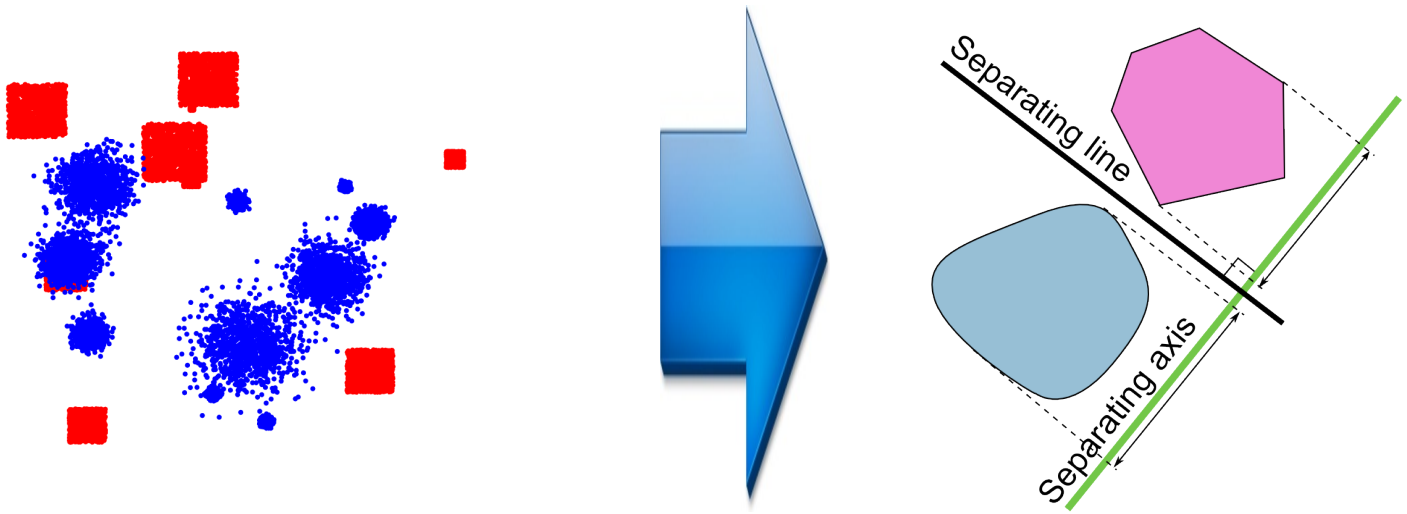
Relevance in Experimental Setting

- Classification model

$$f(h_{\#}\mu) = f(\mu) \text{ for } h \in \mathcal{G}_{\varepsilon,R}$$

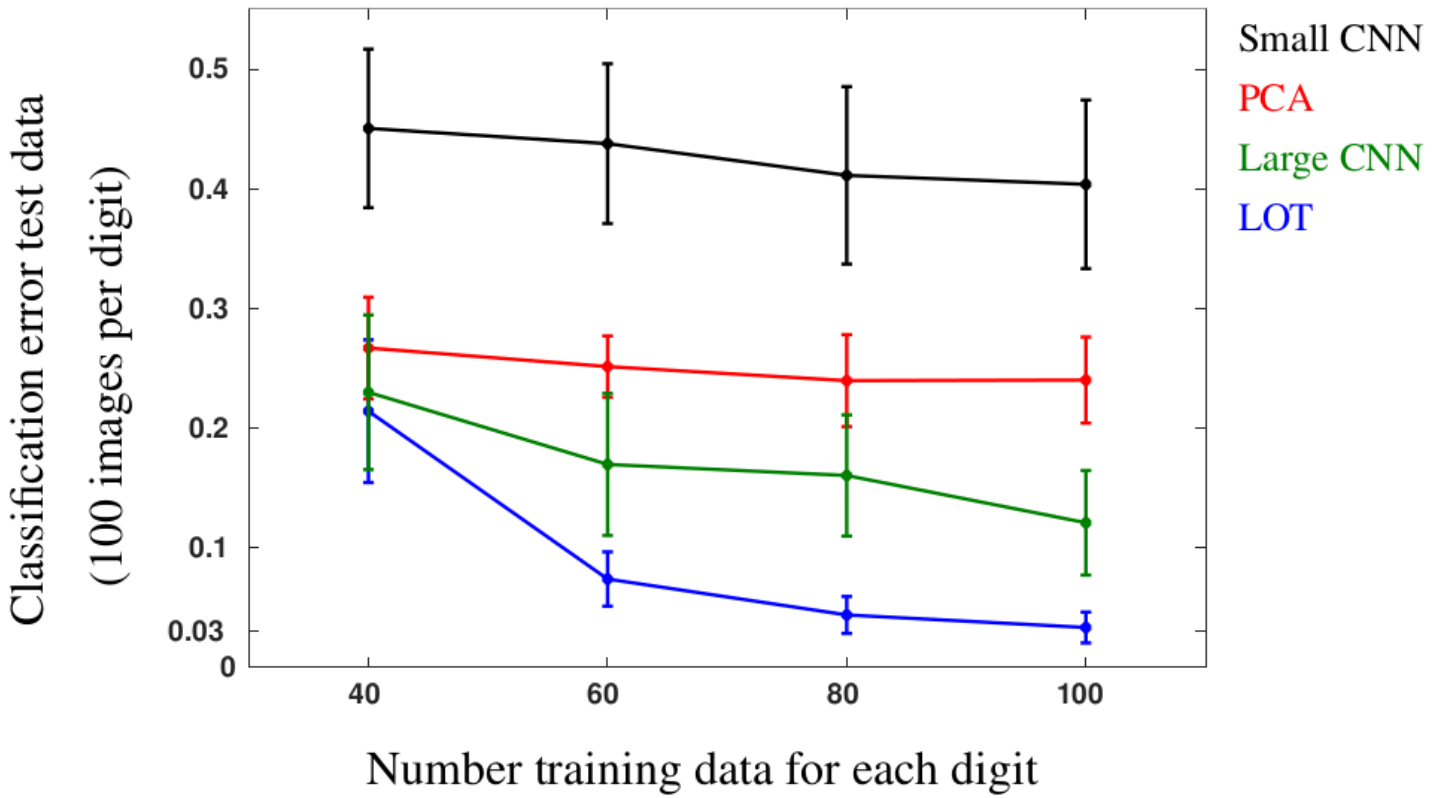
- Hyperplane classifier f_θ between $\mathcal{H} \star \mu$ and $\mathcal{H} \star \nu$
  - Sample σ to project T_σ^μ to T_Z^X
  - Reduce dimension through:
    - PCA if data on shared computer
    - JL-embedding if data on distributed computers
- **Note:** Equivalent approximation results for deterministic Voronoi cells (Hamm, Khurana 2024)



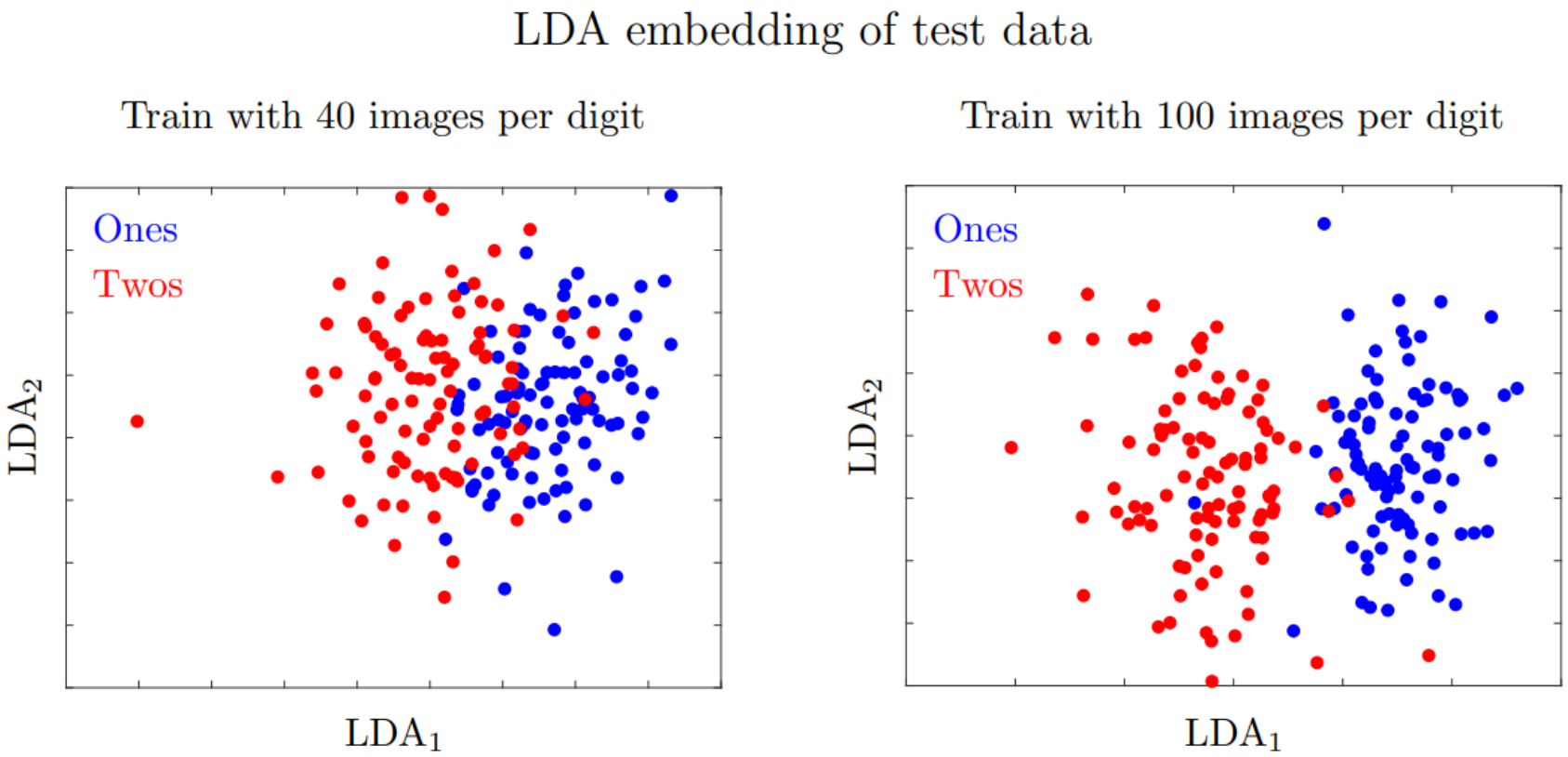
Experimental Validation (shifts and scalings)

Modified MNIST Classification Between 1's and 2's

- Data sampled from MNIST images
- Each image additionally augmented by random shift and scaling
- Sample k labeled examples of each class for training
- σ is centered normal distribution

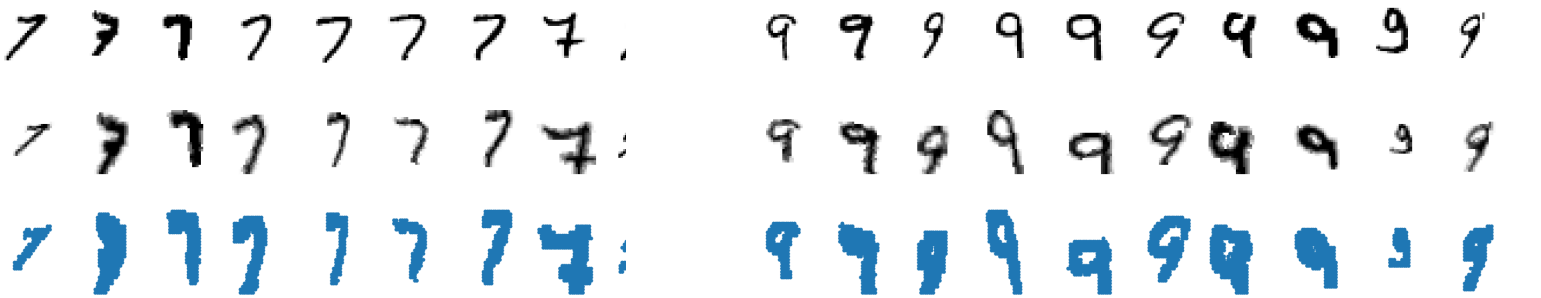


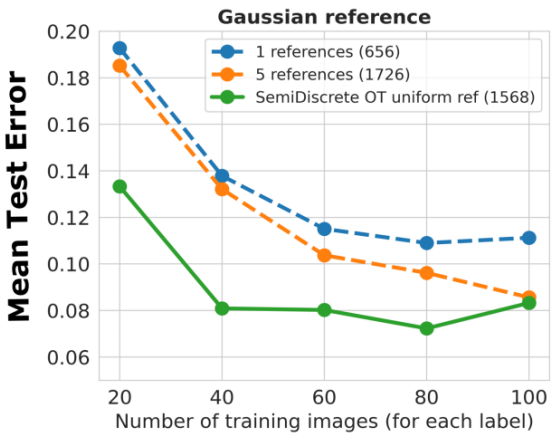
Experimental Validation (shifts and scalings)

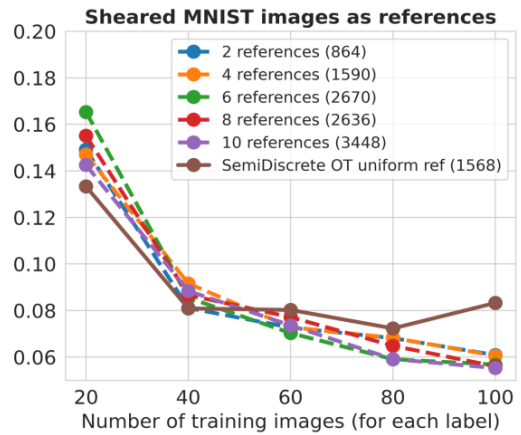


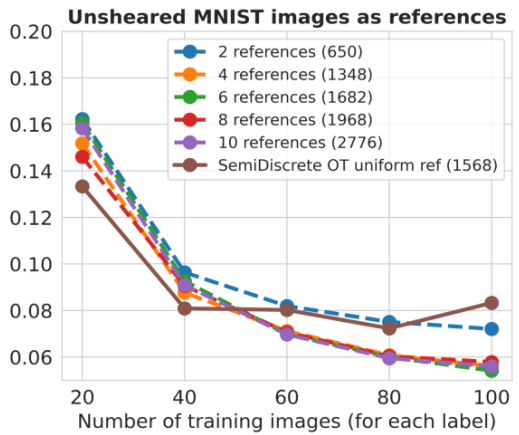
Experimental Validation (shearing)

Modified MNIST Classification Between 7's and 9's





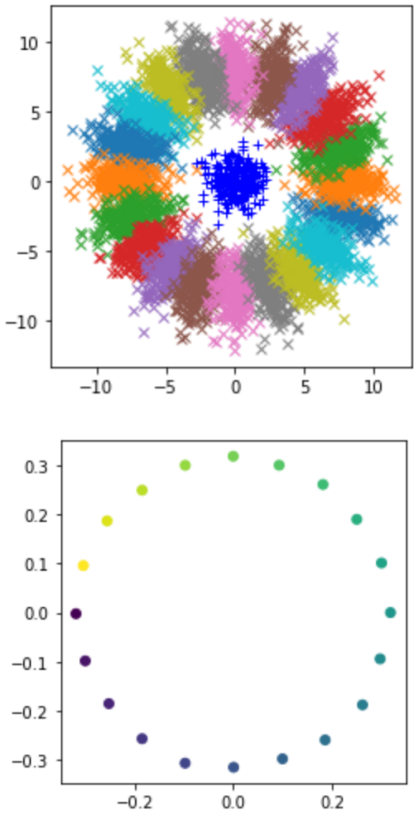




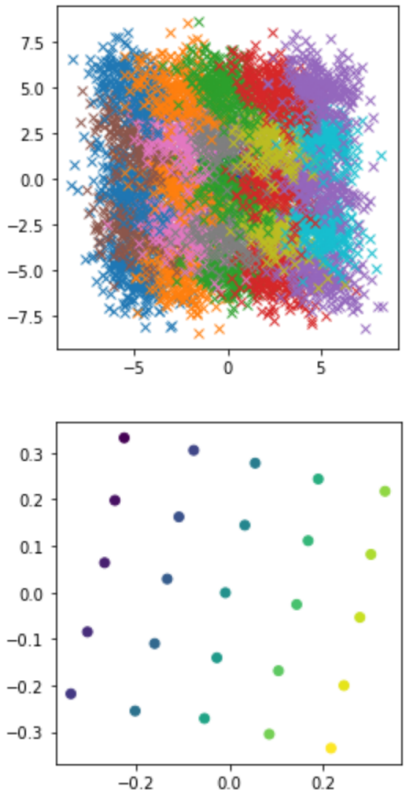
PCA Embedding

- $\text{supp}(\mu_k) \subset \mathbb{R}^2$
- $\mathcal{H}$ made up of almost compatible transformations

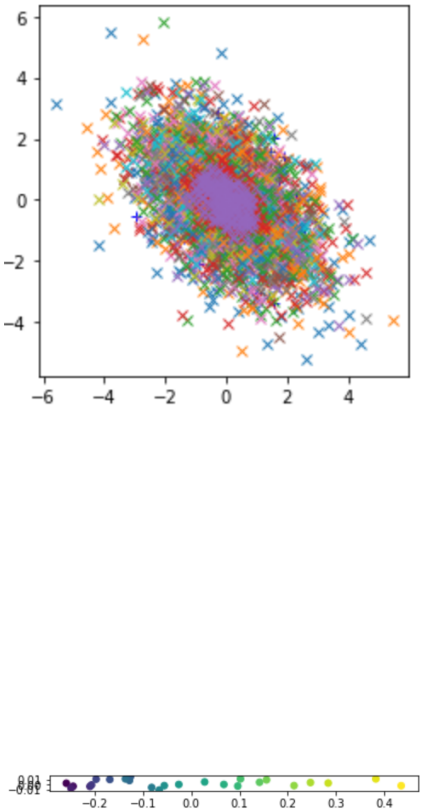
Shifts/Shears



Shifts



Global Scaling



Analysis of Lemuridae Teeth

Scans of teeth of different species of lemur

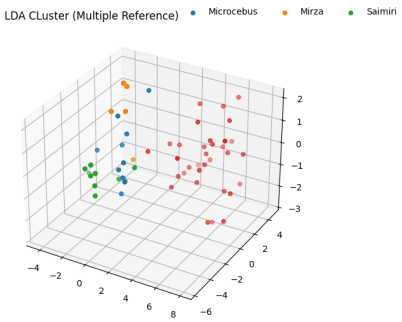
- 3D point clouds, oriented correctly, no point-to-point registration
- Classic ML questions: classification, clustering, data generation
- 4 classes, 5-12 examples per class
- Reference of either Gaussian or multi-reference of LOT barycenters for each class (only calculated on training set)

Classification: Linear SVM: 100% accuracy on test set

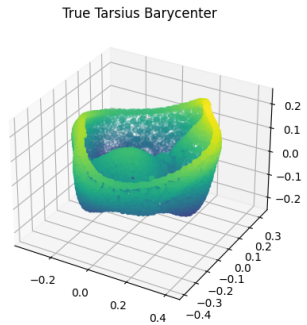
Dimension Reduction: LDA (multi-reference)

Data Generation: Random barycenters per class $\approx$ 1% relative error to true barycenter (multi-reference)

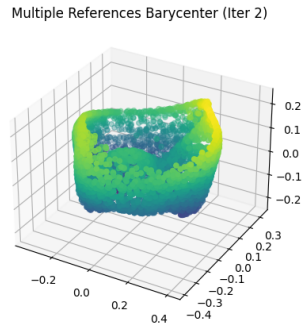
- Approximately 1000$\times$ faster



LDA



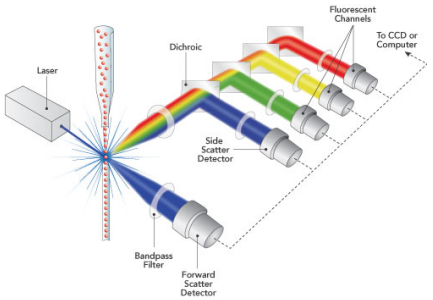
True Tarsius barycenter

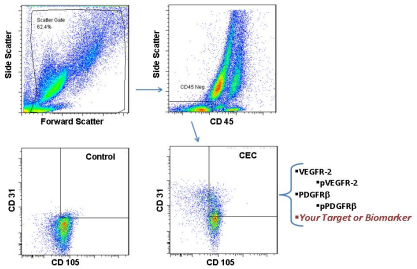


LOT Tarsius barycenter

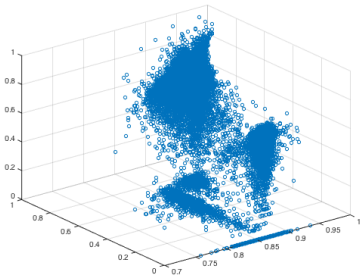
Flow Cytometry - Preliminary Results

- Flow cytometry: each patient is represented by 9D point cloud of cells

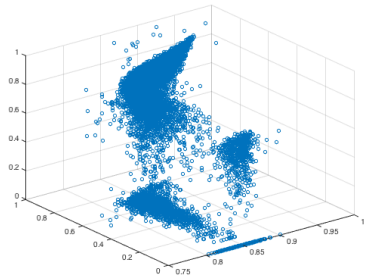




- Used to tell if people have blood disease
  - Medical test is to look at every 2D slice



Healthy



AML

Flow Cytometry Distributions

- Detecting AML: $n_{sick} = 43$, $n_{healthy} = 315$
- Comparison to Deep Set
- Binary cross-entropy loss
- Easy to do bagging / resampling of σ to build ensemble classifier

	DeepSets	DS Bagging	LOT Classifier	LOT Ensemble
precision	0.17	0.23	0.67	**0.67**
recall	0.33	0.44	0.44	**0.67**
accuracy	0.84	0.86	0.95	**0.96**

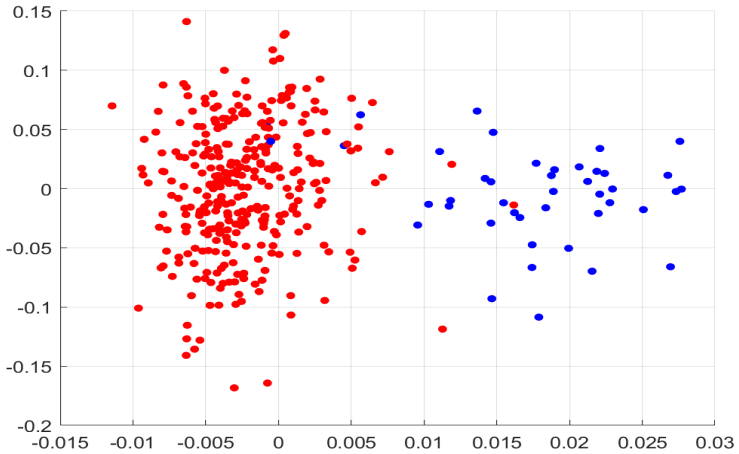
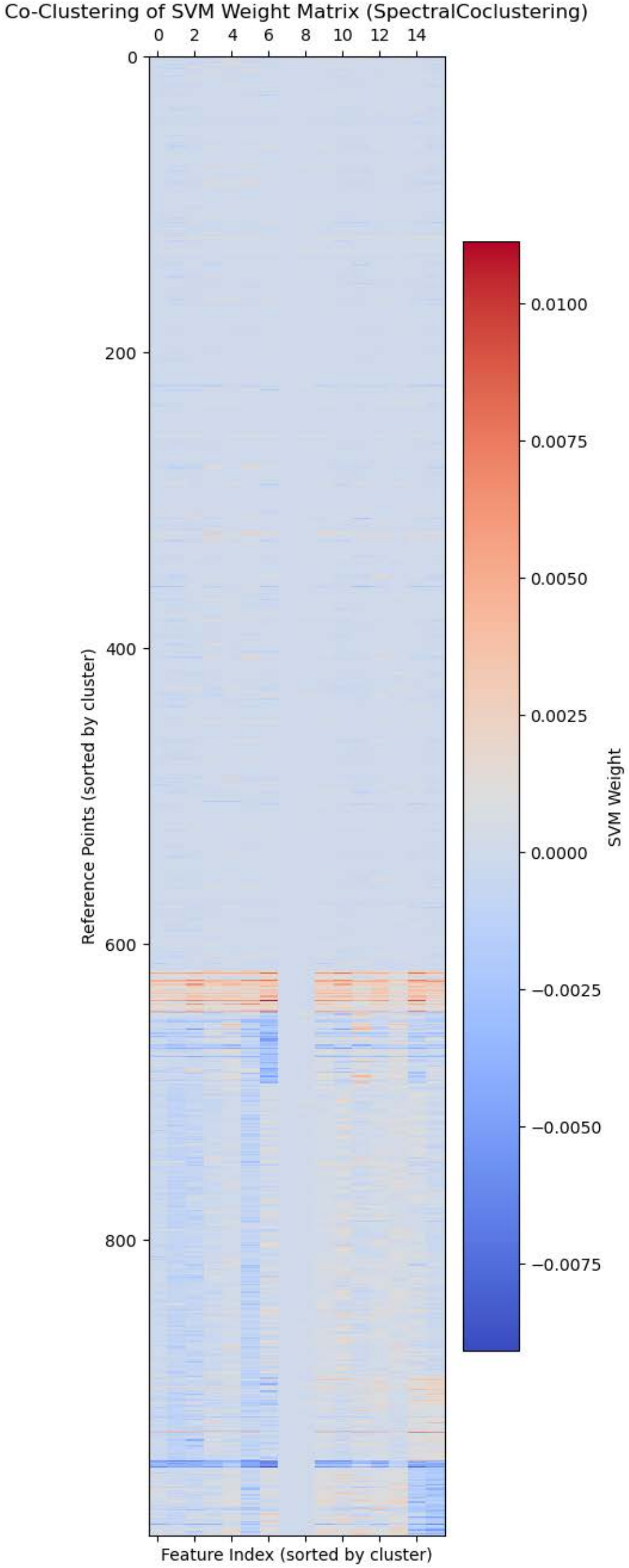


Figure: LDA Embedding


Co-Clustering of SVM Weight Matrix (SpectralCoclustering)
0
2
4
6
8
10
12
14
0
200
400
600
800
Reference Points (sorted by cluster)
Feature Index (sorted by cluster)
0.0100
0.0075
0.0050
0.0025
0.0000
−0.0025
−0.0050
−0.0075
SVM Weight

Shameless Plug for Conference

- **What:** Stand-alone conference at UCSD
- **When:** December 1-2, 2025 (just before NeurIPS in San Diego)
- **Includes:**
  - Paper submission / review / proceedings in PMLR
  - Keynotes / talks / parallel sessions
  - Posters / student presentations
- **Workshop Website:**
  `https://www.tagds.com/events/tag-ds-2025`