Breakout group: interpretability

Susan Holmes

Christine Wells, Elana Fertig, Aedin Culhane, Lauren Hsu ,Sehyun Oh, Pratheepa Jeganathan, Davide Risso, Alexis Coullomb, Ayshwarya Subramanian, Vera , Genevieve Stein-O’Brien, Kris Sankaran, Mike Love,

Key Points

- Communicating within the field: what are we talking about?
- Using extra (contiguous - meta) data to interpret output.
- Biological Interpretations: bridges to data bases such as KEGG, Gene Ontology, HCA.
- Visualizations: using several sources.
- Validation through complementary data.
- Explaining results to biologists through generative models (Factor Analysis).

Nonlinear dimensionality reduction wiki

Glossary

Here we spoke of multimodal and multiview and used modalities.

Factor Analysis as a generative model not in the French sense (MFA).

Brushing : interactive exploration through a window.

Interpretation of results

What is the biological meaning of what we see?

What does this method do, what can’t it do (deblack-boxing).

How robust are the results across different views or modalities.

Standards for metrics sharing of results/data may help with over-interpreting results.

Relevance to biology/the original question may require additional available resources (like Gene ontology) or defining new metrics.

Issue of discrete cell type

Can we get away from the notion of a discrete cell type that does not exist biologically but exists in the history of experimental designs.

E.g., among immunologists, focus on gating, limitation in discretizing away uncertainty.

This is the data being used to develop new methods

Problem with loss of information in the desire to simplify.

Communication challenge with biologists about tradeoffs between focusing on rare cell types vs. more “continuous” view on cell types.

Important Examples

- Some mistakes made in overinterpretation: counterexamples.
- Figures that help (brushed UMAP plots).
- Figures that are confusing (tSNE).
- Concordance between different domains brings stronger evidence than more of the same (pertains also to ”alignment” between different modalities.
- How much each result is believed (strength of evidence).
- Eric Lock (UMN) non generalization of certain latent factors across different cancers (but good biological reason, which was cancers with male and female vs predominantly one sex alone) - BIDIFAC+ paper. TCGA meta COCA hierarchical analyses: counter-examples.
- Uniform Manifold Approximation (McInnes, Healy, and Melville, 2018) and brushing with covariates (Kris) GitHub: `krisrs1128/birs_mini/master`.

Problems with contiguous data:
Reliance on GSEA on for biological inference and ’validation’ of our data. But a large number genes are poorly annotated if at all. a lot of pathways share the same small set of genes. Some classes of gene/transcript lack an adequate ontology, and pathway enrichment doesn’t consider isoforms which will alter protein partnerships and pathway outcomes.

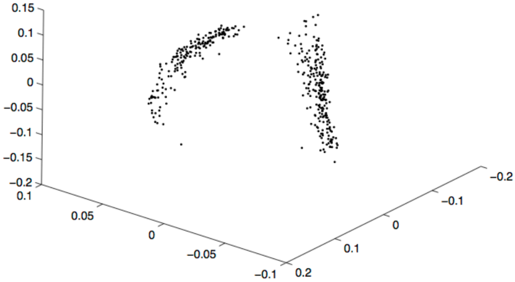
How wrong can we be?

Example of interpretation of voting data

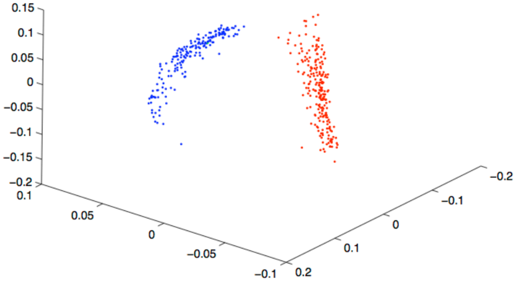
Data from 2005 U.S. House of Representatives roll call votes. We further restricted our analysis to the 401 Representatives that voted on at least 90% of the roll calls (220 Republicans, 180 Democrats and 1 Independent) leading to a 401×669 matrix V of voting data.

```
   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10
1  -1 -1  1 -1  0  1  1  1  1   1
2  -1 -1  1 -1  0  1  1  1  1   1
3   1  1 -1  1 -1  1  1 -1 -1  -1
4   1  1 -1  1 -1  1  1 -1 -1  -1
5   1  1 -1  1 -1  1  1 -1 -1  -1
6  -1 -1  1 -1  0  1  1  1  1   1
7  -1 -1  1 -1 -1  1  1  1  1   1
8  -1 -1  1 -1  0  1  1  1  1   1
9   1  1 -1  1 -1  1  1 -1 -1  -1
10 -1 -1  1 -1  0  1  1  0  0   0
```

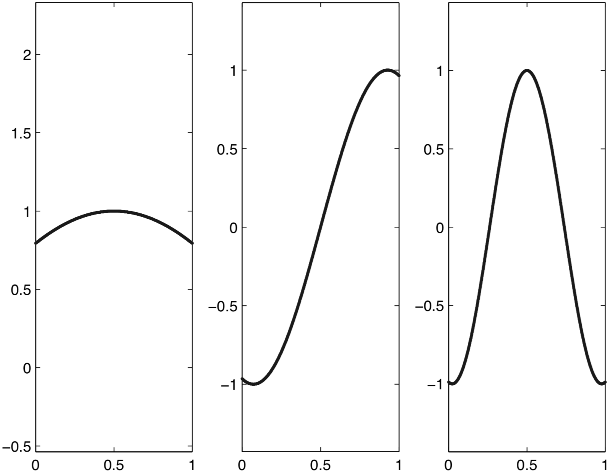
Lower dimensional projection by kernel MDS



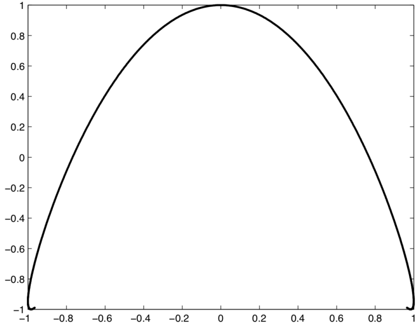
3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes.



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. Color has been added to indicate the party affiliation of each representative.

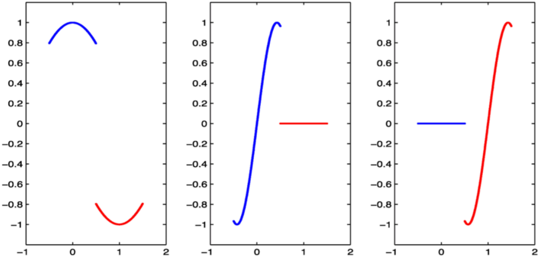


Approximate eigenfunctions f_1, f_2 and f_3.



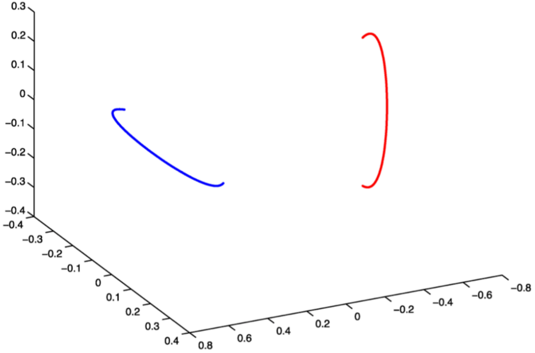
A horseshoe that results from plotting $\Lambda : x_i \mapsto (f_2(x_i), f_3(x_i))$.

It is not in general possible to determine the absolute order knowing only that Λ comes from the eigenfunctions.
You need a crib!



Approximate eigenfunctions g_1, g_2, g_3 and g_4 for the Gram matrix arising from the two population model.

Model eigenfunctions

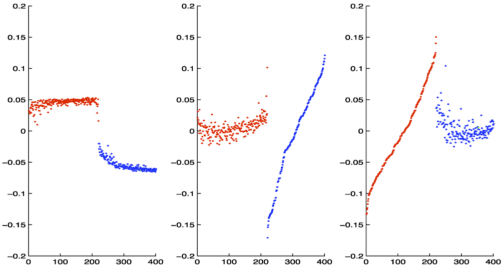


Twin horseshoes that result from plotting
$\Lambda : x_i \mapsto (g_2(x_i), g_3(x_i)g_4(x_i))$.

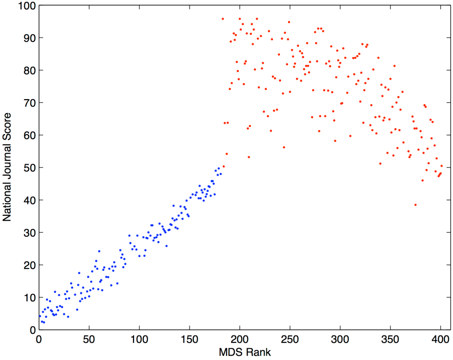
Effect of adding normal $N(0, 1/5)$ noise to the matrix K_{200} before normalizing by the average row sum. The specific form of the noise does not noticeably affect the results.

Numerically obtained eigenfunctions for a noisy K_n (example of parametric Bootstrap).

Data eigenfunctions

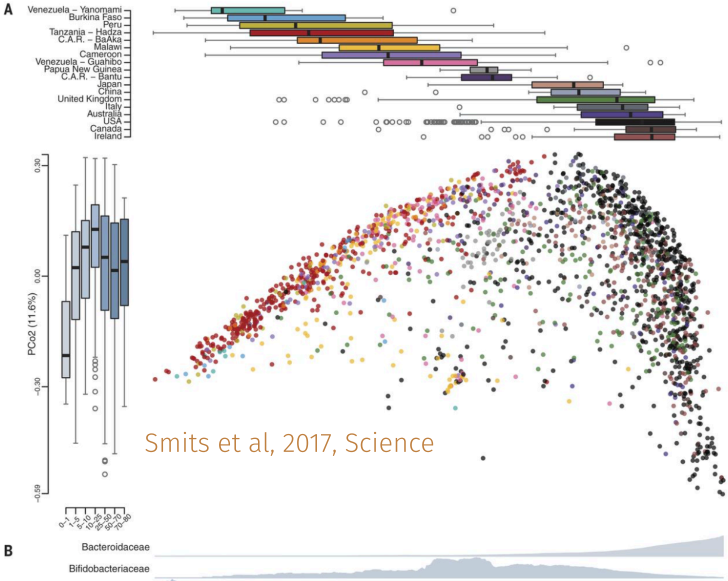


The re-indexed second, third and fourth eigenfunctions outputted from the MDS algorithm applied to the 2005 U.S. House of Representatives roll call votes. Colors indicate political parties.



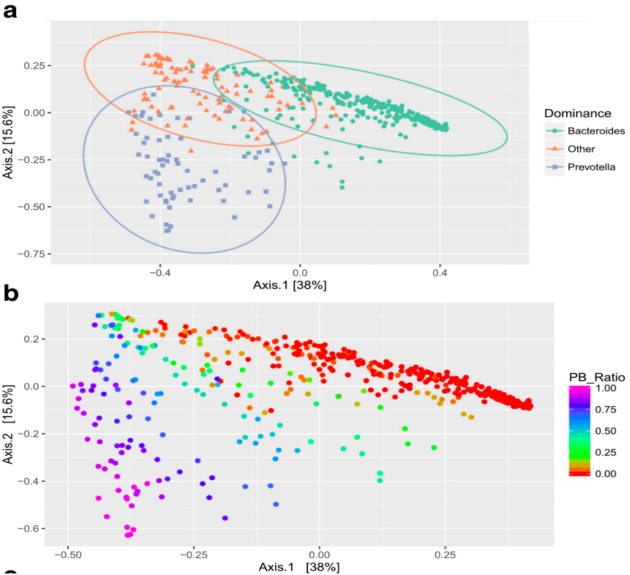
Comparison of the MDS derived rank for Representatives with the National Journal's liberal score

Hazda data

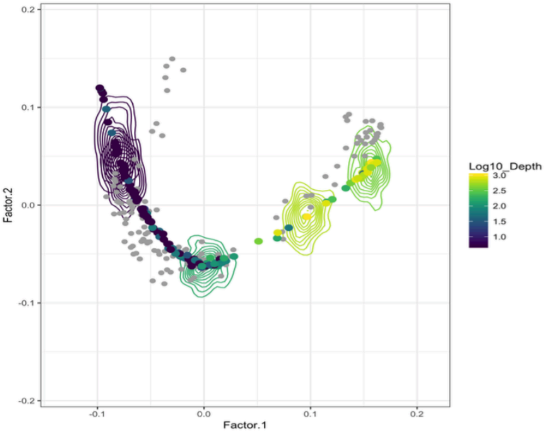


Smits et al, 2017, Science

Gorvitovskaia, Holmes, Huse, 2016, Microbiome



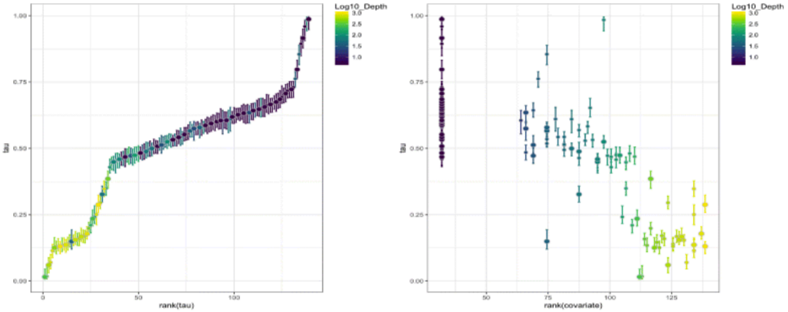
Gradient analysis :Tara Ocean microbiome (with uncertainties)



(b) Datapoint location confidence contours

`buds` package on github: `https://github.com/nlhuong/buds`.
See BMC paper, Nguyen and Holmes (2017).

Tara Ocean rankings data



(a)

(b)

Fig. 4

Latent ordering in TARA Oceans dataset shown with uncertainties. The differences in the slope of plot (**a**) indicate varying data coverage along the underlying gradient. Correlation between the water depth and the latent ordering in microbial composition data is shown in (**b**). Coloring corresponds to log10 of the water depth (in meters) at which the ocean sample was collected

Uncertainty vs missing data - also biological vs technical uncertainty.

Can we use multi-modal data to account for data costs and deeper profiling of rare, expensive samples and then broader profiling of larger populations with cheaper technologies?

How to interpret or assess biological relevance?

Even when we use Bioconductor, how can we be confident on the databases it uses if they were done with “old” method instead of scRNAseq or CITEseq?

How to go to nonlinearity?

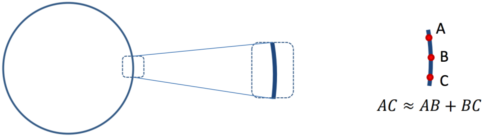
1) Transformations can get us back to linearity (log from multiplicative to additive).

2) Advanced VAE: Interpretation for nonlinearity (Emily Fox and her group).

3) Latent variables become response, back to supervised (covariates can explain

What is a manifold ?

Informally, the manifold is a subset of points in the high dimensional space that locally looks like a low-dimensional space: Example: arc of a circle



Consider a tiny bit of a circumference (2D) can treat as line (1D)

What is a manifold ?

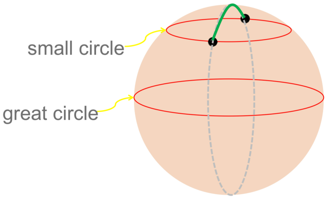
(Link to wikipedia) In mathematics, a manifold is a topological space that locally resembles Euclidean space near each point. More precisely, each point of an ℓ-dimensional manifold has a neighborhood that is homeomorphic to the Euclidean space of dimension ℓ (ℓ-manifold).

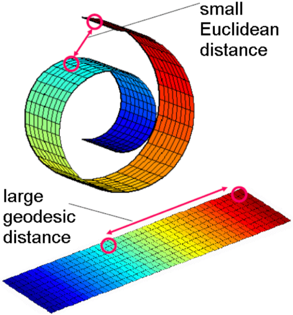
One-dimensional manifolds include lines and circles, but not figure eights (because no neighborhood of their crossing point is homeomorphic to Euclidean 1-space).

Two-dimensional manifolds are also called surfaces. Examples include the plane, the sphere, and the torus, which can all be embedded (formed without self-intersections) in three dimensional real space, but also the Klein bottle and real projective plane, which will always self-intersect when immersed in three-dimensional real space.

Suppose that X is a subset of some ambient Euclidean space $\boldsymbol{R}^m$. Then X is an ℓ -dimensional manifold if each point $x \in X$ possesses

Geodesics





Euclidean distance in the embedding space is often not a good measure of distance between two points on a manifold. Length of geodesic along the manifold is more appropriate.

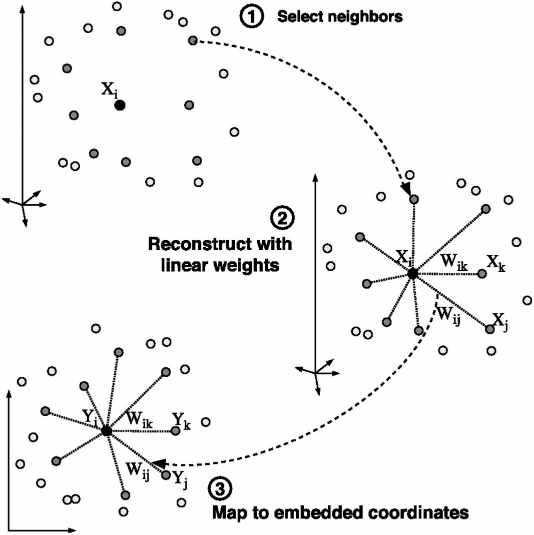
Local Linear Embedding

Nonlinear dimensionality reduction by locally linear embedding. Sam Roweis & Lawrence Saul. Science, v.290 no 5500, Dec.22, 2000. pp.2323– 2326.
A Riemannian manifold is locally linear, model local neighborhoods as linear patches and then embed in a lower dimensional manifold.

LLE also begins by finding a set of the nearest neighbors of each point.

Local Linear Embedding

Compute a set of weights for each point that best describes the point as a linear combination of its neighbors.



Local Linear Embedding

LLE computes the barycentric coordinates of a point x_i based on its neighbors x_j.

The original point is reconstructed by a linear combination, given by the weight matrix W_{ij}, of its neighbors.

The reconstruction error is given by the cost function E(W).

$$E(W) = \sum_i \left| \mathbf{X}_i - \sum_j \mathbf{W}_{ij} \mathbf{X}_j \right|^2$$

reconstructed only from its neighbors, thus enforcing W_{ij} to be zero ii point x_j is not a neighbor of the point x_i and

(b) The sum of every row of the weight matrix is made to be 1

$$\sum_j \mathbf{W}_{ij} = 1$$

Local Linear Embedding

The original data points are collected in a p dimensional space and the goal of the algorithm is to reduce the dimensionality to d such that $p >>> d$. The same weights W in the p dimensional space will be used to reconstruct the same point in the lower d dimensional space. A neighborhood preserving map is created based on this idea. Each point x_i in the p dimensional space is mapped onto an output point Y_i in the d dimensional space by minimizing the cost function

$$C(Y) = \sum_i \left| \mathbf{Y}_i - \sum_j \mathbf{W}_{ij} \mathbf{Y}_j \right|^2$$

In this cost function, the weights $\mathrm{W_{ij}}$ are kept fixed and the minimization is done on the points Y_{i} to optimize the coordinates.

Local Linear Embedding

This minimization problem can be solved by solving a sparse $N \times N$ eigenvalue problem (N being the number of data points), whose bottom nonzero eigenvectors provide an orthogonal set of coordinates.

This is done by taking the lower eigenvectors of $(I - W)(I - W)^t$.

The data points are reconstructed from K nearest neighbors, as measured by Euclidean distance.

For such an implementation the algorithm has only one free parameter K, which can be chosen by cross validation.

LLE tends to handle non-uniform sample densities poorly because there is no fixed unit to prevent the weights from drifting as various regions differ in sample densities. LLE has no internal model.

Nonlinear embedding: general method

- Find a representation of points on a manifold.
- Use NN graphs.
- Create a low dimensional embedding.

Problems occur if the densities of points are unequally sampled on the manifold, if the data come from a very high dimensional space or are not from a connected manifold (support problems). `tSNE, UMAP`.

t-SNE is a graph based method: executive summary

K-Nearest Neighbours weighted by a kernel with bandwidth adapted to the K neighbours

Normalize outgoing edge weights to sum to one.

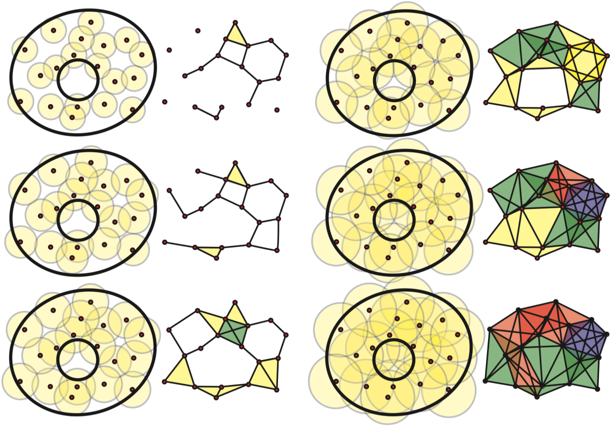
Symmetrize by averaging edge weights between each pair of vertices

Renormalize so the total edge weight is one

Use a force directed graph layout (almost).

Mathematicians also have baseline measure problems

Ghrist, R. Barcodes: persistent toology of data, AMS, 2008

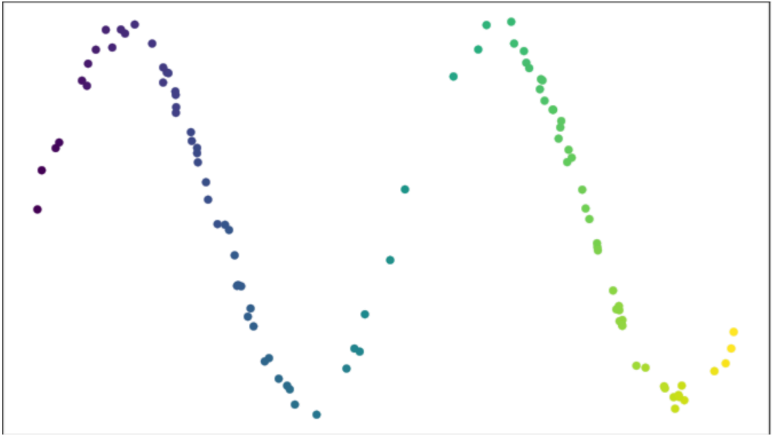


UMAP: topology to extend the graph theoretical view

Mclnnes, L., Healy, J., and Melville, J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Arxiv-preprint

How to use tSNE: examples
Using UMAP on the tSNE examples

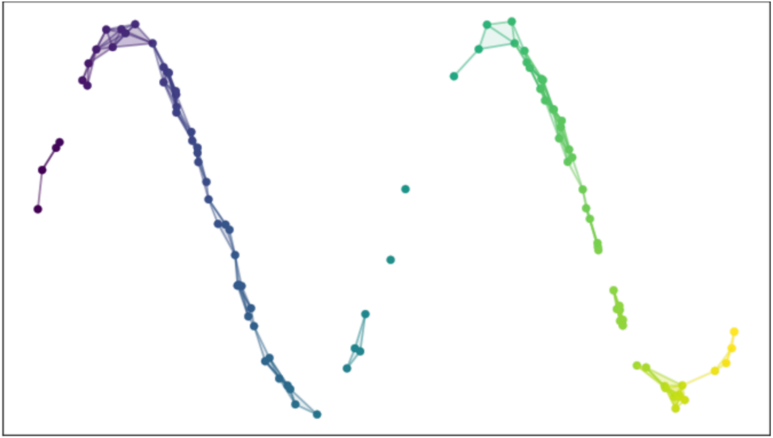
A data example



Nerve Theorem

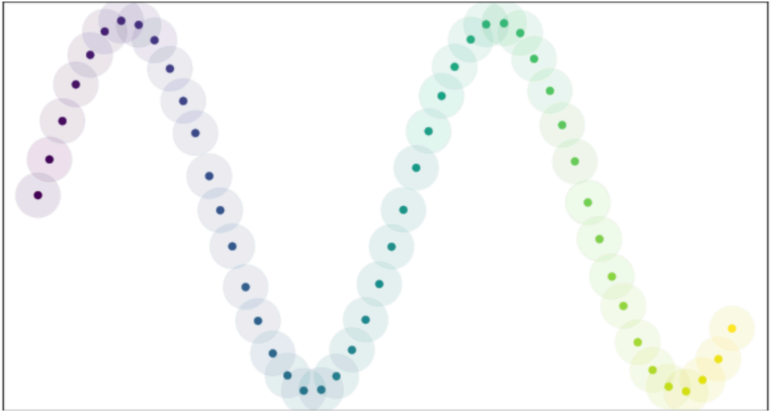
Let $\mathcal{U} = \{U_i\}_{i\in I}$ be a cover of a topological space X. If, for all $\sigma \subset I$ $\bigcap_{i\in\sigma} U_i$ is either contractible or empty, then $\mathcal{N}(\mathcal{U})$ is homotopically equivalent to X

Real data are more problematic



No Problem for uniform

If the data is uniformly distributed on the manifold then the cover will be “good”



More maths needed than just the simplex and cover.

Functor: A function between domains of discourse (categories).

Adjunction: A near equivalence between domains of discourse.

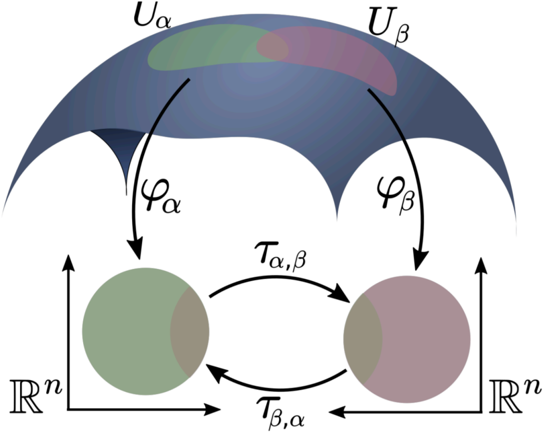
Limit: A solution to a system of constraints.

Colimit: Gluing together a system of objects.

Best data are uniform

Data is not uniformly distributed on the manifold.

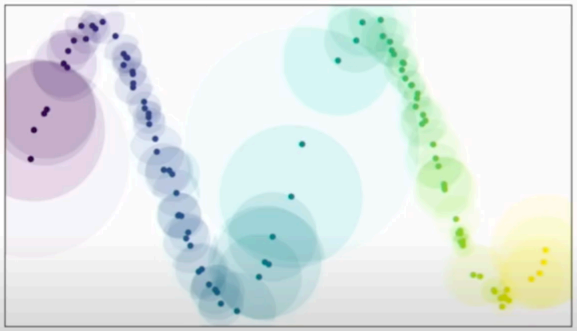
So, change the metric:



Riemannian metric on the manifold.

Assumption

The manifold is locally connected.



In practice: cross entropy

$$\sum_{a \in A} \mu(a) \log\left(\frac{\mu(a)}{\nu(a)}\right) + (1 - \mu(a)) \log\left(\frac{1 - \mu(a)}{1 - \nu(a)}\right)$$

In practice: cross entropy

$$\sum_{a \in A} \overbrace{\mu(a) \log \left(\frac{\mu(a)}{\nu(a)} \right)}^{\text{Get the clumps right}} + \underbrace{(1 - \mu(a)) \log \left(\frac{1 - \mu(a)}{1 - \nu(a)} \right)}_{\text{Get the gaps right}}$$

$$CE(X, Y) = \sum_i \sum_j \left[p_{ij}(X) \log \left(\frac{p_{ij}(X)}{q_{ij}(Y)} \right) + (1 - p_{ij}(X)) \log \left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)} \right) \right]$$

Probability used instead of Student Kernel

UMAP uses exponential probability distribution in high dimensions but not necessarily Euclidean distances like tSNE but rather any distance can be plugged in. In addition, the probabilities are not normalized:

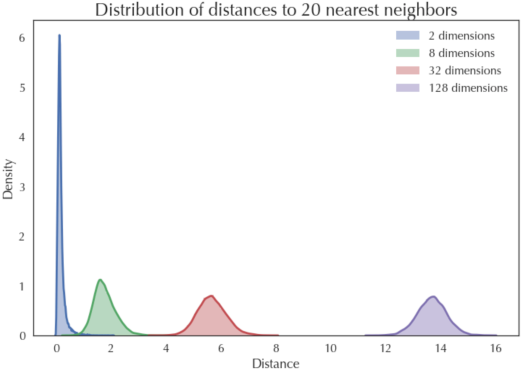
$$p_{i|j} = exp\{-\frac{d\left(x_i, x_j\right) - \rho_i}{\sigma_i}\}$$

Here ρ is an important parameter that represents the distance from each i- th data point to its first nearest neighbor.

This ensures the local connectivity of the manifold. In other words, this gives a locally adaptive exponential kernel for each data point, so the distance metric varies from point to point.

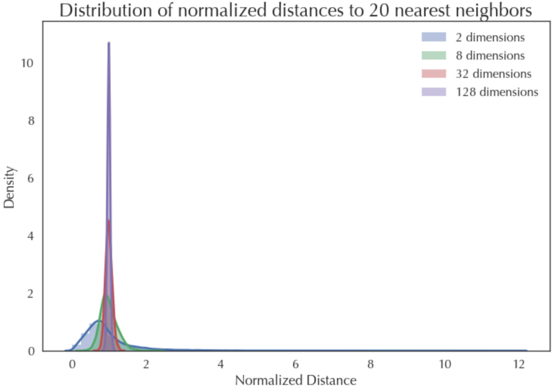
Avoiding curse of dimensionality problems

Un-normalized distances:



Avoiding curse of dimensionality problems

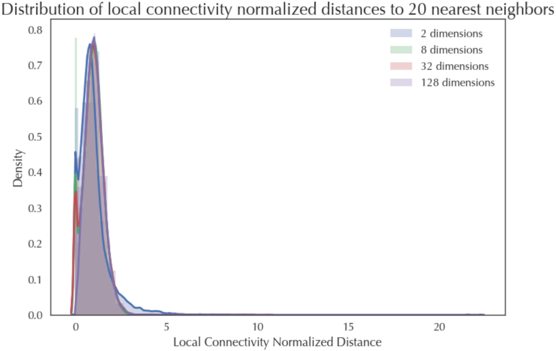
Un-normalized distances:



Normalization loses information.

Avoiding curse of dimensionality problems

Un-normalized distances:



UMAP does not apply normalization to either high- or low-dimensional probabilities.

Rescaled distances

See details here.

UMAP uses the family of curves $1/\left(1+a * y^{\wedge}(2 b)\right)$ for modelling distance probabilities in low dimensions, not exactly Student t-distribution but very-very similar, please note that again no normalization is applied:

$$q_{ij}=\left(1+a\left(y_i-y_j\right)^{2b}\right)^{-1}$$

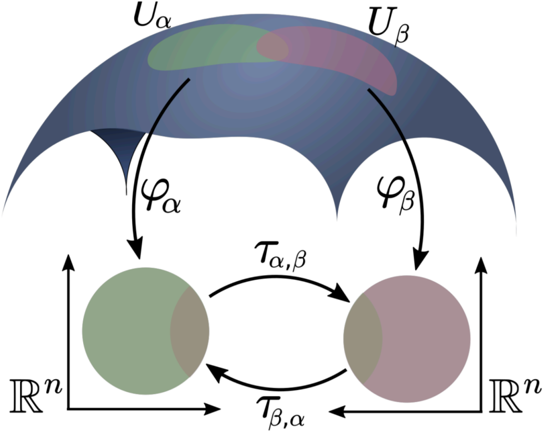
where $a \approx 1.93$ and $b \approx 0.79$ for default UMAP hyperparameters (in fact, for min_dist $=0.001$).

In practice, UMAP finds a and b from non-linear least square fittings to the piecewise function with the `min_dist` hyperparameter

$$\left(1+a\left(y_i-y_j\right)^{2b}\right)^{-1} \approx \begin{cases} 1 & \text{if } y_i-y_j \leq \min\_\text{dist} \\ e^{-\left(y_i-y_j\right)-\min\_\text{dist}} & \text{if } y_i-y_j > \min\_\text{dist} \end{cases}$$

Problem

The local metrics are all incompatible.



Combining different patch distances