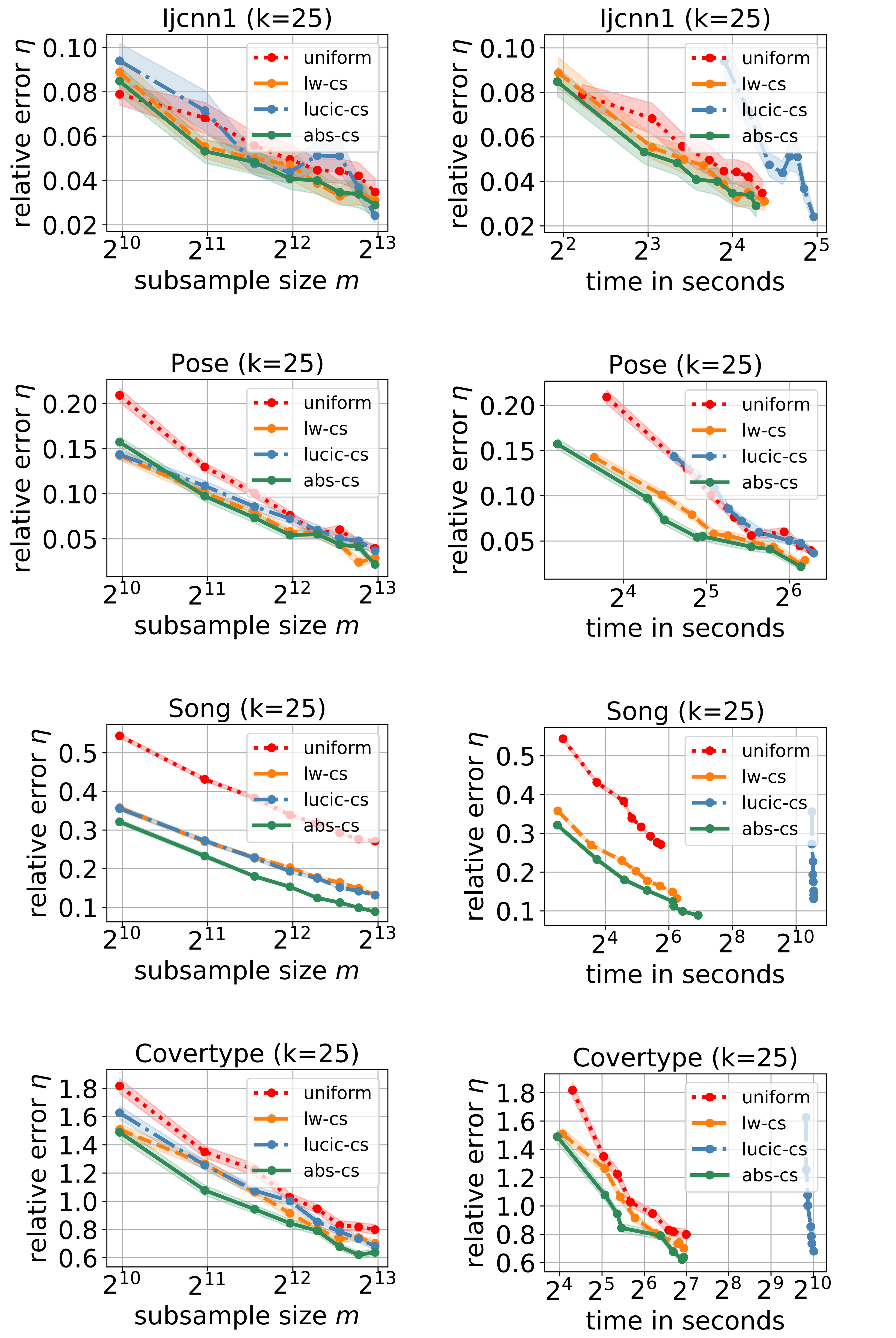


Contributions

- every coreset for k-means is also a coreset for archetypal analysis
- a simple and efficient sampling strategy to compute a coreset
- rigoros theoretical analysis of the derived coreset
- empricial evaluation support the theoretical derivation

LEUPHANA
UNIVERSITY OF LÜNEBURG

Experiments



- uniform sampling performs consistently worse than its peers
- our coreset often yields the best results
- consistently lower relative errors in shorter time