Towards a generative model for Stochastic Neighbor Embedding

Julien Chiquet, Thibault Espinasse, Francois Gindraud, Franck Picard, Hugues van Assel

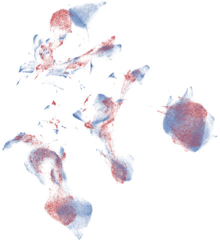
Institut Camille Jordan, CNRS Univ. Lyon
AgroParisTech INRA - MIA, Paris
Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

franck.picard@ens-lyon.fr

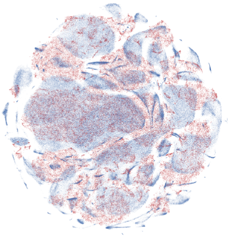
Outline

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]

- $(X_1, \dots, X_n)$ 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:

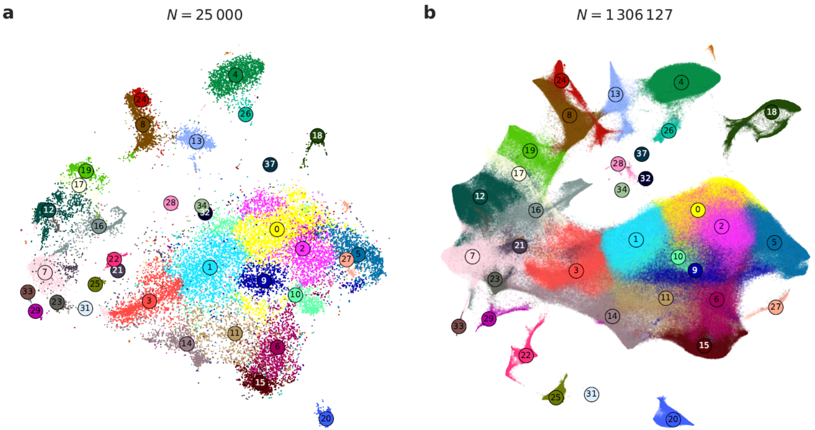
$$\forall (i,j) \in [n]^2, \quad q_{ij} = \left(1 + a\|X_i - X_j\|_2^{2b}\right)^{-1}$$

UMAP solves the following problem:

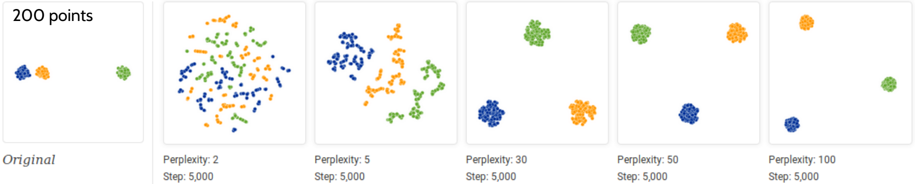
$$\min_{Z \in \mathbb{R}^{n \times d}} \quad -\sum_{i<j} p_{ij} \log q_{ij} + (1 - p_{ij}) \log(1 - q_{ij})$$

Outline

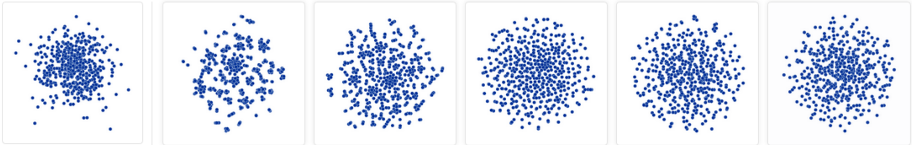
tSNE on single cell Gene Expression data [1]



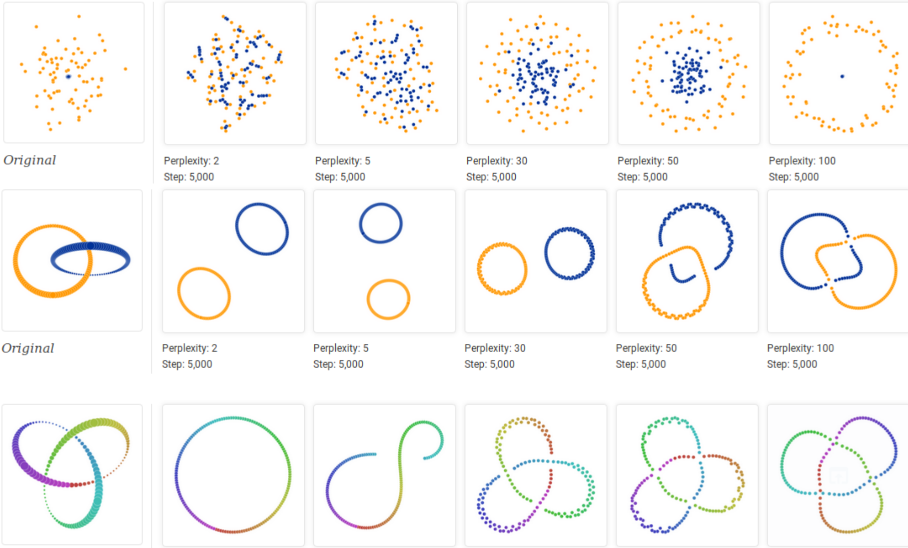
tSNE does not account for between-cluster distance



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)
- Sensitive 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

Motivations

- tSNE is defined by a quantity to optimize: Minimize the KL between p and q so that the data representation z minimizes:

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

- What is the underlying model ? p_{ij} proba of ?
- Could we improve the optimization algorithm if the underlying model was better defined ?
- Could we estimate the hyperparameters (smoothing) using ML ?
- Could we perform model selection ?

Markov Processes on a Graph for X

- Consider $G_X = (\mathcal{V}, \mathcal{E}_X)$ with $\mathcal{V} = \{1, \dots, n\}$ a set of nodes
- Nodes have attributes $(X_1, \dots, X_n)$ in $\mathbb{R}^p$
- **Main idea**: to any reversible Markov Process one can associate a symmetric graph, (reciprocal true).
- Introduce Y_X, a MP taking values in $\mathcal{V}$, s.t.

$$\mathbb{P}(Y_X(t+1) = j \mid Y_X(t) = i, X = x) = \Pi_X(i,j)$$

- X is fixed, no distribution assumption (kernel method)

Gaussian Transition Kernel on X

- We suppose that the transition kernel is of the form

$$\Pi_X(i,j) = \frac{k(x_i, x_j)}{d_X(i)}, \quad d_X(i) = \sum_{j=1}^{n} k(x_i, x_j)$$

- Π_X is not symmetric but has the conservation property:

$$\sum_{j=1}^{n} \Pi_X(i,j) = 1.$$

- Π_X is the 1-step transition matrix between points
- Stationary distribution of Y_X:

$$\mu_X \Pi_X = \mu_X, \quad \mu_X(i) = \frac{d_X(i)}{\bar{d}_X}, \quad \bar{d}_X = \sum_j d_X(j)$$

Markov Process on a Graph for Z

- Consider another graph $G_Z = (\mathcal{V}, \mathcal{E}_Z)$ with $\mathcal{V} = \{1, \dots, n\}$ (same)
- Z is the set of new attributed in $\mathbb{R}^q$ (unknown).
- Introduce a new MP Y_Z defined on $\{1, \dots, n\}$ s.t.

$$\mathbb{P}(Y_Z(t+1) = j \mid Y_Z(t) = i, Z = z) = \frac{h(z_i, z_j)}{d_Z(i)} = \Pi_Z(i, j)$$

- Z is fixed and considered as a parameter, but the form of the transition is specified

Gaussian or Student transition kernel on Z

- Suppose the new transition is of the form (Z unknown)

$$\Pi_Z(i,j) = \frac{h(z_i, z_j)}{d_Z(i)}$$

- We get close to tSNE by choosing

$$\begin{aligned} k(x_i, x_j) &= \exp\left(-\frac{1}{2\sigma}\|x_i - x_j\|^2\right) \\ h(z_i, z_j) &= \frac{1}{1 + \|z_i - z_j\|^2} \end{aligned}$$

- Suppose the two chains are conditionally independent

$$Y_X \perp Y_Z | X, Z$$

Maximum Coupling between Markov Processes

- Once the two chains specified, find Z by coupling the two processes

$$Z(X) = \max_{Z} \Big(\log \mathbb{P}(Y_X = Y_Z \mid X, Z) \Big)$$

- Maximizing the coupling between Y_X and Y_Z $\Leftrightarrow$ Minimizing the KL between Y_X and Y_Z

$$\mathbb{E}_{Y_X \sim \mu_X} \Big(\log \mathbb{P}(Y_Z = Y_X \mid X, Z) \Big) = \mathbb{E}_{Y \sim \mu_X} \Big(\log \mathbb{P}(Y_Z = Y \mid X, Z) \Big)$$

Minimum KL and Maximum Coupling

- The KL divergence between Markov Process

$$KL(Y_X, Y_Z) = \mathbb{E}_{Y \sim \mu_X}\Big(\log \mathbb{P}(Y_X = Y) \Big) - \mathbb{E}_{Y \sim \mu_X}\Big(\log \mathbb{P}(Y_Z = Y) \Big)$$

- Connection with the probability of coupling

$$\mathbb{E}_{Y_X \sim \mu_X}\Big(\log \mathbb{P}(Y_Z = Y_X) \Big) = \mathbb{E}_{Y \sim \mu_X}\Big(\log \mathbb{P}(Y_Z = Y) \Big)$$

- Minimizing the KL between chains wrt Z maximizes the probability of coupling

$$KL(Y_X, Y_Z) \quad = \quad -H_{\mu_X}(Y_X) - \mathbb{E}_{Y_X \sim \mu_X}\Big(\log \mathbb{P}(Y_Z = Y_X \mid X, Z) \Big)$$

Empirical Maximum Coupling

- To retrieve the hidden components:

$$Z_n(X) \quad = \quad \arg\max_{Z} \left[\widehat{\mathrm{H}}_{\mu_X}(Y_Z \mid X) \right],$$

- $\mathrm{H}_{\mu_X}(Y_Z \mid X, Z)$ stands for the entropy of chain Y_Z under μ_X with empirical version (fixed X)

$$\begin{aligned} \widehat{\mathrm{H}}_{\mu_X}(Y_Z \mid X) \quad &= \quad \sum_{i=1}^{n} \mu_X(i) \log \mu_Z(i) \\ &+ \quad \sum_{i=1}^{n} \mu_X(i) \left(\sum_{j=1}^{n} \Pi_X(i,j) \log \Pi_Z(i,j) \right) \end{aligned}$$

Specified transitions induce simplifications

$$d_X(i) = \sum_{j=1}^{n} k(x_i, x_j), \;\; d_Z(i) = \sum_{j=1}^{n} h(z_i, z_j)$$

$$\begin{aligned} \mu_X(i) &= d_X(i)/\bar{d}_X \quad \bar{d}_X = \sum_i d_X(i) \\ \mu_Z(i) &= d_Z(i)/\bar{d}_Z \quad \bar{d}_Z = \sum_i d_Z(i) \end{aligned}$$

and

$$\Pi_X(i,j) = \frac{k(X_i, X_j)}{d_X(i)}, \quad \Pi_Z(i) = \frac{h(Z_i, Z_j)}{d_Z(i)}$$

Then

$$\widehat{\mathrm{H}}_{\mu_X}(Y_Z \mid X) = \sum_{i,j} \frac{k(X_i, X_j)}{\bar{d}_X} \log \frac{h(Z_i, Z_j)}{\bar{d}_Z}$$

tSNE maximizes the coupling between Markov Processes

- If considering only KL minimization, the new representation would be such that:

$$\widehat{Z}_n(X) \quad = \quad \arg\max_Z \left[\sum_{i,j} \frac{k(X_i, X_j)}{\bar{d}_X} \log \frac{h(Z_i, Z_j)}{\bar{d}_Z} \right],$$

- $\bar{d}_X, \bar{d}_Z$ are normalization terms (different in tSNE - for now)
- The criterion is conditional to X
- interpretability of Z ? Representation of new Xs ?

Outline

Hidden Graph to structure observations

- Let us suppose that observations (rows) are structured thanks to a hidden random Graph
- $G = (V, E)$ with $V = \{1, \cdots, n\}$ the vertices

$$A_{ij} = \sum_{(k,\ell)\in E} \mathbb{1}_{(i,j)=(k,\ell)}, \quad L_G = D - A, \quad \text{where} \quad D_{ii} = \sum_j A_{ij}$$

- L_G, the Laplacian of G has the following property:

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

Conditional distribution of X on a graph

- Conditional model of the observations given the graph

$$X \mid G \sim \mathcal{MN}\left(0, L_G^{-1},\ R^{-1}\right),$$

- L_G^{-1} between-cell variability, R^{-1} between-genes correlation.
- Consider the Gaussian kernel for X

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

- Conditional distribution of $X \mid G$:

$$\mathbb{P}(X \mid G) \quad \propto \quad |L_G|^{p/2} \prod_{i,j=1}^{n} k(X_i, X_j)^{A_{ij}}$$

Conditional distribution of Z on a graph

- Consider that the low-dimensional representation is also structured according to a graph
- Consider the Gaussian kernel for Z

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

- Conditional distribution of $Z \mid G$:

$$\mathbb{P}(Z \mid G) \quad \propto \quad |L_G|^{q/2} \prod_{i,j=1}^{n} k(Z_i, Z_j)^{A_{ij}}$$

Embedding with Graph Coupling

- Consider two graphs G_X and G_Z
- Coupling with $G_X = G_Z$

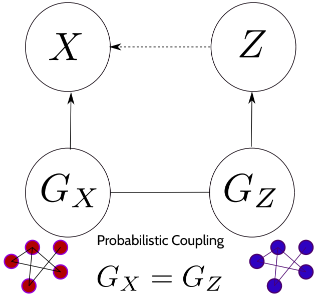
$$\mathbb{E}_{G \sim G_X}\Big(\log \mathbb{P}(G_Z = G_X \mid X, Z) \Big)$$

- which is equivalent to

$$\mathbb{E}_{G \sim G_X}\Big(\log \mathbb{P}(G_Z = G \mid X, Z) \Big)$$

- which is the entropy of G_Z under G_X

$$\mathrm{H}_{G_X}(G_Z \mid X, Z)$$



Graph Coupling with Z as a parameter

- Find the best Z such that the two graphs G_X and G_Z are as close as possible:

$$Z(X) \quad = \quad \arg\min_{Z} \left[\mathrm{H}_{G_X}(G_Z \mid X, Z) \right]$$

- The cross entropy between distribution of G_X and G_Z, which writes

$$\mathrm{H}_{G_X}(G_Z) = -\sum_{g} \mathbb{P}(G_X = g \mid X) \log \mathbb{P}(G_Z = g \mid Z).$$

- Challenge : define a prior distribution and deduce the posterior

Bernoulli prior distribution for G_X

- Let A_X be the adjacency matrix of G_X, with $A_{X,ij} \in \{0,1\}$

$$\mathbb{P}(G_X; \pi_X) = \frac{\left|L_X\right|^{-a_X/2} \times \prod_{i,j} \pi_{X,ij}^{A_{X,ij}}}{\sum_{A' \in \{0,1\}} \left|L_X(A')\right|^{-a_X/2} \times \prod_{i',j'} \pi_{X,i'j'}^{A'_{X,i'j'}}}$$

- $\left|L_{G_X}\right|^{-a_X/2}$ catches the dependency of connections wrt the graph.
- Retrieves conjugacy with the Gaussian conditional model
- Setting $a_X = 0$ leads to an independent Bernoulli prior

$$\mathbb{P}(A_{X,ij} = 1; \pi_X) = \frac{\pi_{X,ij}}{1 + \pi_{X,ij}}$$

Induced Posterior Distribution for G_X

- The posterior writes

$$\begin{aligned}\mathbb{P}(G_X \mid X; \pi_X) &\propto \mathbb{P}(G_X; \pi_X)\mathbb{P}(X \mid G_X; R) \\ &\propto \left|L_X\right|^{(p-a_X)/2} \prod_{ij} \Big(\pi_{X,ij} k(X_i, X_j; R)\Big)^{A_{X,ij}}\end{aligned}$$

- When $a_X = p$ we get independent Bernoulli posteriors

$$\mathbb{P}(A_{ij} = 1 \mid X; \pi) = \frac{\pi_{ij} k(X_i, X_j)}{1 + \pi_{ij} k(X_i, X_j)} = q_{\mathcal{B}}(X_i, X_j)$$

- When $a_X = 0$ we get an independent prior, but an intractable posterior

Maximum Coupling with the Bernoulli prior

$$
\begin{aligned}
\mathrm{KL}(G_X, G_Z) &= \sum_{ij} p_{\mathcal{B}}(X_i, X_j) \log \frac{p_{\mathcal{B}}(X_i, X_j)}{q_{\mathcal{B}}(Z_i, Z_j)} \\
&+ \sum_{ij} \Big(1 - p_{\mathcal{B}}(X_i, X_j)\Big) \log \frac{1 - p_{\mathcal{B}}(X_i, X_j)}{1 - q_{\mathcal{B}}(Z_i, Z_j)} \\
&= \mathrm{H}^{\mathcal{B}}_{G_X}(G_Z) \\
&+ \sum_{ij} p_{\mathcal{B}}(X_i, X_j) \log p_{\mathcal{B}}(X_i, X_j) \\
&+ \sum_{ij} \Big(1 - p_{\mathcal{B}}(X_i, X_j)\Big) \log \Big(1 - p_{\mathcal{B}}(X_i, X_j)\Big)
\end{aligned}
$$

$\rightarrow$ UMAP computes a KL (and not a cross entropy)

Fixed-degree prior distribution for G_X

- Denote by $D_{X,i}$ the degree of node i, consider

$$\mathbb{P}(G_X;\pi,D_X) \;\propto\; \left|L_{G_X}\right|^{-a_X/2}\prod_{i=1}^{n}\prod_{\ell=1}^{D_i}\pi_{i,e_{i\ell}}, \quad A_{X,ij}=\sum_{\ell=1}^{D_i}\mathbb{1}_{\{e_{i\ell}=j\}}$$

- Choosing $a_X = 0$ corresponds to a multinomial model:

$$A_{X,i1},\ldots,A_{X,in};D_{X,i} \;\sim\; \mathcal{M}\left\{D_{X,i};\left(\frac{\pi_{X,ij}}{\sum_{\ell=1}^{n}\pi_{X,i\ell}}\right)_j\right\},$$

- Choosing $a_X = p$ leads to

$$A_{X,i1},\ldots,A_{X,in}\mid X;D_{X,i} \;\sim\; \mathcal{M}\left\{D_{X,i};\left(\frac{\pi_{X,ij}k(X_i,X_j)}{\sum_{\ell=1}^{n}\pi_{X,ik}k(X_i,X_\ell)}\right)_j\right\},$$

tSNE and the Fixed-degree model

- In the following we will write:

$$p_D(X_i, X_j) = \frac{\pi_{ij} k(X_i, X_j)}{\sum_{\ell=1}^{n} \pi_{ij} k(X_i, X_\ell)}, \quad q_D(Z_i, Z_j) = \frac{\pi_{ij} k(Z_i, Z_j)}{\sum_{\ell=1}^{n} \pi_{ij} k(Z_i, Z_\ell)}.$$

- We retrieve the non-symmetric normalization term (Markov-like)
- With this prior we obtain the tSNE-like criterion

$$\mathrm{H}^{D}_{G_X}(G_Z) = -\sum_{i,j} D_{Xi} \Big\{ p_D\big(X_i, X_j\big) \log q_D(Z_i, Z_j) \Big\}$$

tSNE is defined for fixed X

- In the original method, the distribution of X is not modelled
- All quantities are defined conditionally to X
- This helps to choose $a_X = p$ and $a_z = q$ so that the posteriors p and q are factorized
- This allows to compute the cross entropy (sum)
- Master's internship:

$\rightarrow$ impact on Z of the different priors

$\rightarrow$ induced momentum algorithms for each prior

Outline

Symmetrization and directed graphs

- In the original formulation : $p_{ij} = (p_{i|j} + p_{j|i})/2N$
- What probabilistic model should we consider to obtain the same symmetrization with our posteriors ?
- Considering an oriented graph with symmetrized Laplacian

$$\begin{cases} L_{ij} = -(A_{ij} + A_{ji})/2 & \text{if } i \neq j \\ L_{ii} = (A_{i+} + A_{+i})/2 \end{cases}$$

- How to get to a symmetrized posterior from here ?
- interpretation of the underlying directed graph ?

Kernel calibration and Perplexity

- tSNE strongly depends on the calibration of the kernel

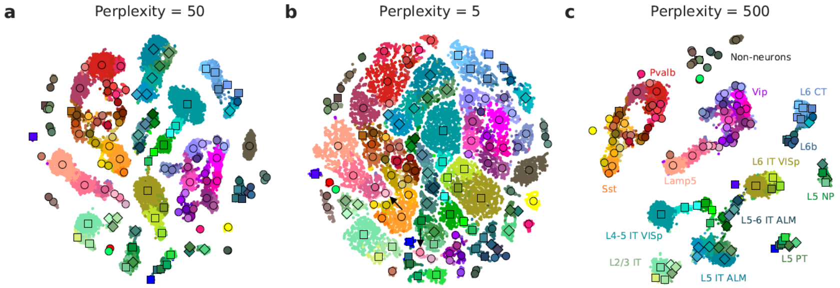
$$k(X_i, X_j; \sigma_i) = \exp\left(-\frac{1}{2\sigma_i}\|X_i - X_j\|_R^2\right),$$

- 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[1]



Connecting the kernel bandwidth with the graph model

- Consider $D = \text{diag}(d_1, \dots, d_n)$ the matrix of degrees
- Consider the random walk laplacian is defined by:

$$L^{RW} = D^{-1}L$$

- The following property holds:

$$\forall X \in \mathbb{R}^n, \ \text{tr}(X^T L^{RW} X) = \frac{1}{2}\sum_{i,j} A_{i,j}\frac{\|X_i - X_j\|^2}{d_i}$$

- Hence we can consider

$$X_{n,p} \mid G_X \sim \mathcal{MN}_{n,p}\left(0, \left(L^{RW}\right)^{-1}, R^{-1}\right)$$

Back to the coupling strategy

- Maximizing the probability of coupling by minimizing the KL

$$\mathrm{KL}(G_X, G_Z) = \mathrm{H}_{G_X}(G_Z) - \mathrm{H}_{G_X}(G_X)$$

- $\mathrm{H}_{G_X}(G_X)$ is exactly the perplexity parameter
- Constrained coupling with a given degree of entropy

$$\begin{aligned} Z(X) &= \underset{Z, \mathrm{H}_{G_X}(G_X) = \mathrm{Perp}}{\arg\min} \Big[\mathrm{KL}(G_X, G_Z) \Big] \\ &= \arg \min_{Z, \mathrm{H}_{G_X}(G_X) = \mathrm{Perp}} \Big(\mathrm{H}_{G_X}(G_Z) - \mathrm{Perp} \Big) \end{aligned}$$

Connection with Nearest Neighbors Graphs and Manifold Learning

- 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

$$\widehat{Z}_{\text{Perp}}(X) = \arg\min_{Z} \left(\mathrm{H}_{\widehat{G}_{X,\text{Perp}}}(G_Z) \right)$$

Maximum Likelihood inference for SNE ?

- Define the observed X and hidden G, Z variables
- Define the observed-data likelihood : $\mathbb{P}(X)$
- Define the conditional distribution : $\mathbb{P}(X \mid G, Z)$
- Define the prior distribution $\mathbb{P}(G, Z)$
- Compute the conditional expectation of the complete-data loglik

$$\mathcal{Q} = \mathbb{E}_{G,Z|X}\Big(\log \mathbb{P}(X, G, Z) \Big)$$

- Compute the posterior

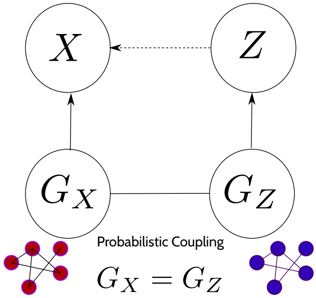
$$\log \mathbb{P}(G, Z \mid X)$$

The two-graph model is not identifiable

- Coupling with $G_X = G_Z$

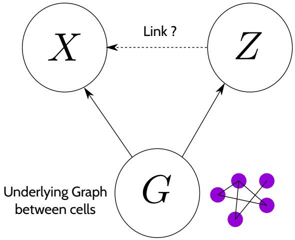
  $$\log \mathbb{P}(X, Z, G_X, G_Z, G_X = G_Z)$$

- Discrepancy between two priors and posterior
- Difficult to model a link between X and Z
- Non identifiable model



The one-graph model

- One prior that rules them all
- Different priors for G (Bernoulli, fixed number of edges, fixed degree)
- Identifiable model but computational issues
- tSNE strategy : Z is a parameter



When a_X and a_Z come back

- The joint likelihood of the model:

$$\log \mathbb{P}(X, G \mid Z) \quad = \quad \log \mathbb{P}(X \mid G, Z) + \log \mathbb{P}(G \mid Z)$$

- In the EM framework, $\mathcal{Q}$ becomes

$$\mathcal{Q}_Z = \mathbb{E}_{G|X}\Big(\log \mathbb{P}(X \mid G, Z) + \log \mathbb{P}(G \mid Z) \Big)$$

- $\widehat{Z}$ maximizes the posterior probability of connection

$$\widehat{Z} = \arg\max_Z \Big(\mathcal{Q}_Z \Big) = \arg\max_Z \Big\{ \mathbb{E}_{G|X}\Big(\log \mathbb{P}(G \mid Z) \Big) \Big\}$$

- Involves the tricky term

$$\mathbb{E}_{G|X}\Big(|L_G| \Big)$$

Connections with the fixed graph model [2]

- Consider the Multivariate Gaussian Model

$$X_i \sim \mathcal{N}(\mu_i, \Sigma), \quad \mu_i \in \mathbb{R}^p \quad \Sigma \in \mathcal{S}_+^p \quad i = 1, 2, ..., n$$

- Consider that the observations are connected by a given graph G
- Regularized Mean estimation problem:

$$\widehat{M}_\alpha = \underset{M}{\text{argmin}} \, \|X - M\|_F^2 + \alpha \text{tr}(M^T \mathcal{L}_\mathcal{S} M)$$

where $\mathcal{L}_\mathcal{S} = \frac{D-A}{\frac{1}{n}\sum_i d_i}$

- In our setting, would it be $X \mid \mu$, $\mu \sim \mathcal{N}\big(0, \tau\big)$?

References

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

[2] Tianxi Li, Cheng Qian, Elizaveta Levina, and Ji Zhu. High-dimensional gaussian graphical models on network-linked data. *Journal of Machine Learning Research*, 21(74):1–45, 2020.

[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.