Visualizing High-Dimensional Spaces

Basic Training in Condensed Matter
02/09/2024

Katherine Quinn
kq57@georgetown.edu

Outline

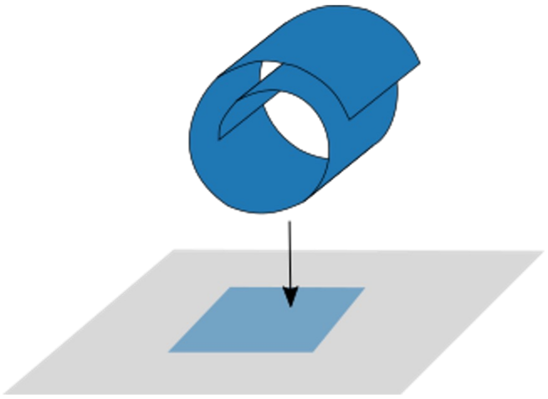
1) Standard visualization methods
2) Probabilistic models and data
3) Intensive Principal Component Analysis (InPCA)

Visualizing high-dimensional spaces is hard. We need to find an embedding space which captures our features of interest.

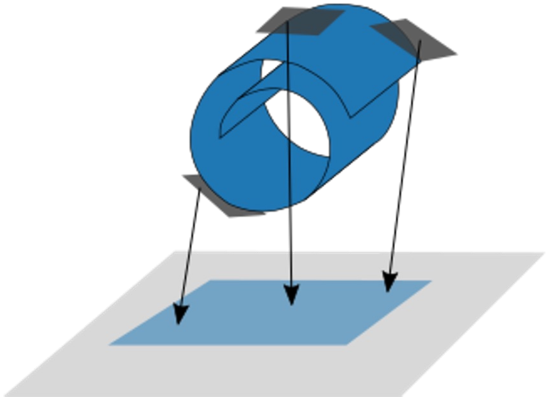
Visualizing High-Dimensional Spaces

Want to represent an n-dimensional space on an 2D plane in a way that keeps features of interest.

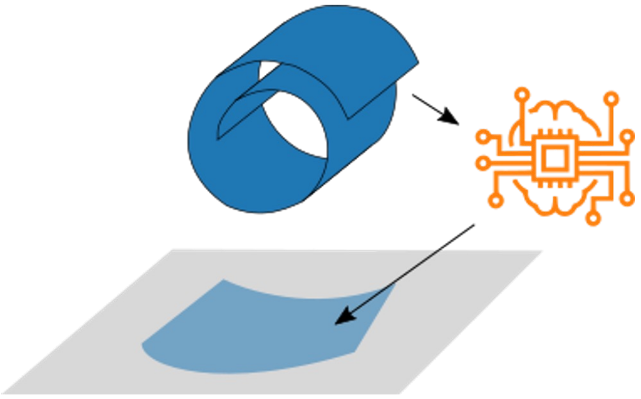
Global Projection



Local Projection

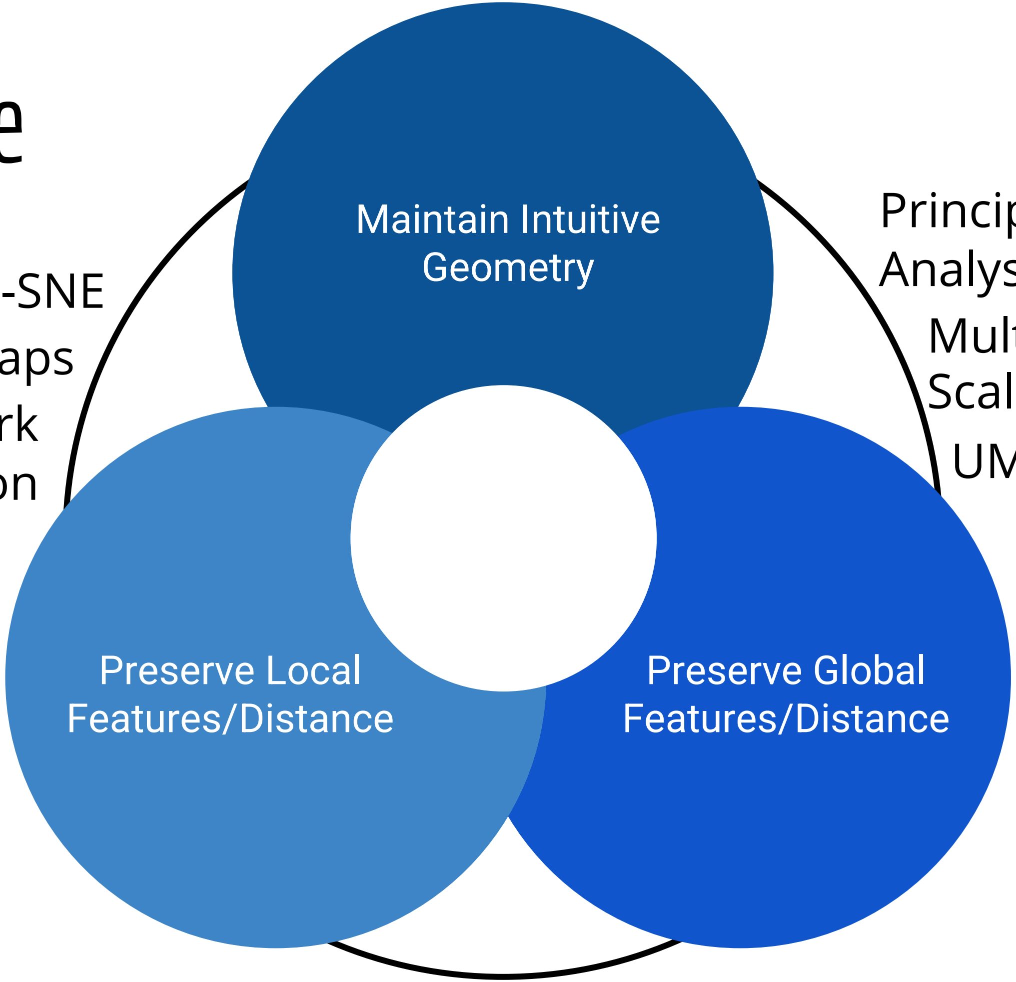


Manifold Learning



Compromise

t-SNE
Diffusion Maps
Graph/Network visualization



Principal Component Analysis (PCA)
Multidimensional Scaling (MDS)
UMAP

Example: Visualizing Text Documents

Patent Documents

Natural language processing with a large language model (LLM)

Visualize with PCA

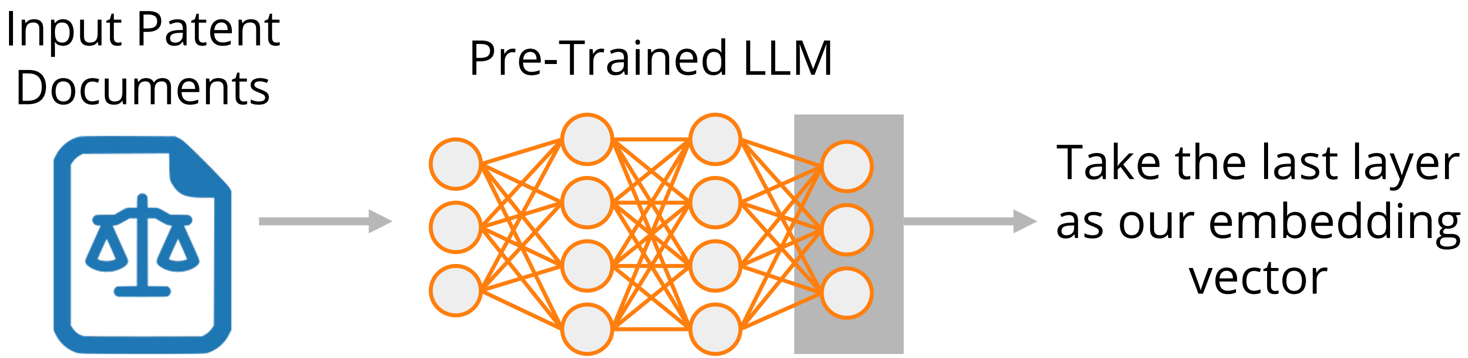
Scholarly Literature

Simple citation network

Visualize with igraph

Patents

- Feed each patent into a pre-trained neural network (*e.g.* SentenceTranformer).
- Take the 768-dimensional embedding space as the model outputs.
- Treat the embedding space as a simple, Euclidean space.



Patents with PCA

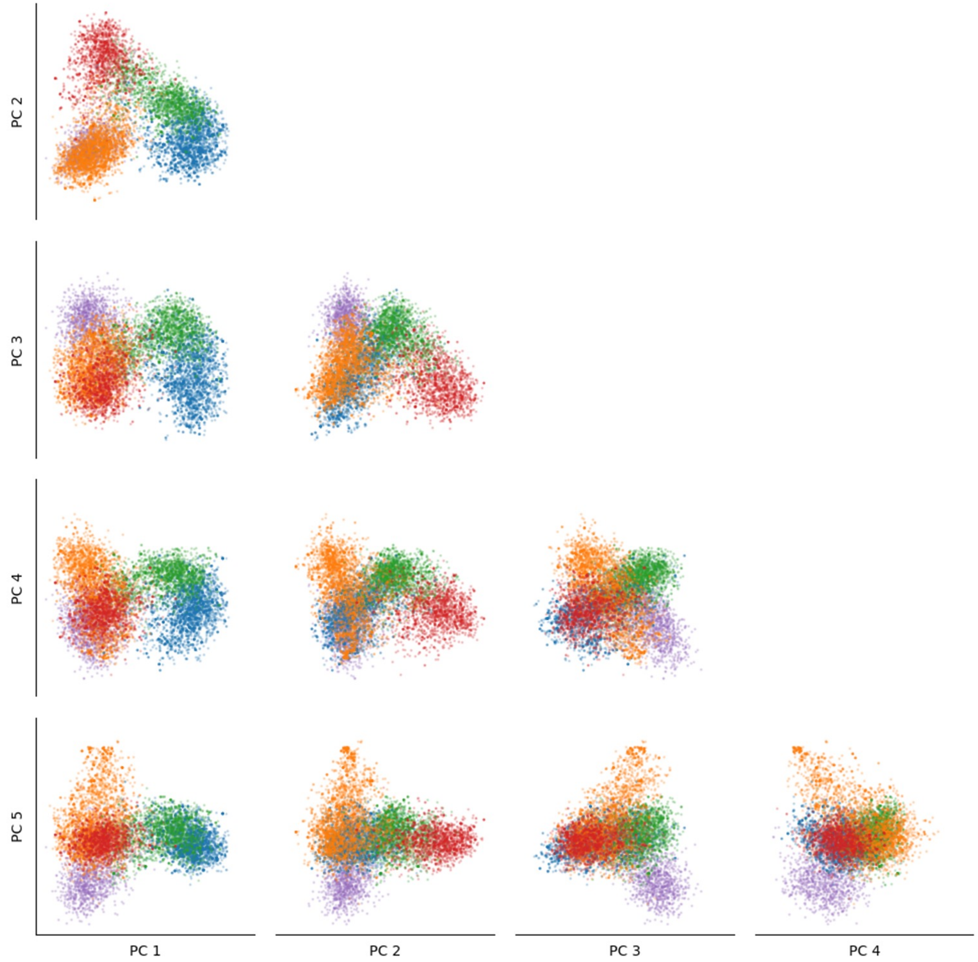
Takes orthogonal directions of maximal variance.

Viewing 11,000 patents in 768D space.

Five dominant components (out of 768).

Colors represent patent categories.

Biotechnology - blue
Telecommunications - orange
Food_and_tobacco - green
Mining_and_quarrying - red
Real_estate - purple



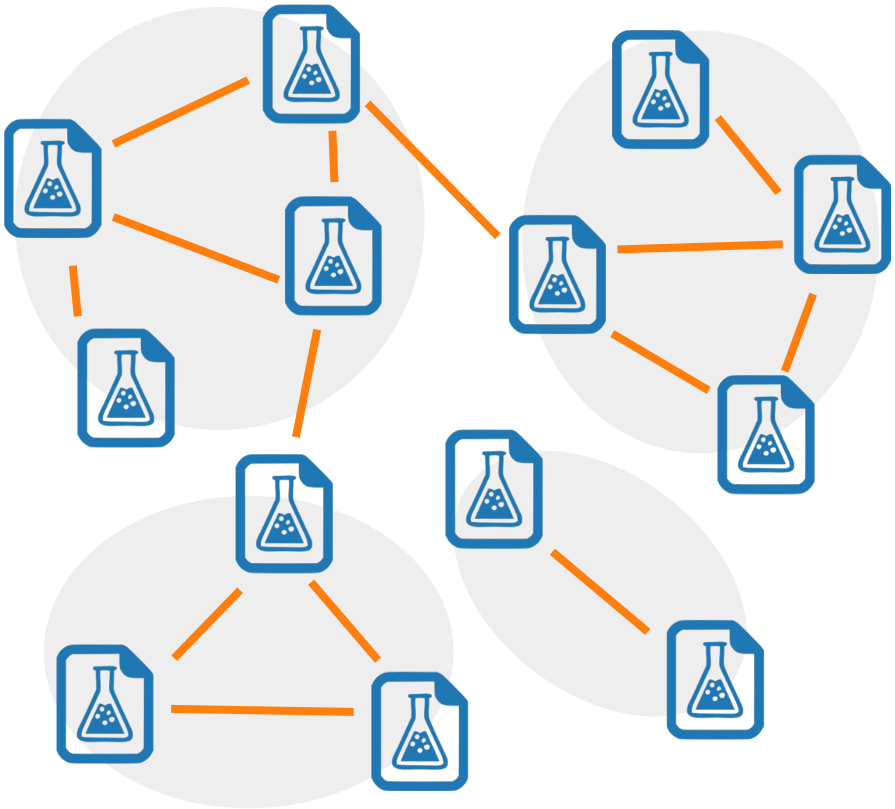
Scholarly Literature

Generate a citation-based network between articles in our merged academic corpus.

Weights are calculated using outgoing citation fractions.

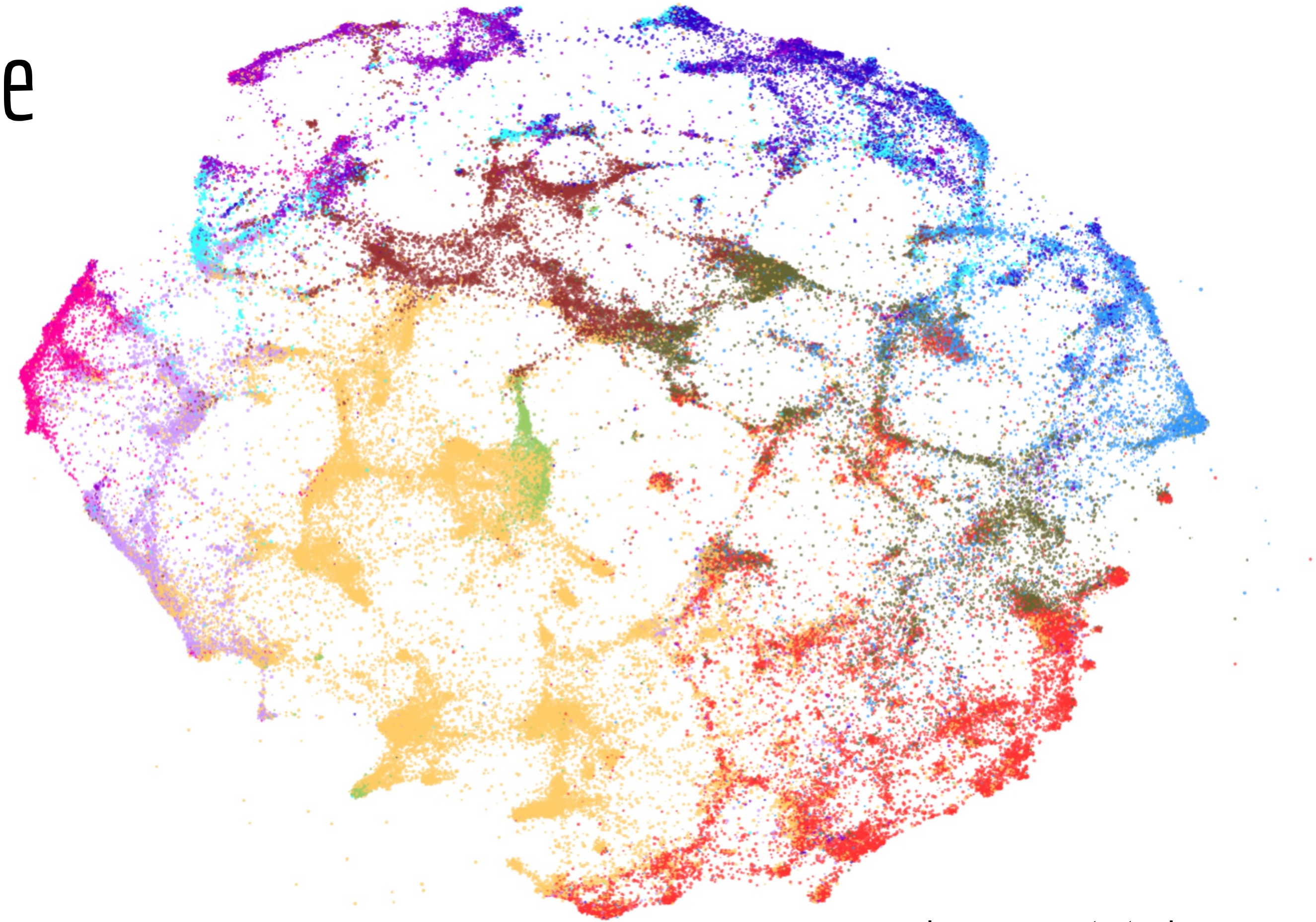
Cluster this network using the Leiden algorithm.

Visualize resulting clusters with igraph.



Map of Science

~85,000 Clusters



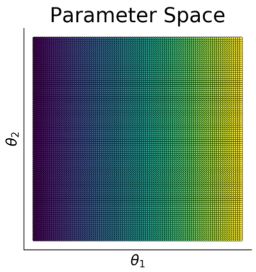
sciencemap.eto.tech

Distances and Embedding Spaces

How we create the low-dimensional visualization determines our *embedding space*, *i.e.* the space in which we view our model or data of interest (*e.g.* parameter space, prediction or behavior space, etc).

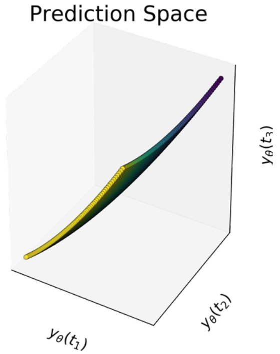
Distance measures can be impacted by the embedding space.

Parameter space may be highly distorted and have non-Euclidean distances.



Simple Model

$$y_\theta(t) = e^{-\theta_1 t} + e^{-\theta_2 t}$$



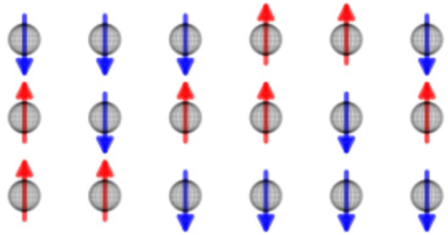
Model prediction space may have a simple Euclidean distance.

Probability Distributions

Any measurement with uncertainty can be seen as a probability distribution.

Many models have uncertainty, and produce model prediction that are probabilistic.

Ising Model

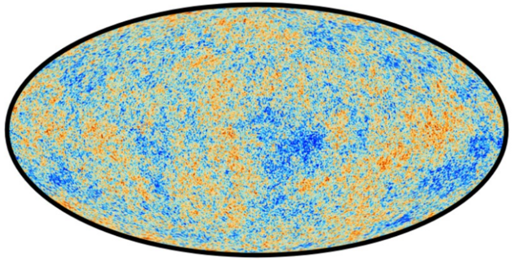


$\mathcal{L}(S \mid \theta)$

Simple ML Classifier

$\mathcal{L}(D \mid \theta)$

Cosmic Microwave Background



$\mathcal{L}(M \mid \theta)$

How to Visualize Collections of Distributions?

Embedding Space

Need a space with an intuitive geometry to visualize distributions.

Ideally something not too warped (will come back to this).

Want low-dimensional representations.

Distance

Need to measure how similar two distributions are from each other.

Divergences: Way of measuring difference between two probability distributions.

Intensive Principal Component Analysis (InPCA)

Combine two known techniques.

1. PCA
   - Extract orthogonal directions of maximal variance.
2. Replica Theory
   - Tune the dimensionality of the system by considering replicas, *i.e.* drawing multiple samples from the same distribution.

Resulting embedding space will be Minkowski-like (timelike and spacelike components).

Hellinger

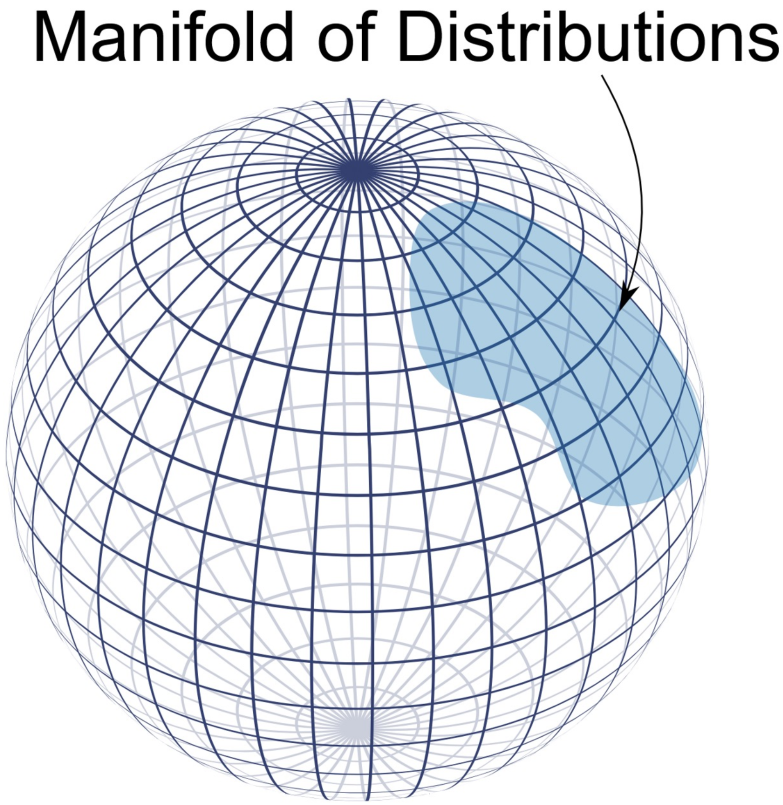
Probability distributions are normalized, so their square roots have length one.

The set of all probability distributions occupy part of the surface of a hypershere.

Euclidean distance, metric is FIM.

Distance is one minus dot product.

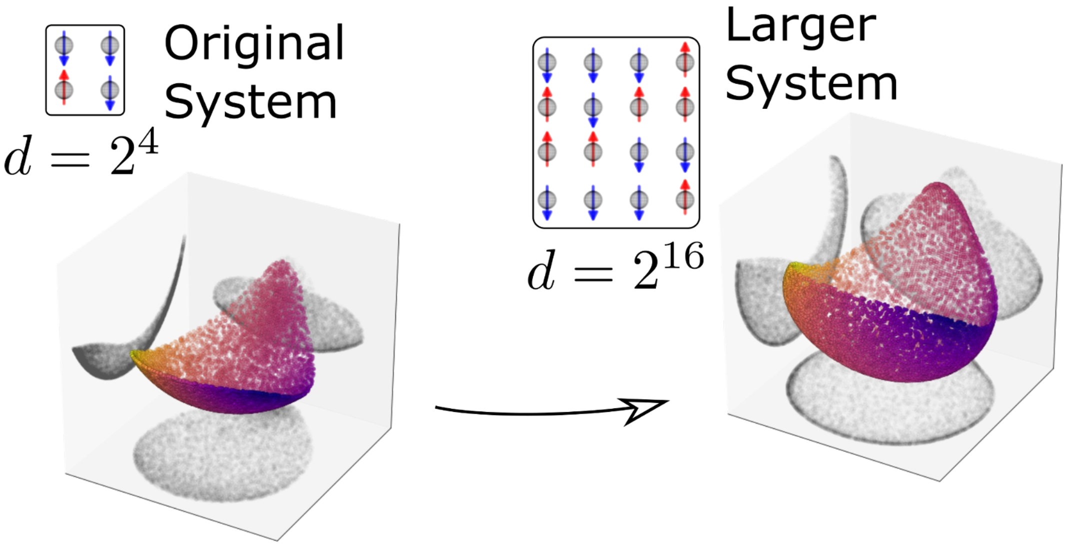
$$1 - \left\langle \sqrt{\mathcal{L}_1(x)}, \sqrt{\mathcal{L}_2(x)} \right\rangle_x$$



$$\left\{ \sqrt{\mathcal{L}(x)} \right\}$$

Curse of Dimensionality

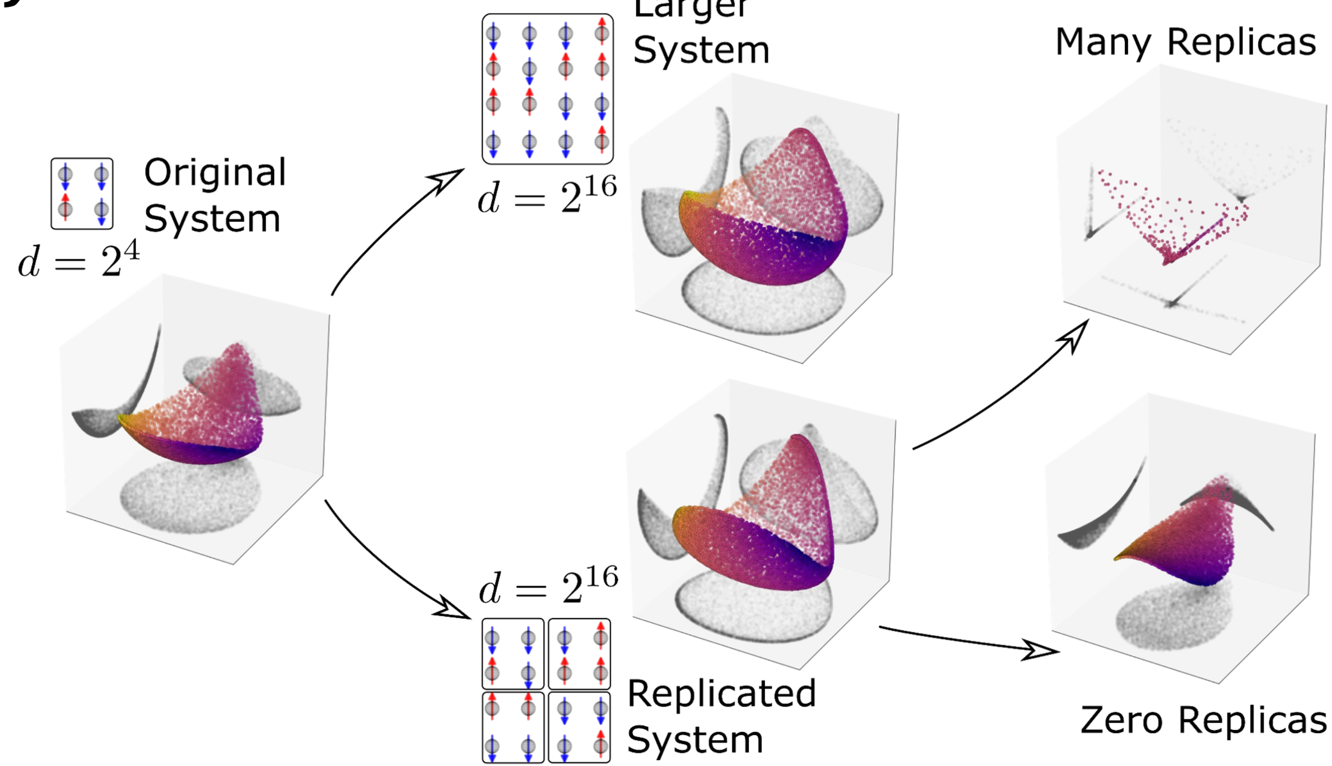
For very high dimensional spaces, distances become saturated and vectors tend to become increasingly orthogonal. Distances will all go to 1.



Replica Theory

Simulate the curse of dimensionality by looking at replicas of the original system.

Consider the limit of zero replicas.



Replica Trick

Replicated Distribution | Hypersphere Distance | Math Trick | Intensive Distance

Replicated Distribution

Looking at replicas of the original distribution (or multiple drawn samples from the same distribution)

$$\mathcal{L}(x) \rightarrow \mathcal{L}(x_1)\mathcal{L}(x_2)\cdots\mathcal{L}(x_n)$$

Hypersphere Distance

Hypersphere distance from Hellinger uses the dot product, which has a nice relationship with replicas.

$$d^2(\mathcal{L}_1, \mathcal{L}_2) = 2\left(1 - \left\langle \sqrt{\mathcal{L}_1(x)}, \sqrt{\mathcal{L}_2(x)} \right\rangle_x\right) \rightarrow 2\left(1 - \left\langle \sqrt{\mathcal{L}_1(x)}, \sqrt{\mathcal{L}_2(x)} \right\rangle_x^n\right)$$

Math Trick

Replica theory relies on the simple math limit.

$$\lim_{n\rightarrow 0} \frac{z^n - 1}{n} = \log(z)$$

Intensive Distance

We obtain a simple distance measure for the limit of zero replicas, related to the Bhattacharyya distance (a known divergence with the FIM as a metric).

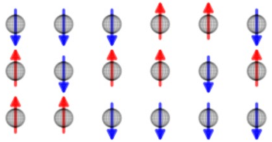
$$\lim_{n\rightarrow 0} \frac{d^2(\mathcal{L}_1, \mathcal{L}_2)}{n} = -2\log\left\langle \sqrt{\mathcal{L}_1(x)}, \sqrt{\mathcal{L}_2(x)} \right\rangle_x$$

Resulting InPCA Algorithm

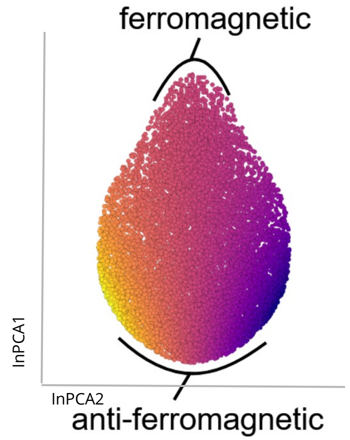
1. **Generate a set of probability distributions** — $\{\mathcal{L}_i(x)\}$
2. **Compute the intensive cross-covariance** — $W_{ij} = \log\left\langle \sqrt{\mathcal{L}_i(x)}, \sqrt{\mathcal{L}_j(x)} \right\rangle + \cdots$
3. **Perform an Eigenvalue decomposition** — $W = U \Sigma U^T$
4. **Find the projection matrix** — $T = U\sqrt{\Sigma}$
5. **Plot the Projections** — $\{(T_{i0}, T_{i1})\}$

Visualizing Probabilistic Manifolds with inPCA

Ising Model

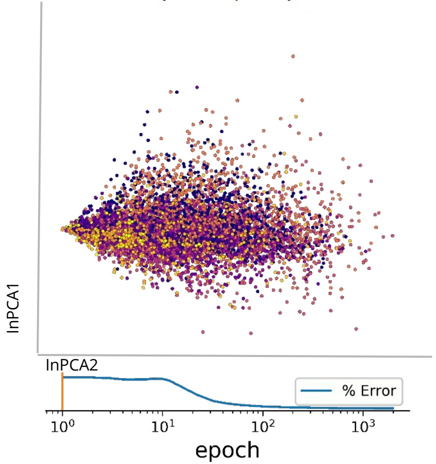


$$\mathcal{L}(S \mid \theta)$$

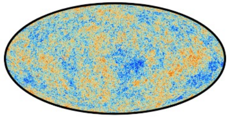


Simple ML Classifier

$$\mathcal{L}(D \mid \theta)$$



Cosmic Microwave Background



$$\mathcal{L}(M \mid \theta)$$