A Probabilistic Graph Coupling View of Dimension Reduction

Hugues Van Assel⋆, Thibault Espinasse†, Julien Chiquet‡ and Franck Picard⋆

⋆ : ENS Lyon, † : Institut Camille Jordan Lyon 1, ‡ : INRAE, Université Paris-Saclay

Dimension Reduction

$$\boldsymbol{X} \in \mathbb{R}^{n \times p} \to \boldsymbol{Z} \in \mathbb{R}^{n \times q}$$

Spectral methods. Performs an eigendecomposition of a similarity matrix, can be framed in the kernel PCA framework.

- Linear : PCA, MDS
- Non-linear : Laplacian Eigenmaps, Isomap, LLE, Diffusion maps ...

Neighbor Embedding (NE) methods. Matches similarities defined in both input and latent spaces.

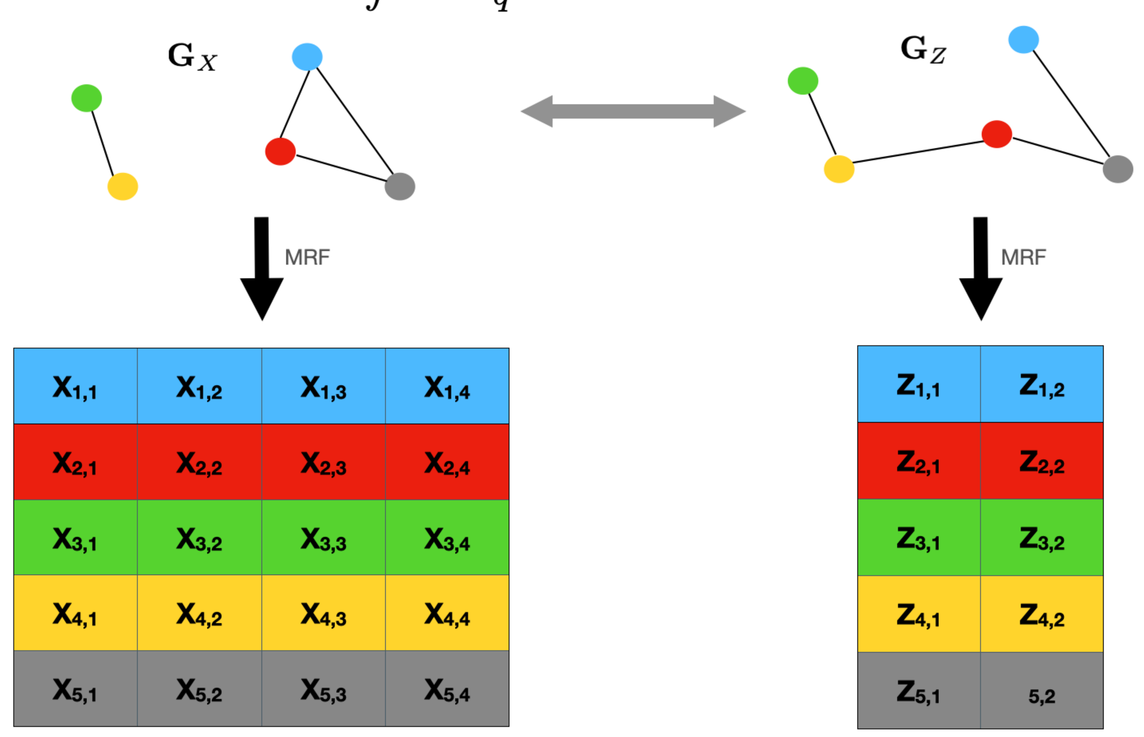
- SNE, t-SNE, UMAP, largeVis

Is there a common probabilistic model?

Graph Coupling Model

$$\min_{\mathbf{Z}} \mathrm{KL}(\mathbb{P}(\mathbf{G}_X \mid \mathbf{X}) || \mathbb{P}(\mathbf{G}_Z \mid \mathbf{Z}))$$

$$\mathrm{KL}(P||Q) = \int p \log \frac{p}{q} d\lambda \quad \text{where} \quad dP = pd\lambda, \quad dQ = qd\lambda$$



Bayesian Model

$$\mathbb{P}(\boldsymbol{G}_X|\boldsymbol{X}) \propto \underbrace{\mathbb{P}(\boldsymbol{X}|\boldsymbol{G}_X)}_{\text{Conditional}} \underbrace{\mathbb{P}(\boldsymbol{G}_X)}_{\text{Prior}}$$

- The conditional takes the same form across all methods (pairwise MRF).
- The graph priors characterize each method. There are two types:
  - discrete graphs with simple topological constraints (NE).
  - positive definite matrices (Spectral).

Neighbor Embedding Methods

Algorithm	Input Similarity	Latent Similarity	Loss Function
SNE	$P^D_{ij} = \frac{k_x(\boldsymbol{X}_i - \boldsymbol{X}_j)}{\sum_\ell k_x(\boldsymbol{X}_i - \boldsymbol{X}_\ell)}$	$Q^D_{ij} = \frac{k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}{\sum_\ell k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_\ell)}$	$-\sum_{i \neq j} P^D_{ij} \log Q^D_{ij}$
Sym-SNE	$\overline{P}^D_{ij} = P^D_{ij} + P^D_{ji}$	$Q^E_{ij} = \frac{k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}{\sum_{\ell,t} k_z(\boldsymbol{Z}_\ell - \boldsymbol{Z}_t)}$	$-\sum_{i<j} \overline{P}^D_{ij} \log Q^E_{ij}$
LargeVis	$\overline{P}^D_{ij} = P^D_{ij} + P^D_{ji}$	$Q^B_{ij} = \frac{k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}{1 + k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}$	$-\sum_{i<j} \overline{P}^D_{ij} \log Q^B_{ij} + \left(2 - \overline{P}^D_{ij}\right) \log(1 - Q^B_{ij})$
UMAP	$\widetilde{P}^B_{ij} = P^B_{ij} + P^B_{ji} - P^B_{ij} P^B_{ji}$	$Q^B_{ij} = \frac{k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}{1 + k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}$	$-\sum_{i<j} \widetilde{P}^B_{ij} \log Q^B_{ij} + \left(1 - \widetilde{P}^B_{ij}\right) \log(1 - Q^B_{ij})$

NE Methods as Graph Coupling

Let k be even and positive, we consider the conditional:

$$\mathbb{P}(\boldsymbol{X}|\boldsymbol{W}) \propto \prod_{ij} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{W_{ij}} .$$

Gaussian kernel. $k : \boldsymbol{x} \mapsto \exp\left(-\|\boldsymbol{x}\|_2^2\right)$. In this case, the pairwise MRF is a matrix normal distribution with among row precision $\boldsymbol{L}$ (graph Laplacian of $\boldsymbol{W}$): $vec(\boldsymbol{X}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{L}^\dagger \otimes \boldsymbol{I}_p)$.

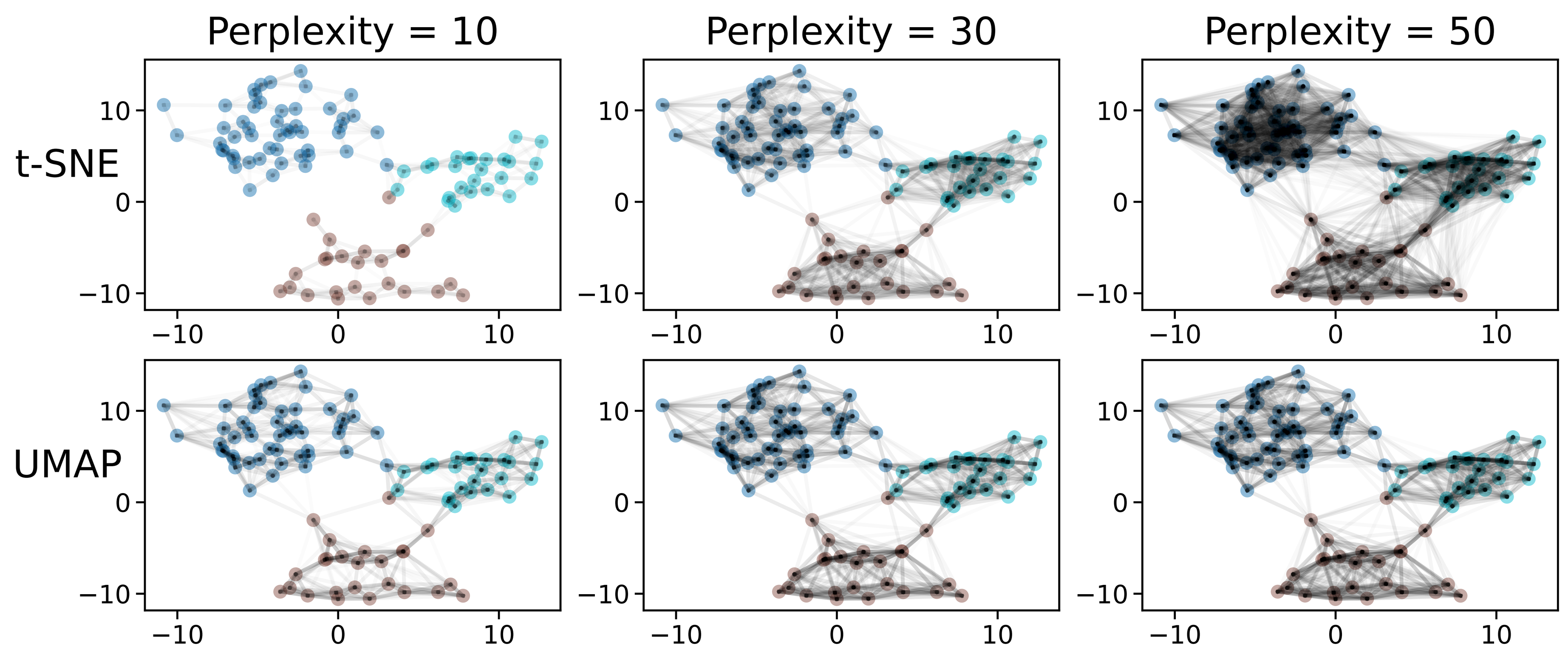
Priors. For $\boldsymbol{W}_X$ and $\boldsymbol{W}_Z$, we consider priors that are conjugate with the pairwise MRF likelihood plus the following topological constraints.

- B : binary edges.
- D : outdegree 1 for each node.
- E : n total edges.

$\mathcal{P}_Z$, $\mathcal{P}_X$	B	D	E
B	UMAP		
D	LargeVis	SNE	Sym-SNE

One can retrieve the losses of Neighbor Embedding methods as (visualization of posteriors below)

$$-\mathbb{E}_{\boldsymbol{W}_X \sim \mathbb{P}(\cdot|\boldsymbol{X})} \left[\log \mathbb{P}(\boldsymbol{W}_Z = \boldsymbol{W}_X|\boldsymbol{Z})\right] .$$



Large Scale Deficiency

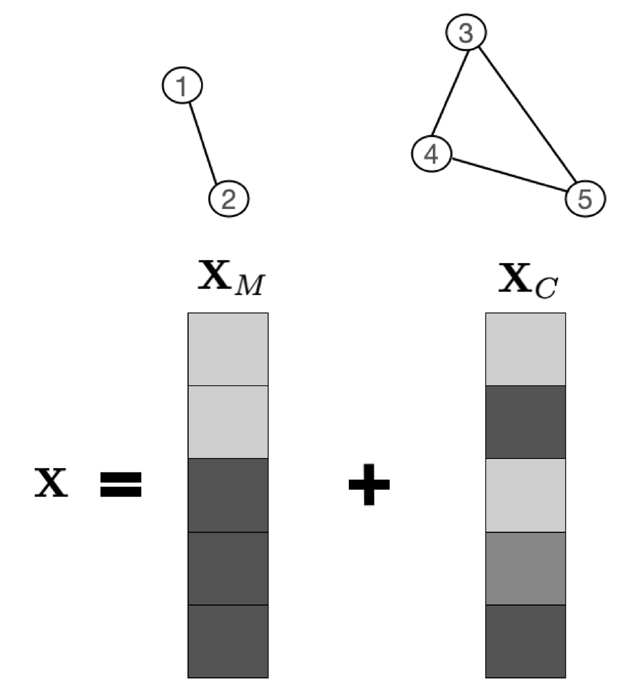
Integrability. If k is $\mathbb{R}^p$-integrable and bounded above, then $\boldsymbol{X} \mapsto \prod_{ij} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{W_{ij}}$ is integrable on $(\ker \boldsymbol{L})^\perp \otimes \mathbb{R}^p$ where $\boldsymbol{L}$ is the graph Laplacian of $\boldsymbol{W}$.

Gaussian kernel. $\mathcal{N}(\mathbf{0}, \boldsymbol{L}^\dagger \otimes \boldsymbol{I}_p)$ only defines a probability on $(\ker \boldsymbol{L})^\perp \otimes \mathbb{R}^p$.

Let $\boldsymbol{X}_M = \mathrm{Proj}_{(\ker \boldsymbol{L}) \otimes \mathbb{R}^p}(\boldsymbol{X})$ and $\boldsymbol{X}_C = \mathrm{Proj}_{(\ker \boldsymbol{L})^\perp \otimes \mathbb{R}^p}(\boldsymbol{X})$.

- $\boldsymbol{X}_M$ is the mean of $\boldsymbol{X}$ on $\boldsymbol{W}$'s CCs.
- $\boldsymbol{X}_C$ is centered on the CCs of $\boldsymbol{W}$.

$\boldsymbol{X}_C$ is structured by the model unlike $\boldsymbol{X}_M$.



PCA as Graph Coupling

Wishart distribution: denoted $\boldsymbol{\Theta} \sim \mathcal{W}(\nu, \boldsymbol{\Pi})$ for

$$\mathbb{P}(\boldsymbol{\Theta}; \nu, \boldsymbol{\Pi}) \propto |\boldsymbol{\Theta}|^{\frac{\nu}{2}} e^{-\frac{1}{2}\langle \boldsymbol{\Pi}, \boldsymbol{\Theta} \rangle} .$$

Let $\nu > 0$, $\boldsymbol{\Theta}_X \sim \mathcal{W}(\nu, \boldsymbol{I}_n)$ and $\boldsymbol{\Theta}_Z \sim \mathcal{W}(\nu + p - q, \boldsymbol{I}_n)$. If $\boldsymbol{\Theta}_X$ and $\boldsymbol{\Theta}_Z$ structure the rows of respectively $\boldsymbol{X}$ and $\boldsymbol{Z}$ such that:

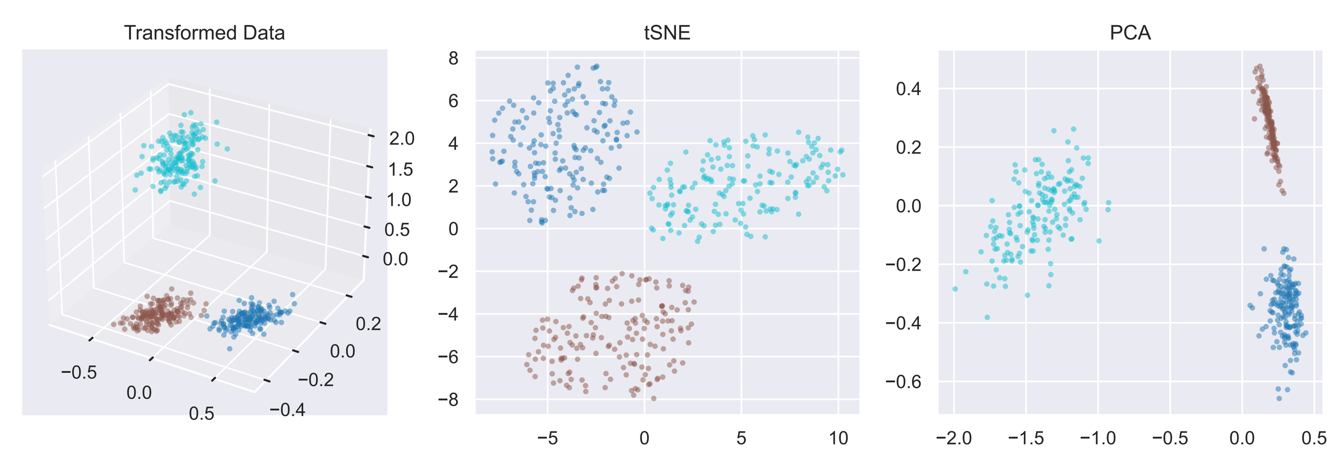
$$\mathrm{vec}(\boldsymbol{X})|\boldsymbol{\Theta}_X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_X^{-1} \otimes \boldsymbol{I}_p)$$
$$\mathrm{vec}(\boldsymbol{Z})|\boldsymbol{\Theta}_Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q) .$$

Then the solution of the precision coupling problem:

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{n \times q}} \mathrm{KL}(\mathbb{P}(\boldsymbol{\Theta}_X|\boldsymbol{X}) || \mathbb{P}(\boldsymbol{\Theta}_Z|\boldsymbol{Z}))$$

is a PCA embedding of $\boldsymbol{X}$ with q components.

PCA vs t-SNE



- PCA represents better the clusters' positions (inter-cluster variability).
- t-SNE is better at representing the local structure (intra-cluster variability).