Geometric Disentanglement by Random Polytopes

arXiv:2009.13987

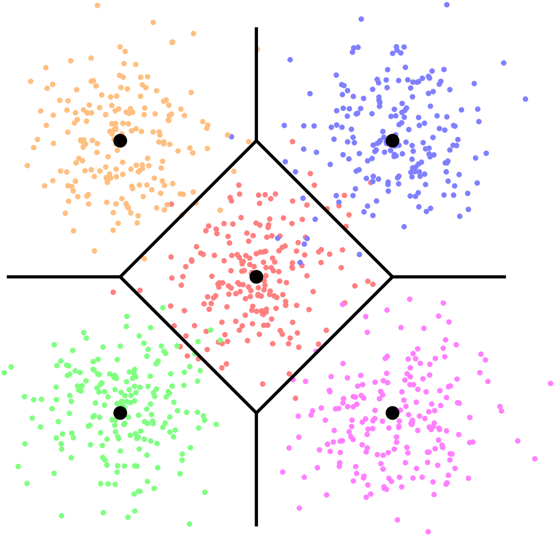
Michael Joswig

TU Berlin & MPI-MIS Leipzig

Leipzig, 06 April 2021

joint w/ Marek Kaluba
Lukas Ruff

Geometric Disentanglement of “The Data”



- What is a good description of the data?
  - do we see like five clusters?
- Suppose this is labeled training data
- *k*-means clustering
  - Voronoi diagram

Outline

1. Random Polytopes
   - Random polytope descriptors
   - Scaling distance and anomaly scores
2. Experiments
   - Standardized data sets
   - (Variational) autoencoder neural networks
   - Out of distribution attacks

Random Polytope Descriptors

Let $X \subset \mathbb{R}^d$ be finite, with $N := |X|$.

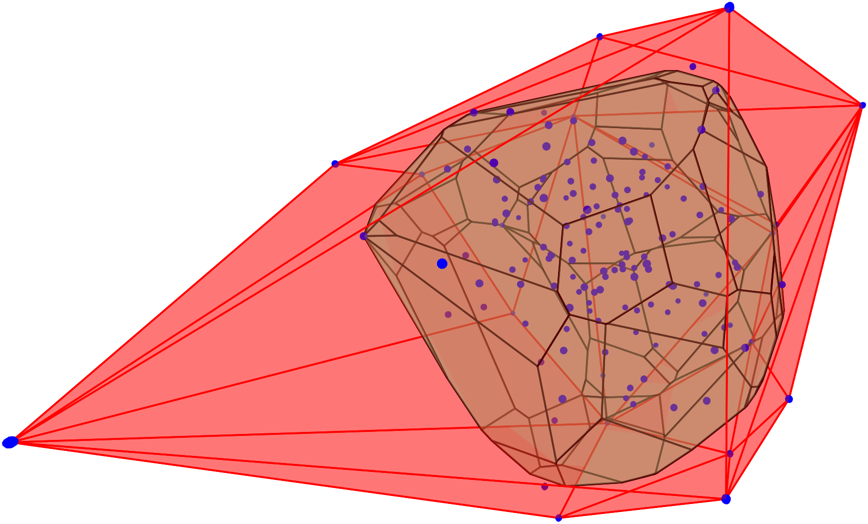
- pick a set $Y \subset \mathbb{S}^{d-1}$ of *directions* uniformly at random, with $m := |Y|$, and let ℓ be a positive integer
- the Random Polytope Descriptor (RPD) is the polyhedron

$$\mathrm{RPD}_{m,\ell}(X) \ := \ \left\{ v \in \mathbb{R}^d \ \middle| \ \langle v, y \rangle \leq \ell\text{-}\max_{x \in X} \langle x, y \rangle, \ y \in Y \right\} \ ,$$

  where ℓ- max $= \ell$-th largest scalar product
- assume: $\mathrm{RPD}_{m,\ell}(X)$ is bounded $\iff$ Y positively spanning
- $\mathrm{RPD}_{m,1}(X) = D_Y(X)$ *dual bounding body*

Example

$N = 150,\ d = 3,\ m = 60,\ \ell = 3$



Controlled Combinatorics

Proposition (J., Kaluba & Ruff 2020+)

Let $X \sim N(0, I)$ be normally distributed with mean zero (and identity covariance matrix).
Then, for arbitrary m and ℓ, the number of vertices of $\mathrm{RPD}_{m,\ell}(X)$ is of order $\Theta(m)$, with high probability, for d constant.

Proof.

- normal distribution and uniform choice of directions $Y \subset \mathbb{S}^{d-1}$ rotationally invariant.
- $\mathrm{RPD}_{m,\ell}(X)$ follows the *Rotation-Symmetry Model (RSM)* of Borgwardt (1987), with m inequalities

$\square$

- similar result for $X \sim \mathbb{S}^{d-1}$

Scaling Distance and Anomaly Scores

Let $P \subset \mathbb{R}^d$ be a d-polytope with m facets and interior point c.

- *scaling distance* of $x \in \mathbb{R}^d$ to P with respect to c:

$$\mathrm{sd}_c(x, P) \ := \ \min \{ \alpha \geq 0 \mid x \in \alpha(P - c) + c \}$$

- $x \in P \iff \mathrm{sd}_c(x, P) \leq 1$

- several natural choices for *central point* c
  - centroid = center of gravity
    - Rademacher (2007): hard to approx.
  - vertex barycenter
    - Elbassoni & Tiwari (2009): hard to compute exactly
  - Chebyshev center = center of largest sphere inscribed
    - Eaves & Freund (1992); Renegar (1988): $O(\sqrt{m})$ by LP

Main Theoretical Result

Theorem (J., Kaluba & Ruff 2020+)

Let $P = \operatorname{conv}(X) \sim P(d,n)$, *and let* $Y \subset \mathbb{S}^{d-1}$ *be a set of* m *directions chosen uniformly at random. Fix* $\epsilon > 0$ *and* $0 < p < 1$.
The following holds almost surely for $m \to \infty$: *the mean of* s *randomly chosen vertices of* $D_Y(X)$ *is at distance* $\leq \epsilon$ *from the origin with probability at least* $1-p$ *if*

$$s > \left(1 + \frac{2}{d}\log\left(\frac{2}{p}\right)\frac{e}{e-1}\cdot\frac{1}{\epsilon^2(1-h_0)^2}\right) ,$$

where h_0 *is the Hausdorff distance of* P *to the sphere* $\mathbb{S}^{d-1}$.

- proof uses results of Newman (2020)
- sampling a few vertices from $D_Y(X)$ to approximate vertex barycenter superior to computing Chebychev center in practice

MNIST & Fashion-MNIST

Modified National Institute of Standards and Technology datasets

MNIST

- 10 labeled classes of handwritten digits
- 60,000 grayscale images with 28×28 pixels

FMNIST

- Zalando's article images
- same paramaters as MNIST

false positives for AE/k-means

Autoencoder Neural Networks

dimensionality reduction / feature learning / unsupervised learning

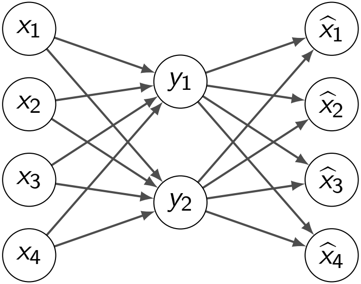
Given finitely many data points $X \subset \mathbb{R}^n$ and *latent dimension* d, find

- *encoder* $\phi : \mathbb{R}^n \to \mathbb{R}^d$ and
- *decoder* $\psi : \mathbb{R}^d \to \mathbb{R}^n$

such that

$$\| X - \psi(\phi(X)) \|$$

is minimal.



$n = 4$ and $d = 2$

For instance, $\phi = \sigma(Wx + b)$ and $\psi = \sigma(\widehat{W}y + \widehat{b})$, where $\sigma(x) = \frac{1}{1+e^{-x}}$.
In that case: find $W, b, \widehat{W}, \widehat{b}$.

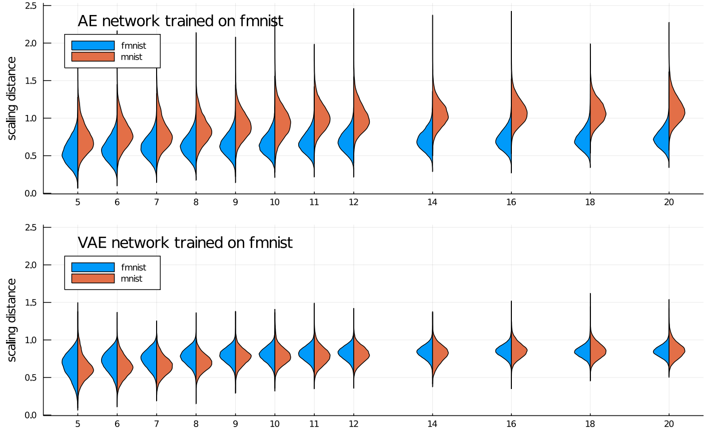
Comparison of RPD vs k-means on MNIST

- AUC score
  - Area Under Receiver Operating Characteristic Curve
- $N = 6000,\ d = 16,\ m = 640,\ \ell = 1$

	AE		VAE	
Class	RPD	k-means	RPD	k-means
0	**99.6**	99.3	**99.2**	96.1
1	**99.8**	99.5	**99.8**	98.6
2	**94.6**	94.0	**96.3**	93.4
3	95.8	**95.9**	**96.2**	95.5
4	**95.4**	92.3	**98.0**	94.3
5	**93.4**	88.8	**96.0**	90.3
6	**98.3**	94.4	**99.4**	95.8
7	**95.7**	94.4	**97.0**	95.2
8	**96.6**	94.0	**96.4**	92.4
9	**97.5**	95.1	**97.6**	93.5

Out of Distribution Detection

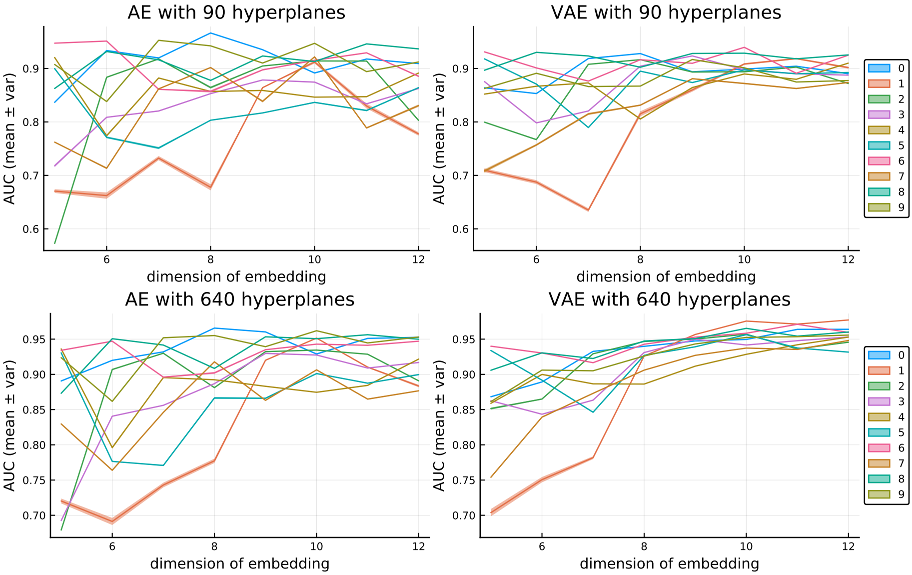
AE/VAE networks trained FMNIST data, used to embed MNIST



- distribution of minimal scaling distance to one of ten FMNIST RPD
  - various dimensions up to 20 for fixed $(m, \ell) = (640, 1)$
- 5 distinct AE and VAE networks trained and 5 RPD per dimension

Dependence on m and d

MNIST data AUC scores per class



Conclusion

- fast generic method for classification
  - e.g., computing H-description of $\mathrm{RPD}_{640,2}(X)$ for $N = 6000$, $d = 20$, Chebychev center, takes < 3s on Laptop (i5-6200U)
  - drops to < 1s with approximate vertex barycenter
  - evaluating scaling distance < 0.001s
- all polyhedral computations verifiable in polynomial time, by exact computations
  - useful for analyzing existing ML methods