A Graph Coupling View of Dimension Reduction

Hugues van Assel[*], Thibault Espinasse, Julien Chiquet[†], Franck Picard[*]

[*]Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

franck.picard@ens-lyon.fr

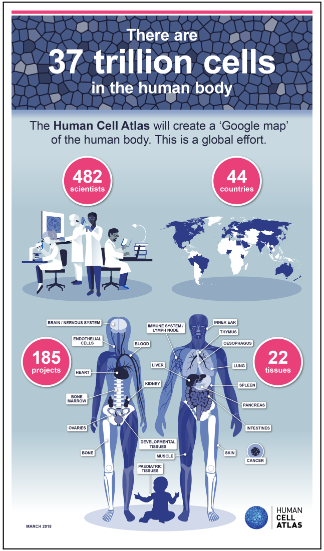
Groupe StatMath SFdS, Janvier 2023

Outline

1. **Single-Cell Genomics and Biology**
2. Presentation of Neighbor Embedding Methods
3. Empirical properties of tSNE
4. First steps in the definition of the graph coupling strategy
5. Prior and Posterior distributions for graph coupling
6. Open questions and research challenges

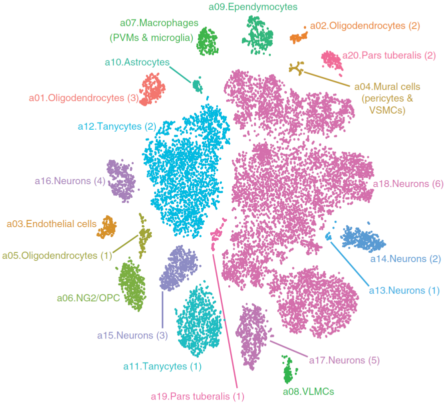
The Single-Cell Revolution

- Cells are the basic unit of living organisms
- Recent technological breakthroughs allow the molecular characterization of cells
- Describe cell population with high dimensional molecular features



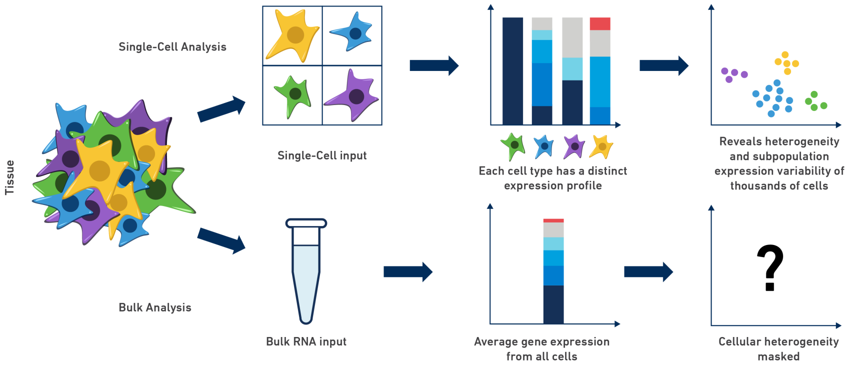
Cell biology goes genome-wide

- Investigate shapes, locations, interactions, functions of cell types
- Classify cells into distinct cell types
- Account for the between-cell variability and heterogeneities



[1]

Single-Cell from a statistician's perspective


Tissue
Single-Cell Analysis
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

From 10X Genomics

An unprecedented challenge

- Genomics was precursor for data representation and visualization

Publication	cells	tissue	Seq. protocol	clusters
Cadwell et al. (2016)	46	visual cortex	Smart-seq2	2
Tasic et al. (2016)	1,679	visual cortex	SMARTer	49
Macosko et al. (2015)	44,808	retina	Drop-seq	39
10x Genomics	1,306,127	brain cells	10x Gen.Chrom.	39

- Dimension reduction is mandatory for any analysis (clustering, visualization, Regulatory networks inference, etc)

High-dimensional count data

$$x_{ij} = \text{expression of gene } j \text{ in cell } i$$

$$\mathbf{X}_{n\times p} = \begin{array}{c} \left[\begin{array}{c|c|c|c|c|c|c} & & & & & & \\ \hline & & & x_{ij} & & & \\ \hline & & & & & & \end{array}\right] \begin{array}{c} 1 \\ \vdots \\ n \end{array} \left.\vphantom{\begin{array}{c} 1 \\ \vdots \\ n \end{array}}\right\} \text{cells} \\ \underbrace{1 \quad \dots \quad \dots \quad \dots \quad \dots \quad p}_{\text{genes}} \end{array}$$

- **High dimension:** n grows but $\ll p$ & **Big Data:** n and p grow
- **Count data** with ove-rdispersion and excess of zeros

Outline

1. Single-Cell Genomics and Biology

2. **Presentation of Neighbor Embedding Methods**

3. Empirical properties of tSNE

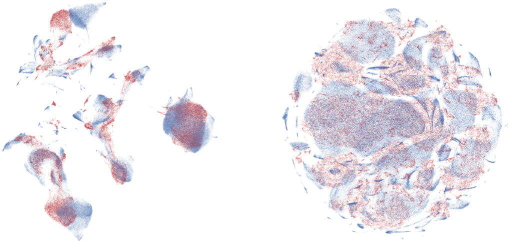
4. First steps in the definition of the graph coupling strategy

5. Prior and Posterior distributions for graph coupling

6. Open questions and research challenges

Beyond Linear methods

- Linear methods like PCA are robust but badly shaped for complex geometries
- High-dim. datas are characterized by multiscale properties (local / global structures)
- Non-Linear projection methods aim at preserving local characteristics of distances
- Many proposed methods such as LargeVis, tSNE, UMAP



(a) UMAP

(b) t-SNE

from [3]

Stochastic Neighbor Embedding (SNE) [4]

- $(X_1, \ldots, X_n)$ are the points in the high-dimensional space $\mathbb{R}^p$,
- Consider a similarity between points:

$$p_{i|j} = \frac{\exp(-\|X_i - X_j\|^2/2\sigma_i^2)}{\sum_{\ell \neq i} \exp(-\|X_\ell - X_j\|^2/2\sigma_\ell^2)}$$

- Further symmetrized

$$p_{ij} = (p_{i|j} + p_{j|i})/2N$$

- Hyper-parameter σ_i locally smooths the data, to be tuned
- Linked to the regularity of the target manifold

tSNE and Student / Cauchy kernels

- Consider $(Z_1, \dots, Z_n)$ are points in the low-dimensional space $\mathbb{R}^2$
- Consider a similarity between points in the new representation:

$$q_{i|j} = \frac{\exp(-\|Z_i - Z_j\|^2)}{\sum_{\ell \neq i} \exp(-\|Z_\ell - Z_j\|^2)}$$

- Robustify this kernel by using Student(1) kernels (ie Cauchy)

$$q_{i|j} = \frac{(1 + \|Z_i - Z_j\|^2)^{-1}}{\sum_{\ell \neq i} (1 + \|Z_i - Z_\ell\|^2)^{-1}}$$

Optimizing tSNE by Gradient descent

- Minimize the KL between p and q to find $Z \in \mathbb{R}^2$ such that:

$$C(Z) = \sum_{ij} KL(p_{ij}, q_{ij})$$

$$\left[\frac{\partial C(Z)}{\partial Z}\right]_i = \sum_j (p_{ij} - q_{ij})(Z_i - Z_j)$$

- Gradient update (adaptive learning rate η)

$$Z^{(t)} = Z^{(t-1)} + \eta \frac{\partial C(Z)}{\partial Z} + \alpha(t)(Z^{(t-1)} - Z^{(t-2)})$$

- $\alpha(t)$ momentum to speed up and improve convergence
- Initialization $Z_i^{(0)} \sim \mathcal{N}(0, \delta I)$, δ small.

Uniform Manifold Approximation and Projection [3]

$$\forall (i,j) \in [n]^2, \quad p_{j|i} = \exp\left(-\frac{\|X_i - X_j\|_2^2 - \rho_i}{\sigma_i}\right)$$

with $\rho_i = \min_{j \neq i} \|X_i - X_j\|^2$. Let us define

$$p_{ij} = p_{j|i} + p_{i|j} - p_{j|i} p_{i|j}$$

and:

$$\forall (i,j) \in [n]^2, \quad q_{ij} = \left(1 + a\|X_i - X_j\|_2^{2b}\right)^{-1}$$

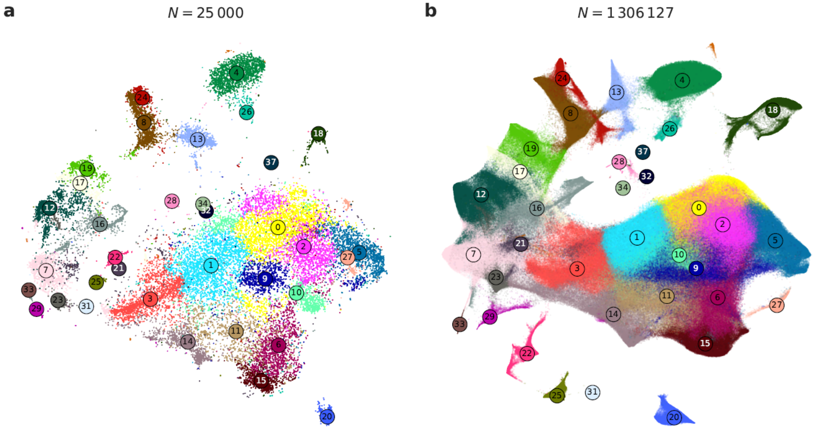
UMAP solves the following problem:

$$\min_{Z \in \mathbb{R}^{n \times d}} \quad -\sum_{i<j} p_{ij} \log q_{ij} + (1 - p_{ij}) \log(1 - q_{ij})$$

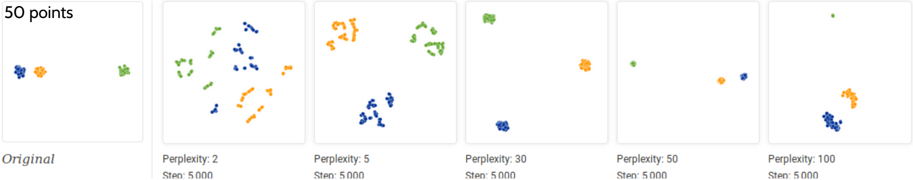
Outline

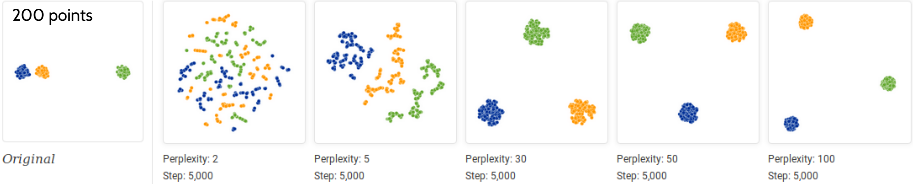
1. Single-Cell Genomics and Biology
2. Presentation of Neighbor Embedding Methods
3. **Empirical properties of tSNE**
4. First steps in the definition of the graph coupling strategy
5. Prior and Posterior distributions for graph coupling
6. Open questions and research challenges

tSNE on single cell Gene Expression data [2]

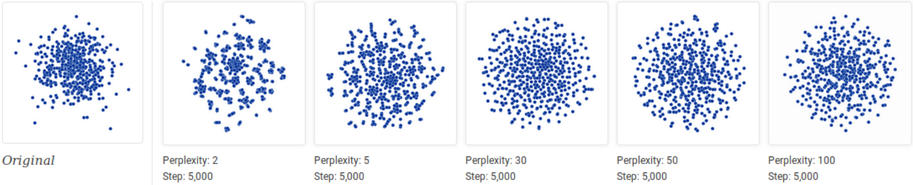


tSNE does not account for between-cluster distance

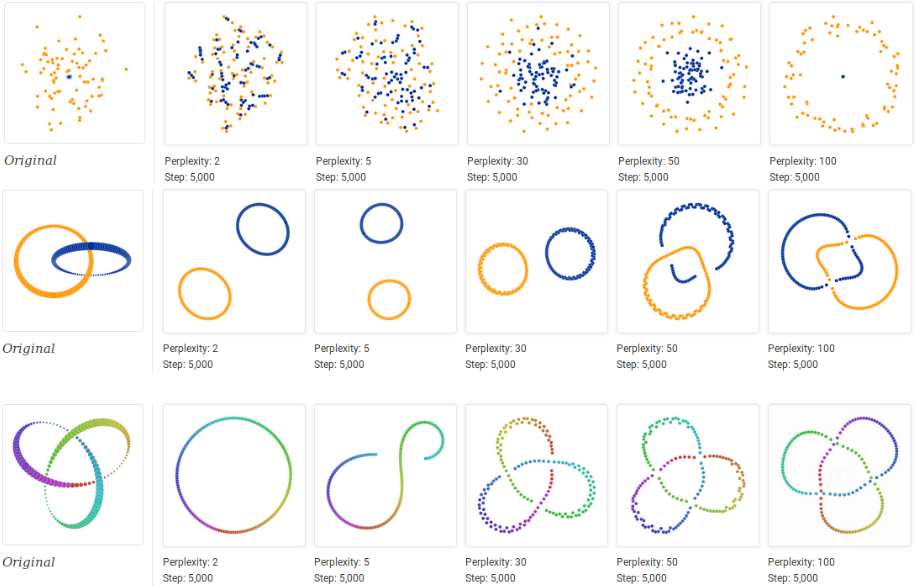




What about random noise ?



Catching Complex Geometries



Properties of t-SNE

- Good at preserving local distances (intra-cluster variance)
- Not so good for global representation (inter-cluster variance)
- Good at creating clusters of points that are close, but bad at positioning clusters wrt each other
- Does not handle well high dimensional data (preliminary PCA and feature selection)
- Sensistive to the calibration of the hyperparameter (smoothing)
- Reproducibility of results due to stochastic optimization

$\rightarrow$ What are the statistical / probabilistic foundations of Stochastic Neighbor Embedding ?

Outline

1. Single-Cell Genomics and Biology
2. Presentation of Neighbor Embedding Methods
3. Empirical properties of tSNE
4. **First steps in the definition of the graph coupling strategy**
5. Prior and Posterior distributions for graph coupling
6. Open questions and research challenges

Hidden Graph to structure observations

- Consider W the adjacency matrix of a hidden random graph
- The graph Laplacian operator is the map L such that for $(i,j) \in [n]^2$:

$$L(W)_{ij} = \begin{cases} -W_{ij} & \text{if } i \neq j \\ \sum_{k \in [n]} W_{ik} & \text{otherwise .} \end{cases}$$

- $L = L(W)$ has the following property:

$$\forall X \in \mathbb{R}^{n \times p}, \quad \sum_{i,j} W_{ij} \|X_i - X_j\|^2 = \mathrm{tr}(X^T L X).$$

- In a first step, consider a graph with one connected component

Conditional distribution of X on a graph W_X

- Consider a Matrix Normal model with row and column dependencies

$$X \mid W_X \sim \mathcal{MN}\left(0, L_X^{-1}, \Sigma^{-1}\right),$$

- L_X^{-1} between-cell correlation, Σ^{-1} between-genes correlation.
- The conditional density relates to the Gaussian kernel

$$k(X_i - X_j) = \exp\left(-\frac{1}{2}\|X_i - X_j\|_\Sigma^2\right),$$

- Which can be generalized to translation invariant kernels:

$$\mathbb{P}(X \mid W_X) \propto \prod_{(i,j)\in[n]^2} k(X_i - X_j)^{W_{X,ij}}.$$

Conditional distribution of Z on a graph W_Z

- Consider that the low-dimensional representation is also structured according to a graph

$$Z \mid W_Z \sim \mathcal{MN}\left(0, L_Z^{-1}, I_q\right),$$

- Consider the Gaussian kernel for Z

$$k(Z_i - Z_j) = \exp\left(-\frac{1}{2}\|Z_i - Z_j\|_{I_q}^2\right),$$

- Conditional distribution of $Z \mid W_Z$:

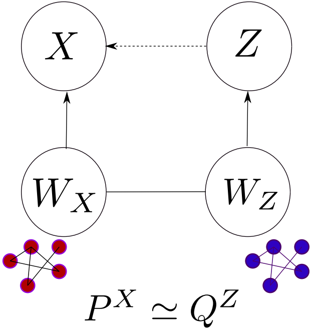
$$\mathbb{P}(Z \mid W_Z) \quad \propto \quad \prod_{(i,j)\in[n]^2} k(Z_i - Z_j)^{W_{Z,ij}}$$

Embedding with Graph Coupling

- Consider two hidden graphs W_X and W_Z
- Couple W_X with W_Z in a probabilistic way
- Match their posterior distributions

$$\begin{aligned} P^X &= \mathbb{P}(W_X \mid X) \\ Q^Z &= \mathbb{P}(W_Z \mid X; Z) \end{aligned}$$

- Z becomes a parameter to be estimated



Probabilistic Coupling

Graph Coupling with Z as a parameter

- Consider the cross entropy between posteriors

$$\mathcal{H}\big(P^X, Q^Z\big) = -\mathbb{E}_{W_X \sim P^X}\Big(\log \mathbb{P}(W_Z = W_X \mid X; Z)\Big)$$

- Find the best low-dimensional representation such that the two graphs match

$$Z(X) \quad = \quad \arg\min_{Z} \Big\{\mathcal{H}\big(P^X, Q^Z\big)\Big\}$$

- Connection with the KL between posteriors

$$\mathrm{KL}\big(P^X, Q^Z\big) = \mathcal{H}\big(P^X, Q^Z\big) - \mathcal{H}\big(P^X, P^X\big)$$

First Outline

Done...

- Consider two hidden random graphs W_X, W_Z
- Define a conditional model $X \mid W_X, Z \mid W_Z$
- Consider pairwise similarity distributions (Pairwise Markov Random Field)
- Find Z by matching the posteriors using a cross entropy criterion

...to be done :

- Define/Construct the priors for W_X, W_Z
- Deduce/Induce the posteriors for W_X, W_Z
- Carefully inspect the case with more than one connected component

Outline

1. Single-Cell Genomics and Biology
2. Presentation of Neighbor Embedding Methods
3. Empirical properties of tSNE
4. First steps in the definition of the graph coupling strategy
5. **Prior and Posterior distributions for graph coupling**
6. Open questions and research challenges

Construction of conjugate priors for hidden graphs

- Consider a prior distribution for the hidden graph in the general form ($\alpha = 0$ later on)

$$\mathbb{P}_{\mathcal{P}}(W;\pi) \propto \mathcal{C}_k(W)^{\alpha}\,\Omega_{\mathcal{P}}(W) \prod_{(i,j)\in[n]^2} \pi_{ij}^{W_{ij}}$$

- $\mathcal{P}$ stands for a family of priors s.t:

$\mathcal{P}$		$\Omega_{\mathcal{P}}(W)$	Prior for W
$\mathcal{B}$	Bernoulli	$\prod_{ij} 1_{W_{ij}\leq 1}$	$\mathcal{B}\left(\frac{\pi_{ij}}{1+\pi_{ij}}\right)$
$\mathcal{D}$	Unitary Fixed degree	$\prod_{i} 1_{W_{i+}=1}$	$\mathcal{M}\left(1, \frac{\pi_i}{\pi_{i+}}\right)$
$\mathcal{E}$	Fixed Number of edges	$\prod_{ij}(W_{ij}!)^{-1}$	$\mathcal{M}\left(n, \frac{\pi}{\pi_{++}}\right)$

Deducing the limit posterior for hidden graphs

- We show that the posterior distribution $\mathbb{P}_{\mathcal{P}}(W \mid X; \pi, k)$ converge to (details later)

$\mathcal{P}$		Approximate Posterior for W
$\mathcal{B}$	Bernoulli	$\mathcal{B}\left(\frac{\pi_{ij} k_{ij}}{1+\pi_{ij} k_{ij}}\right)$
$\mathcal{D}$	Unitary Fixed degree	$\mathcal{M}\left(1, \frac{[\pi k]_i}{[\pi k]_{i+}}\right)$
$\mathcal{E}$	Fixed Number of edges	$\mathcal{M}\left(n, \frac{\pi k}{[\pi k]_{++}}\right)$

- $\pi_{ij} k_{ij} = \pi_{ij} k(X_i - X_j)$ is the posterior strength of edges (normalized or not)

Mixing Prior distributions for coupling

- Priors for W_X and W_Z induce the approximate posteriors

$$\mathbb{P}_{\mathcal{P}_X}(W_X \mid X; \pi_X, k_X) = P^{\mathcal{P}_X}$$

$$\mathbb{P}_{\mathcal{P}_Z}(W_Z \mid X; \pi_Z, k_Z) = Q^{\mathcal{P}_Z}$$

- Match the approximate posteriors

$$\mathcal{H}\left(P^{\mathcal{P}_X}, Q^{\mathcal{P}_Z}\right) = -\mathbb{E}_{W_X \sim P^{\mathcal{P}_X}}\left\{\log \mathbb{P}_{\mathcal{P}_Z}(W_Z = W_X; \pi_Z, k_Z)\right\}$$

Model based Stochastic Neighbor Embedding

- Choosing $\mathcal{P}_X = \mathcal{P}_Z = \mathcal{D}$:

$$\mathcal{H}_{D,D} = -\sum_{i \neq j} P_{ij}^D \log Q_{ij}^D \ .$$

$$P_{ij}^D = \frac{\pi_{ij} k(X_i - X_j)}{\sum_{\ell=1}^n \pi_{i\ell} k(X_i - X_\ell)}, \quad Q_{ij}^D = \frac{\pi_{ij} k(Z_i - Z_j)}{\sum_{\ell=1}^n \pi_{i\ell} k(Z_i - Z_\ell)}.$$

- We defined the generative model for SNE !
- Can be generalized to symmetric graphs

Model based UMAP [3]

- Choose $\mathcal{P}_X = \mathcal{P}_Z = \mathcal{B}$ and define the symmetrized graph

$$\widetilde{W}_X = 1_{W_X + W_X^T \geq 1}$$

- By independence of the symmetrized edges,

$$\widetilde{W}_{X,ij} \sim \mathcal{B}\left(\widetilde{P}^B_{ij}\right) \quad \text{with} \quad \widetilde{P}^B_{ij} = P^B_{ij} + P^B_{ji} - P^B_{ij} P^B_{ji}$$

- Coupling $\widetilde{W}_X$ and W_Z gives:

$$\mathcal{H}_{\widetilde{B},B} = -2 \sum_{i<j} \widetilde{P}^B_{ij} \log Q^B_{ij} + \left(1 - \widetilde{P}^B_{ij}\right) \log\left(1 - Q^B_{ij}\right)$$

General Approach for Graph Coupling

Algorithm	Input Similarity	Latent Similarity	Loss Function
SNE	$P^D_{ij} = \frac{k_x(X_i - X_j)}{\sum_\ell k_x(X_i - X_\ell)}$	$Q^D_{ij} = \frac{k_z(Z_i - Z_j)}{\sum_\ell k_z(Z_i - Z_\ell)}$	$-\sum_{i \neq j} P^D_{ij} \log Q^D_{ij}$
Sym-SNE	$\overline{P}^D_{ij} = P^D_{ij} + P^D_{ji}$	$Q^E_{ij} = \frac{k_z(Z_i - Z_j)}{\sum_{\ell,t} k_z(Z_\ell - Z_t)}$	$-\sum_{i<j} \overline{P}^D_{ij} \log Q^E_{ij}$
LargeVis	$\overline{P}^D_{ij} = P^D_{ij} + P^D_{ji}$	$Q^B_{ij} = \frac{k_z(Z_i - Z_j)}{1 + k_z(Z_i - Z_j)}$	$-\sum_{i<j} \overline{P}^D_{ij} \log Q^B_{ij} + \left(2 - \overline{P}^D_{ij}\right) \log(1 - Q^B_{ij})$
UMAP	$\widetilde{P}^B_{ij} = P^B_{ij} + P^B_{ji} - P^B_{ij} P^B_{ji}$	$Q^B_{ij} = \frac{k_z(Z_i - Z_j)}{1 + k_z(Z_i - Z_j)}$	$-\sum_{i<j} \widetilde{P}^B_{ij} \log Q^B_{ij} + \left(1 - \widetilde{P}^B_{ij}\right) \log(1 - Q^B_{ij})$

Second Outline

Done...

- Consider two hidden random graphs W_X, W_Z
- Define a conditional model $X \mid W_X, Z \mid W_Z$
- Consider pairwise similarity distributions (Pairwise Markov Random Field)
- Find Z by matching the posteriors using a cross entropy criterion
- Define/Construct the priors for W_X, W_Z
- Deduce/Induce the posteriors for W_X, W_Z

...to be done :

- Carefully inspect the case with more than one connected component

The model is not fully integrable

- Suppose the graph has R connected components of size $n_r = \mathrm{Card}(C_r)$.
- By the spectral theorem $L = U \Lambda U^T$ where $U = (U_1, ..., U_n)$ is orthogonal

$$\forall r \in \{1, \dots, R\}, \quad \lambda_r = 0 \quad \text{and} \quad U_r = \left(n_r^{-1/2} 1_{i \in C_r} \right)_{i \in [n]}$$

- $(U_1, ..., U_R)$ is an orthogonal basis of $\ker(L)$
- The projection of X on $\ker(L)$ is the empirical mean by connected components

$$X_{M,i} = \frac{1}{n_r} \sum_{r \in [R]} \left\{ 1_{i \in C_r} \left(\sum_{\ell \in C_r} X_\ell \right) \right\}$$

- $\mathbb{P}(X \mid W_X)$ is not fully integrable on $\mathbb{R}^{n \times p}$ but only on $\ker(L)^\perp$

$$X - X_M : \text{relative position of points within CC}$$

Diffuse Conditional and Integrability

- To overcome the integrability issue, we introduce a distribution on CC means

$$\mathbb{P}(X \mid W_X) = \mathbb{P}(X_M \mid W_X) \times \mathbb{P}(X - X_M \mid W_X)$$

- We choose a distribution on CC means such that:

$$X_M | \Theta \sim \mathcal{MN}\left(0, \left[\varepsilon U_{1:R} \Theta U_{1:R}^T\right]^{-1}, \Sigma\right)$$

- When $\varepsilon \to 0$, the position of CCs is not informative anymore

Completed model and posterior computations

- Posterior computations are complex wrt to CC membership
- $\mathbb{P}_{\mathcal{P}}(W_X \mid X; \pi, k)$ can not be computed easily
- Taking $\varepsilon \to 0$ compensates for the uninformative diffuse conditional on X_M
- This full model at the limit allows to retrieve an approximate tractable posterior

Outline

1. Single-Cell Genomics and Biology
2. Presentation of Neighbor Embedding Methods
3. Empirical properties of tSNE
4. First steps in the definition of the graph coupling strategy
5. Prior and Posterior distributions for graph coupling
6. **Open questions and research challenges**

Kernel calibration and Perplexity

- tSNE strongly depends on the calibration of the kernel

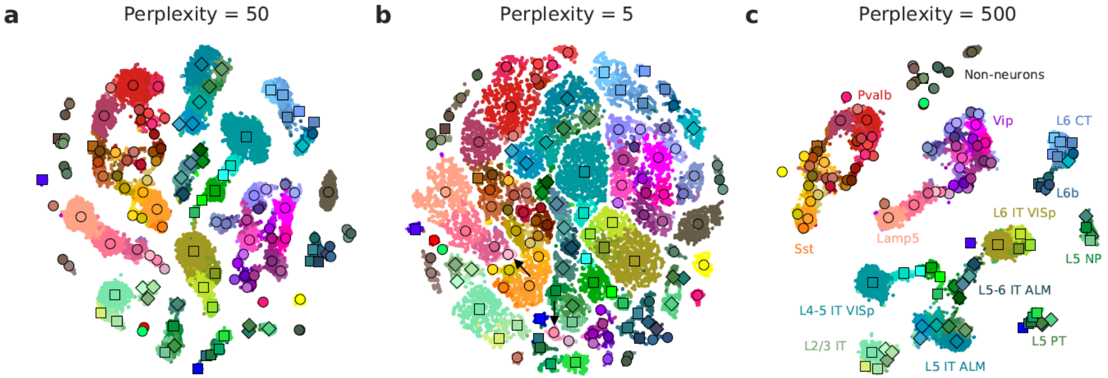
$$k(X_i - X_j; \sigma_i) = \exp\left(-\frac{1}{2\sigma_i}\|X_i - X_j\|_\Sigma^2\right),$$

- σ_i should adjust to local densities (neighborhood of point i)
- In practice, the method is tuned by fixing a given amount of entropy

$$H(p_i) = -\sum_{j=1}^{n} p_{ij} \log_2 p_{ij}$$

- Find σ_i such that $2^{H(p_i)} = \text{perp}$ (user defined)
- Interpreted as the smoothed effective number of neighbors.

Visual inspection of the influence of σ[2]



Back to the coupling strategy

- Maximizing the probability of coupling by minimizing the KL

$$\mathrm{KL}\big(P^X, Q^Z\big) = \mathcal{H}\big(P^X, Q^Z\big) - \mathcal{H}\big(P^X, P^X\big)$$

- $\mathcal{H}\big(P^X, P^X\big)$ is exactly the perplexity parameter
- Constrained coupling with a given degree of entropy

$$\begin{aligned} Z(X) &= \underset{Z, \mathcal{H}\left(P^X, P^X\right) = \mathrm{Perp}}{\arg\min} \left\{ \mathrm{KL}\big(P^X, Q^Z\big) \right\} \\ &= \arg \min_{Z, \mathcal{H}\left(P^X, P^X\right) = \mathrm{Perp}} \left\{ \mathcal{H}\big(P^X, Q^Z\big) - \mathrm{Perp} \right\} \end{aligned}$$

Perspectives

- The method is based on a preliminary smoothing of the data to retrieve a graph with controlled complexity
- This is related (how ?) to manifold learning and density estimation on manifolds
- The output $\widehat{Z}(X)$ strongly depends on this preliminary step

- Can we generalize the approach by matching arbitrary priors (power-law)
- Introduce clustering and spatial information in the framework

- How graph coupling could be restated in the RKHS ?

A Probabilistic Graph Coupling View of Dimension Reduction, van Assel, H. and Espinasse, T. and Chiquet, J. and Picard, F., NEURIPS 2022
`https://arxiv.org/pdf/2201.13053.pdf`

References

[1] J. N. Campbell, E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, E. D. Rosen, B. B. Lowell, and L. T. Tsai. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.*, 20(3):484–496, Mar 2017.

[2] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *bioRxiv*, 2018.

[3] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *Arxiv*, (1802.03426):1–63, 2018.

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.