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, UMR MIA Paris-Saclay

Keywords: Probabilistic Modeling, Dimension Reduction, Random Graphs, Graphical Models

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 kernel matrix. These methods can be framed in the kernel PCA framework:

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

SNE-like methods. Defines similarities in both input and latent spaces and matches them through a non-convex loss optimized by gradient descent.

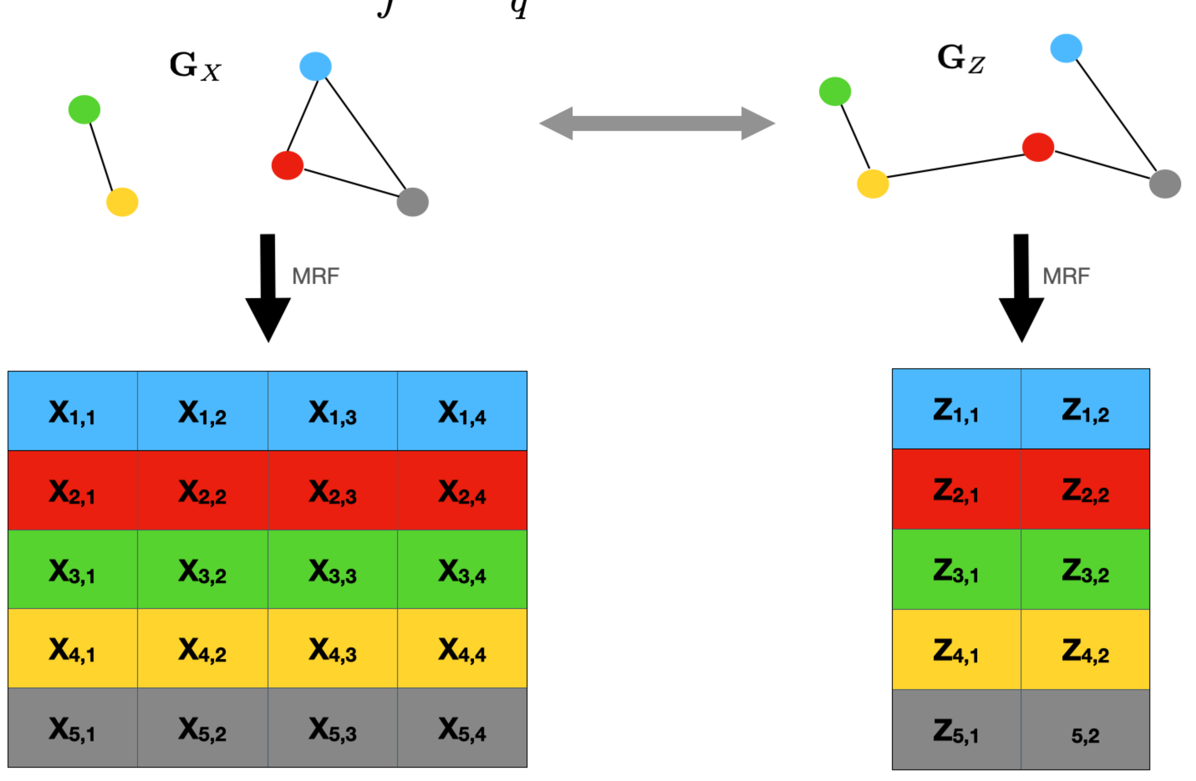
- SNE, t-SNE, UMAP, largeVis

Is there a common probabilistic model?

Graph Coupling Model

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

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



General Idea

Applying Bayes rule:

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

- **The likelihood takes the same form across all the DR methods** but can sometimes be degenerate.
- **What characterize each method are the priors** considered for the latent structuring graphs $\boldsymbol{G}_X$ and $\boldsymbol{G}_Z$.

Pairwise Markov Random Field Likelihood

$$\mathbb{P}(\boldsymbol{X}|\boldsymbol{G}_X) \propto \prod_{i \overset{G_X}{\sim} j} \Psi_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j)$$

- two nodes that are not connected are conditionally independent given all other nodes.

PCA as Graph Coupling

Let $\nu \geq n$, $\boldsymbol{\Theta}_X \sim \mathcal{W}(\nu, \boldsymbol{I}_n)$ 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:

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

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.

SNE-like 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	$\widetilde{P}^B_{ij} = P^B_{ij} + P^B_{ji} - P^B_{ij} P^B_{ji}$	$Q^B_{ij} = \frac{k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}{1 + k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j)}$	$-\sum_{i<j} \widetilde{P}^B_{ij} \log Q^B_{ij} + \left(1 - \widetilde{P}^B_{ij}\right) \log(1 - Q^B_{ij})$

SNE-like Methods as Graph Coupling

Likelihood with shift-invariant kernels:

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

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 $\boldsymbol{W}$.

Graph Priors: Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n \times n}$, $\alpha \in \mathbb{R}$, k be a bounded integrable kernel and $\mathcal{P} \in \{B, D, E\}$.

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

where $\Omega_B(\boldsymbol{W}) = \prod_{ij} 1_{W_{ij} \leq 1}$, $\Omega_D(\boldsymbol{W}) = \prod_i 1_{W_{i+}=1}$ and $\Omega_E(\boldsymbol{W}) = 1_{W_{++}=n} \prod_{ij} (W_{ij}!)^{-1}$ and $\mathcal{C}_k(\boldsymbol{W}) = \int_{\mathcal{X}} \prod_{i \neq j} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{W_{ij}} d\boldsymbol{X}$.

If $\boldsymbol{W} \sim \mathbb{P}_{\mathcal{P},k}(\cdot\,; \boldsymbol{1}, 1)$ then

$$\boldsymbol{W}|\boldsymbol{X} \sim \mathbb{P}^\star_{\mathcal{P}}(\cdot\,; \boldsymbol{K}) \,.$$

which is defined as:

- if $\mathcal{P} = B$, $\forall (i,j) \in [n]^2$, $W_{ij} \overset{\perp}{\sim} \mathcal{B}(K_{ij}/(1 + K_{ij}))$.
- if $\mathcal{P} = D$, $\forall i \in [n]$, $\boldsymbol{W}_i \overset{\perp}{\sim} \mathcal{M}(1, \boldsymbol{K}_i / K_{i+})$.
- if $\mathcal{P} = E$, $\boldsymbol{W} \sim \mathcal{M}(n, \boldsymbol{K}/K_{++})$.

For $(\mathcal{P}_X, \mathcal{P}_Z) \in \{B, D, E\}^2$, we retrieve the losses of SNE-like methods as $\mathrm{KL}(\mathbb{P}^\star_{\mathcal{P}_X}(\cdot; \boldsymbol{K}_X) || \mathbb{P}^\star_{\mathcal{P}_Z}(\cdot; \boldsymbol{K}_Z))$:

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

Large Scale Deficiency

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 the CCs of $\overline{\boldsymbol{W}}$.
- $\boldsymbol{X}_C$ is centered on the CCs of $\overline{\boldsymbol{W}}$.

$\boldsymbol{X}_C$ is structured by the model unlike $\boldsymbol{X}_M$ which is taken from a distribution with infinite variance.