Unsupervised Learning with Probabilistic Modelling and Optimal Transport

Hugues Van Assel
ENS Lyon

February 20, 2024

Contents

Acknowledgements

Allez VA.

1

Introduction

Notation

- $[\![n]\!]$: set of integers $\{1, \dots, n\}$.
- $\mathbf{1}_n$: vector of $\mathbb{R}^n$ with all entries identically set to 1.
- $\mathbf{I}_n$: identity matrix of size $n \times n$.
- exp and log applied to vectors/matrices are taken element-wise.
- $\mathcal{S}^n$: space of $n \times n$ symmetric matrices.
- $[\mathbf{P}]_{ij}$ or P_{ij}: (i, j)-th entry of a matrix $\mathbf{P}$.
- $\mathbf{P}_{i:}$ or $[\mathbf{P}]_{i:}$: i-th row of a matrix $\mathbf{P}$.
- $\odot$ (*resp.* $\oslash$): element-wise multiplication (*resp.* division) between vectors/matrices.

Acronyms.

- DR: Dimensionality Reduction.
- OT: Optimal Transport.
- GW: Gromov-Wasserstein.

Notations. $[\![n]\!]$ denotes the set $\{1, ..., n\}$. exp and log applied to vectors/matrices are taken element-wise. $\mathbf{1} = (1, ..., 1)^\top$ is the vector of 1. $\langle \cdot, \cdot \rangle$ is the standard inner product for matrices/vectors. $\mathcal{S}$ is the space of $n \times n$ symmetric matrices. $\mathbf{P}_{i:}$ denotes the i-th row of a matrix $\mathbf{P}$. $\odot$ (*resp.* $\oslash$) stands for element-wise multiplication (*resp.* division) between vectors/matrices. For $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n, \boldsymbol{\alpha} \oplus \boldsymbol{\beta} \in \mathbb{R}^{n \times n}$ is $(\alpha_i + \beta_j)_{ij}$. The entropy of $\mathbf{p} \in \mathbb{R}^n_+$ is[1] $\mathrm{H}(\mathbf{p}) = -\sum_i p_i(\log(p_i) - 1) = -\langle \mathbf{p}, \log \mathbf{p} - \mathbf{1} \rangle$. The Kullback-Leibler divergence between two matrices $\mathbf{P}, \mathbf{Q}$ with nonnegative entries such that $Q_{ij} = 0 \implies P_{ij} = 0$ is $\mathrm{KL}(\mathbf{P}|\mathbf{Q}) = \sum_{ij} P_{ij} \left(\log(\frac{P_{ij}}{Q_{ij}}) - 1 \right) = \langle \mathbf{P}, \log\left(\mathbf{P} \oslash \mathbf{Q}\right) - \mathbf{1}\mathbf{1}^\top \rangle$.

The i^{th} entry of a vector $\mathbf{v}$ is denoted as either v_i or $[\mathbf{v}]_i$. Similarly, for a matrix $\mathbf{M}$, M_{ij} and $[\mathbf{M}]_{ij}$ both denote its entry (i, j). S_N is the set of permutations of $[\![N]\!]$. $P_N(\mathbb{R}^d)$ refers to the set of discrete probability measures composed of N points of $\mathbb{R}^d$. Σ_N stands for the probability simplex of size N that is $\Sigma_N := \{\mathbf{h} \in \mathbb{R}^N_+ \text{ s.t. } \sum_i h_i = 1\}$. In all the paper, $\log(\mathbf{M}), \exp(\mathbf{M})$ are to be understood element-wise. For $\mathbf{x} \in \mathbb{R}^N, \mathrm{diag}(\mathbf{x})$ denotes the diagonal matrix whose elements are the x_i.

We denote by $\mathcal{U}(\mathbf{h}) = \left\{ \mathbf{T} \in \mathbb{R}^{N \times n}_+ | \mathbf{T}\mathbf{1}_n = \mathbf{h} \right\}$ and $\mathcal{U}(\mathbf{h}, \overline{\mathbf{h}}) = \left\{ \mathbf{T} \in \mathbb{R}^{N \times n}_+ | \mathbf{T}\mathbf{1}_n = \mathbf{h}, \mathbf{T}^\top \mathbf{1}_N = \overline{\mathbf{h}} \right\}$ the set of discrete couplings with respectively one and two marginals. $L_2(x, y) \coloneqq \frac{1}{2}|x - y|^2$ is the quadratic loss, $L_{\mathrm{KL}}(x, y) \coloneqq x \log(x/y)$ the Kullback-Leibler divergence and $L_{\mathrm{GKL}}(x, y) \coloneqq L_{\mathrm{KL}}(x, y) - x + y$ the generalized one. Throughout, $\overline{\mu} = \frac{1}{N} \sum_{i \in [\![N]\!]} \delta_{\boldsymbol{x}_i}$ denotes the empirical data measure. DS is the space of $N \times N$ doubly stochastic matrices. $\mathrm{DS} = \mathcal{U}(\mathbf{1}, \mathbf{1})$ is the space of $N \times N$ doubly stochastic matrices.

For a matrix $\mathbf{Z} \in \mathbb{R}^{N \times d}$, we denote by $\mathbf{C}_\mathbf{Z} = \mathbf{C}_Z(\mathbf{Z})$ its output through a similarity function $\mathbf{C}_Z : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times N}$.

[1] With the convention $0 \log 0 = 0$.

2

Background on Dimensionality Reduction and Optimal Transport

Idées pour Background:

- 2 regards : Discuter approches modélisation probabiliste vs optim.
- Manque de fondements pour le non-supervisé (notamment pas de perte, pas de mesure de performance).

Lien DR et OT:

- minimal Wasserstein estimators (Rosasco et papier de Vayer/Gribonval.)
- liens clustering - Gromov (papier de Chen).
- liens clustering spectral - doubly stochastic (zass sashua).

2.1 Affinity-Based Unsupervised Learning

One major objective of unsupervised learning Hastie et al. [2009] is to provide interpretable and meaningful approximate representations of the data that best preserve its structure *i.e.* the underlying geometric relationships between the data samples. Similar in essence to Occam's principle frequently employed in supervised learning, the preference for unsupervised data representation often aligns with the pursuit of simplicity, interpretability or visualizability in the associated model. These aspects are determinant in many real-world applications, such as cell biology Cantini et al. [2021], Ventre et al. [2023], where the interaction with domain experts is paramount for interpreting the results and extracting meaningful insights from the model.

Dimensionality reduction and clustering. When faced with the question of extracting interpretable representations, from a dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^\top \in \mathbb{R}^{N \times p}$ of N samples in $\mathbb{R}^p$, the machine learning community has proposed a variety of methods. Among them, dimensionality reduction (DR) algorithms have been widely used to summarize data in a low-dimensional space $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)^\top \in \mathbb{R}^{N \times d}$ with $d \ll p$, allowing for visualization of every individual points for small enough d Agrawal et al. [2021], Van Der Maaten et al. [2009]. Another major approach is to cluster the data into n groups, with n typically much smaller than N, and to summarize these groups through their centroids Saxena et al. [2017], Ezugwu et al. [2022]. Clustering is particularly interpretable since it provides a smaller number of points that can be easily inspected. The cluster assignments can also be analyzed. Both DR and clustering follow a similar philosophy of summarization and reduction of the dataset using a smaller size representation.

Two sides of the same coin. As a matter of fact, methods from both families share many similitudes, including the construction of a similarity graph between input samples. In clustering, many popular approaches design a reduced or coarsened version of the initial similarity graph while preserving some of its spectral properties Von Luxburg [2007], Schaeffer [2007]. In DR, the goal is to solve the inverse problem of finding low-dimensional embeddings that generate a similarity graph close to the one computed from input data points Ham et al. [2004], Hinton and Roweis [2002]. Our work builds on these converging viewpoints and addresses the following question: *can DR and clustering be expressed in a common and unified framework ?*

2.2 Dimensionality Reduction

Dimensionality reduction (DR) is of central importance when dealing with high-dimensional data [Donoho, 2000]. It mitigates the curse of dimensionality, allowing for greater statistical flexibility and less computational complexity. DR also enables visualization that can be of great practical interest for understanding and interpreting the structure of large datasets. Most seminal approaches include Principal Component Analysis (PCA) Pearson [1901], multidimensional scaling Kruskal [1978] and more broadly kernel eigenmaps methods such as Isomap Balasubramanian et al. [2002], Laplacian eigenmaps [Belkin and Niyogi, 2003] and diffusion maps [Coifman and Lafon, 2006]. These methods share the definition of a pairwise similarity kernel that assigns a high value to close neighbors and the resolution of a spectral problem. They are well understood and unified in the kernel PCA framework [Ham et al., 2004].

In the past decade, the field has witnessed a major shift with the emergence of a new class of methods. They are also based on pairwise similarities but these are not converted into inner products. Instead, they define pairwise similarity functions

in both input and latent spaces and optimize a cost between the two. Among such methods, the Stochastic Neighbor Embedding (SNE) algorithm Hinton and Roweis [2003], its heavy-tailed symmetrized version t-SNE van der Maaten and Hinton [2008] or more recent approaches like LargeVis Tang et al. [2016] and UMAP McInnes et al. [2018] are arguably the most used in practice. These will be referred to as *SNE-like* or *neighbor embedding* methods in what follows. They are increasingly popular and now considered state-of-art techniques in many fields Li et al. [2017], Kobak and Berens [2019], Anders et al. [2018]. Their popularity is mainly due to their exceptional ability to preserve the local structure, *i.e.* close points in the input space have close embeddings, as shown empirically Wang et al. [2021]. They also demonstrate impressive performances in identifying clusters Arora et al. [2018], Linderman and Steinerberger [2019]. However this is done at the expense of global structure, that these methods struggle in preserving Wattenberg et al. [2016], Coenen and Pearce [2019] *i.e.* the relative large-scale distances between embedded points do not necessarily correspond to the original ones.

Given a dataset $\mathbf{X} \in \mathbb{R}^{n\times p}$ of n samples in dimension p, most DR algorithms compute a representation of $\mathbf{X}$ in a lower-dimensional latent space $\mathbf{Z} \in \mathbb{R}^{n\times q}$ with $q \ll p$ that faithfully captures and represents pairwise dependencies between the samples (or rows) in $\mathbf{X}$. This is generally achieved by optimizing $\mathbf{Z}$ such that the corresponding affinity matrix matches another affinity matrix defined from $\mathbf{X}$. These affinities are constructed from a matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$ that encodes a notion of "distance" between the samples, *e.g.* the squared Euclidean distance $C_{ij} = \|\mathbf{X}_{i:} - \mathbf{X}_{j:}\|_2^2$ or more generally any *cost matrix* $\mathbf{C} \in \mathcal{D} := \{\mathbf{C} \in \mathbb{R}_+^{n\times n} : \mathbf{C} = \mathbf{C}^\top \text{and } C_{ij} = 0 \iff i = j\}$. A commonly used option is the Gaussian affinity that is obtained by performing row-wise normalization of the kernel $\exp(-\mathbf{C}/\varepsilon)$, where $\varepsilon > 0$ is the bandwidth parameter.

Exploring and analyzing high-dimensional data is a core problem of data science that requires building low-dimensional and interpretable representations of the data through dimensionality reduction (DR). Ideally, these representations should preserve the data structure by mimicking, in the reduced representation space (called *latent space*), a notion of similarity between samples. We call *affinity* the weight matrix of a graph that encodes this similarity. It has positive entries and the higher the weight in position (i, j), the higher the similarity or proximity between samples i and j. Seminal approaches relying on affinities include Laplacian eigenmaps [Belkin and Niyogi, 2003], spectral clustering [Von Luxburg, 2007] and semi-supervised learning [Zhou et al., 2003]. Numerous methods can be employed to construct such affinities. A common choice is to use a kernel (*e.g.* Gaussian) derived from a distance matrix normalized by a bandwidth parameter that usually has a large influence on the outcome of the algorithm. Indeed, excessively small kernel bandwidth can result in solely capturing the positions of closest neighbors, at the expense of large-scale dependencies. Inversely, setting too large a bandwidth blurs information about close-range pairwise relations.

Ideally, one should select a different bandwidth for each point to accommodate varying sampling densities and noise levels. One approach is to compute the bandwidth of a point based on the distance from its k-th nearest neighbor Zelnik-Manor and Perona [2004]. However, this method fails to consider the entire distribution of distances. In general, selecting appropriate kernel bandwidths can be a laborious task, and many practitioners resort to greedy search methods. This can be limiting in some settings, particularly when dealing with large sample sizes.

Neighbor embedding methods. An alternative group of methods relies on neighbor embedding techniques which consists in minimizing in $\mathbf{Z}$ the quantity

$$\sum_{(i,j)\in[\![N]\!]^2} L_{\mathrm{KL}}([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}) \,. \tag{NE}$$

Within our framework, this corresponds to eq. (DR) with $L = L_{\mathrm{KL}}$. The objective function of popular methods such as stochastic neighbor embedding (SNE) Hinton and Roweis [2002] or t-SNE Van der Maaten and Hinton [2008] can be derived from eq. (NE) with a particular choice of $\mathbf{C}_X, \mathbf{C}_Z$. For instance SNE and t-SNE both consider in the input space a symmetrized version of the entropic affinity Vladymyrov and Carreira-Perpinan [2013], Van Assel et al. [2023]. In the embedding space, $\mathbf{C}_Z(\mathbf{Z})$ is usually constructed from a "kernel" matrix $\mathbf{K_Z}$ which undergoes a scalar Van der Maaten and Hinton [2008], row-stochastic Hinton and Roweis [2002] or doubly stochastic Lu et al. [2019], Van Assel et al. [2023] normalization. Gaussian kernel $[\mathbf{K_Z}]_{ij} = \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2)$, or heavy-tailed Student-t kernel $[\mathbf{K_Z}]_{ij} = (1 + \|\mathbf{z}_i - \mathbf{z}_j\|_2^2)^{-1}$, are typical choices Van der Maaten and Hinton [2008]. We also emphasize that one can retrieve the UMAP objective McInnes et al. [2018] from eq. (DR) using the binary cross-entropy loss. As described in Van Assel et al. [2022] from a probabilistic point of view, all these objectives can be derived from a common Markov random field model with various graph priors.

Remark 2.1. The usual formulations of neighbor embedding methods rely on the loss $L(x, y) = x \log(x/y)$. However, due to the normalization, the total mass $\sum_{ij}[\mathbf{C}_Z(\mathbf{Z})]_{ij}$ is constant (often equal to 1) in all of the cases mentioned above. Thus the minimization in $\mathbf{Z}$ with the L_{KL} formulation is equivalent.

Hyperbolic geometry. The presented DR methods can also be extended to incorporate non-Euclidean geometries. Hyperbolic spaces Chami et al. [2021], Fan et al. [2022], Guo et al. [2022], Lin et al. [2023] are of particular interest as they can capture hierarchical structures more effectively than Euclidean spaces and mitigate the curse of dimensionality by producing representations with lower distortion rates. For instance, Guo et al. [2022] adapted t-SNE by using the Poincaré distance and by changing the Student's t-distribution with a more general hyperbolic Cauchy distribution. Notions of

projection subspaces can also be adapted, *e.g.* Chami et al. [2021] use horospheres as one-dimensional subspaces.

Entropic Affinities and SNE/t-SNE. Entropic affinities (EAs) were first introduced in the seminal paper *Stochastic Neighbor Embedding* (SNE) Hinton and Roweis [2002]. It consists in normalizing each row i of a distance matrix by a bandwidth parameter ε_i such that the distribution associated with each row of the corresponding stochastic (*i.e.* row-normalized) Gaussian affinity has a fixed entropy. The value of this entropy, whose exponential is called the *perplexity*, is then the only hyperparameter left to tune and has an intuitive interpretation as the number of effective neighbors of each point Vladymyrov and Carreira-Perpinan [2013]. EAs are notoriously used to encode pairwise relations in a high-dimensional space for the DR algorithm t-SNE Van der Maaten and Hinton [2008], among other DR methods including Carreira-Perpinán [2010]. t-SNE is increasingly popular in many applied fields Kobak and Berens [2019], Melit Devassy et al. [2020] mostly due to its ability to represent clusters in the data Linderman and Steinerberger [2019], Cai and Ma [2022]. Nonetheless, one major flaw of EAs is that they are inherently directed and often require post-processing symmetrization.

Entropic Affinities (EAs). Another frequently used approach to generate affinities from $\mathbf{C} \in \mathcal{D}$ is to employ *entropic affinities* Hinton and Roweis [2002]. The main idea is to consider *adaptive* kernel bandwidths $(\varepsilon_i^\star)_{i\in[\![n]\!]}$ to capture finer structures in the data compared to constant bandwidths Van Dijk et al. [2018]. Indeed, EAs rescale distances to account for the varying density across regions of the dataset.

Given $\xi \in [\![n-1]\!]$, the goal of EAs is to build a Gaussian Markov chain transition matrix $\mathbf{P}^{\mathrm{e}}$ with prescribed entropy as

$$\forall i, \forall j,\ P^{\mathrm{e}}_{ij} = \frac{\exp\left(-C_{ij}/\varepsilon_i^\star\right)}{\sum_\ell \exp\left(-C_{i\ell}/\varepsilon_i^\star\right)} \quad \text{with } \varepsilon_i^\star \in \mathbb{R}_+^* \text{ s.t. } \mathrm{H}(\mathbf{P}^{\mathrm{e}}_{i:}) = \log \xi + 1\,. \tag{EA}$$

The hyperparameter ξ, which is also known as *perplexity*, can be interpreted as the effective number of neighbors for each data point Vladymyrov and Carreira-Perpinan [2013]. Indeed, a perplexity of ξ means that each row of $\mathbf{P}^{\mathrm{e}}$ (which is a discrete probability since $\mathbf{P}^{\mathrm{e}}$ is row-wise stochastic) has the same entropy as a uniform distribution over ξ neighbors. Therefore, it provides the practitioner with an interpretable parameter specifying which scale of dependencies the affinity matrix should faithfully capture. In practice, a root-finding algorithm is used to find the bandwidth parameters $(\varepsilon_i^\star)_{i\in[\![n]\!]}$ that satisfy the constraints Vladymyrov and Carreira-Perpinan [2013]. Hereafter, with a slight abuse of language, we call $e^{\mathrm{H}(\mathbf{P}_{i:})-1}$ the perplexity of the point i.

Dimension Reduction with SNE/t-SNE. One of the main applications of EAs is the DR algorithm `SNE` Hinton and Roweis [2002]. We denote by $\mathbf{C}_\mathbf{X} = \left(\|\mathbf{X}_{i:} - \mathbf{X}_{j:}\|_2^2\right)_{ij}$ and $\mathbf{C}_\mathbf{Z} = \left(\|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2\right)_{ij}$ the cost matrices derived from the rows (*i.e.* the samples) of $\mathbf{X}$ and $\mathbf{Z}$ respectively. `SNE` focuses on minimizing in the latent coordinates $\mathbf{Z} \in \mathbb{R}^{n\times q}$ the objective $\mathrm{KL}(\mathbf{P}^\mathrm{e}|\mathbf{Q}_\mathbf{Z})$ where $\mathbf{P}^\mathrm{e}$ solves (EA) with cost $\mathbf{C}_\mathbf{X}$ and $[\mathbf{Q}_\mathbf{Z}]_{ij} = \exp(-[\mathbf{C}_\mathbf{Z}]_{ij})/(\sum_\ell \exp(-[\mathbf{C}_\mathbf{Z}]_{i\ell}))$. In the seminal paper [Van der Maaten and Hinton, 2008], a newer proposal for a *symmetric* version was presented, which has since replaced `SNE` in practical applications. Given a symmetric normalization for the similarities in latent space $[\widetilde{\mathbf{Q}}_\mathbf{Z}]_{ij} = \exp(-[\mathbf{C}_\mathbf{Z}]_{ij})/\sum_{\ell,t}\exp(-[\mathbf{C}_\mathbf{Z}]_{\ell t})$ it consists in solving

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} \mathrm{KL}(\overline{\mathbf{P}^\mathrm{e}}|\widetilde{\mathbf{Q}}_\mathbf{Z}) \quad \text{where} \quad \overline{\mathbf{P}^\mathrm{e}} = \frac{1}{2}(\mathbf{P}^\mathrm{e} + \mathbf{P}^{\mathrm{e}\top})\,. \tag{Symmetric-SNE}$$

In other words, the affinity matrix $\overline{\mathbf{P}^\mathrm{e}}$ is the Euclidean projection of $\mathbf{P}^\mathrm{e}$ on the space of symmetric matrices $\mathcal{S}$.

Proposition 2.1. $\mathrm{Proj}_\mathcal{S}^{\ell_2}(\mathbf{P}) = \arg\min_{\overline{\mathbf{P}}\in\mathcal{S}} \left\|\overline{\mathbf{P}} - \mathbf{P}\right\|_2$.

Proof. For the above problem, the Lagrangian takes the form, with $\mathbf{W} \in \mathbb{R}^{n\times n}$,

$$\mathcal{L}(\mathbf{P}, \mathbf{W}) = \|\mathbf{P} - \mathbf{K}\|_2^2 + \langle \mathbf{W}, \mathbf{P} - \mathbf{P}^\top\rangle\,. \tag{2.1}$$

Cancelling the gradient of $\mathcal{L}$ with respect to $\mathbf{P}$ gives $2(\mathbf{P}^\star - \mathbf{K}) + \mathbf{W} - \mathbf{W}^\top = \mathbf{0}$. Thus $\mathbf{P}^\star = \mathbf{K} + \frac{1}{2}\left(\mathbf{W}^\top - \mathbf{W}\right)$. Using the symmetry constraint on $\mathbf{P}^\star$ yields $\mathbf{P}^\star = \frac{1}{2}\left(\mathbf{K} + \mathbf{K}^\top\right)$. Hence we have:

$$\arg\min_{\mathbf{P}\in\mathcal{S}} \|\mathbf{P} - \mathbf{K}\|_2^2 = \frac{1}{2}\left(\mathbf{K} + \mathbf{K}^\top\right)\,. \tag{2.2}$$

□

Instead of the Gaussian kernel, the popular extension t-SNE [Van der Maaten and Hinton, 2008] considers a different distribution in the latent space $[\widetilde{\mathbf{Q}}_\mathbf{Z}]_{ij} = (1 + [\mathbf{C}_\mathbf{Z}]_{ij})^{-1}/\sum_{\ell,t}(1 + [\mathbf{C}_\mathbf{Z}]_{\ell t})^{-1}$. In this formulation, $\widetilde{\mathbf{Q}}_\mathbf{Z}$ is a joint Student t-distribution that accounts for crowding effects: a relatively small distance in a high-dimensional space can be accurately represented by a significantly greater distance in the low-dimensional space.

2.3 Optimal Transport

Optimal Transport (OT) [Villani et al., 2009, Peyré et al., 2019] is a popular framework for comparing probability distributions and is at the core of Chapter 3 and Chapter 5 of this thesis.

2.3.1 Comparing Distributions with Optimal Transport

Classical OT methods require defining a meaningful transportation cost between the supports of the two distributions. This is however difficult in the context of dimensionality reduction where the two spaces $\mathbb{R}^p$ and $\mathbb{R}^d$ have different dimensions.

Monge formulation. We consider two Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ such that we can define a cost function $c_{\mathcal{XY}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be two probability measures that we aim to compare. The original formulation Monge [1781] of OT seeks the map T satisfying $T_{\#}\mu = \nu$ that minimizes the transportation cost given by

$$M(\mu, \nu) \coloneqq \inf_{T_{\#}\mu=\nu} \int_{\mathcal{X}} c_{\mathcal{XY}}(\boldsymbol{x}, T(\boldsymbol{x}))\mathrm{d}\mu(\boldsymbol{x}) \,. \tag{2.3}$$

The above formulation is highly non-linear in T and the set of admissible maps T is not convex hindering an easy analysis. Moreover, the existence of an optimal map T is not guaranteed in general.

Kantorovich relaxation. To resolve this, a popular relaxation by Kantorovich [1942] consists of optimizing instead over the space of probabilistic couplings with marginals μ and ν

$$W(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c_{\mathcal{XY}}(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{2.4}$$

This formulation is a convex optimization problem and the infimum is well defined under mild assumptions Santambrogio [2015]. If the optimal coupling $\pi^\star$ is supported on a *deterministic* function, *i.e.* $\pi^\star$ is of the form $(\mathrm{id} \times T^\star)\#\mu$, then $T^\star$ solves equation 2.3. This holds under the assumption that one of the inputs is absolutely continuous with respect to the Lebesgue measure for $c_{\mathcal{XY}}(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x} - \boldsymbol{y})$ with h strictly convex Gangbo and McCann [1996]. In the case of discrete measures, the equivalence holds if $\mu = \frac{1}{N}\sum_{i\in[\![N]\!]} \delta_{\boldsymbol{x}_i}$ and $\nu = \frac{1}{N}\sum_{i\in[\![N]\!]} \delta_{\boldsymbol{y}_i}$ as the solution of equation 2.4 is reached at an extremal point of the polytope of doubly stochastic matrices Bertsimas and Tsitsiklis [1997].

2.3.2 Affinity via Symmetric OT

Doubly Stochastic Affinities. Doubly stochastic (DS) affinities are non-negative matrices whose rows and columns have unit ℓ_1 norm. In many applications, it has been demonstrated that DS affinity normalization (*i.e.* determining the nearest DS matrix to a given affinity matrix) offers numerous benefits. First, it can be seen as a relaxation of k-means Zass and Shashua [2005] and it is well-established that it enhances spectral clustering performances Ding et al. [2022], Zass and Shashua [2006], Beauchemin [2015].

Additionally, DS matrices present the benefit of being invariant to the various Laplacian normalizations Von Luxburg [2007]. Recent observations indicate that the DS projection of the Gaussian kernel under the KL geometry is more resilient to heteroscedastic noise compared to its stochastic counterpart Landa et al. [2021]. It also offers a more natural analog to the heat kernel Marshall and Coifman [2019]. These properties have led to a growing interest in DS affinities, with their use expanding to various applications such as smoothing filters Milanfar [2013], subspace clustering Lim et al. [2020] and transformers Sander et al. [2022].

2.3.3 From Symmetric Entropy-Constrained OT to Sinkhorn Iterations

Symmetric Entropy-Constrained Optimal Transport. Entropy-regularized OT [Peyré et al., 2019] and its connection to affinity matrices are crucial components in our solution. In the special case of uniform marginals, and for $\nu > 0$, entropic OT computes the minimum of $\mathbf{P} \mapsto \langle \mathbf{P}, \mathbf{C} \rangle - \nu \sum_i \mathrm{H}(\mathbf{P}_{i:})$ over the space of doubly stochastic matrices $\{\mathbf{P} \in \mathbb{R}_+^{n \times n} : \mathbf{P}\mathbf{1} = \mathbf{P}^\top \mathbf{1} = \mathbf{1}\}$. The optimal solution is the *unique* doubly stochastic matrix $\mathbf{P}^{\mathrm{ds}}$ of the form $\mathbf{P}^{\mathrm{ds}} = \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$ where $\mathbf{K} = \exp(-\mathbf{C}/\nu)$ is the Gibbs energy derived from $\mathbf{C}$ and $\mathbf{u}, \mathbf{v}$ are positive vectors that can be found with the celebrated Sinkhorn-Knopp's algorithm [Cuturi, 2013, Sinkhorn, 1964]. Interestingly, when the cost $\mathbf{C}$ is *symmetric* (*e.g.* $\mathbf{C} \in \mathcal{D}$) we can take $\mathbf{u} = \mathbf{v}$ [Idel, 2016, Section 5.2] so that the unique optimal solution is itself symmetric and writes

$$\mathbf{P}^{\mathrm{ds}} = \exp\left((\mathbf{f} \oplus \mathbf{f} - \mathbf{C})/\nu\right) \text{ where } \mathbf{f} \in \mathbb{R}^n\,. \tag{DS}$$

In this case, by relying on convex duality as detailed in Appendix 2.3.3, an equivalent formulation for the symmetric entropic OT problem is

$$\min_{\mathbf{P} \in \mathbb{R}_+^{n \times n}} \quad \langle \mathbf{P}, \mathbf{C} \rangle \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \mathbf{1},\ \mathbf{P} = \mathbf{P}^\top \text{ and } \sum_i \mathrm{H}(\mathbf{P}_{i:}) \geq \eta\,, \tag{EOT}$$

where $0 \leq \eta \leq n(\log n + 1)$ is a constraint on the global entropy $\sum_i \mathrm{H}(\mathbf{P}_{i:})$ of the OT plan $\mathbf{P}$ which happens to be saturated at optimum (Appendix 2.3.3). This constrained formulation of symmetric entropic OT will provide new insights into entropic affinities, as detailed in the next sections.

In this section, we derive Sinkhorn iterations from the problem (EOT). Let $\mathbf{C} \in \mathcal{D}$.

Proof. We start by making the constraints explicit.

$$\min_{\mathbf{P} \in \mathbb{R}_+^{n \times n}} \quad \langle \mathbf{P}, \mathbf{C} \rangle \tag{2.5}$$

$$\text{s.t.} \quad \sum_{i \in [\![n]\!]} \mathrm{H}(\mathbf{P}_{i:}) \geq \eta \tag{2.6}$$

$$\mathbf{P}\mathbf{1} = \mathbf{1}, \quad \mathbf{P} = \mathbf{P}^\top\,. \tag{2.7}$$

For the above convex problem the Lagrangian writes, where $\nu \in \mathbb{R}_+$, $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{\Gamma} \in \mathbb{R}^{n\times n}$:

$$\mathcal{L}(\mathbf{P}, \mathbf{f}, \nu, \mathbf{\Gamma}) = \langle \mathbf{P}, \mathbf{C} \rangle + \left\langle \nu, \eta - \sum_{i \in [\![n]\!]} \mathrm{H}(\mathbf{P}_i) \right\rangle + 2\langle \mathbf{f}, \mathbf{1} - \mathbf{P}\mathbf{1} \rangle + \langle \mathbf{\Gamma}, \mathbf{P} - \mathbf{P}^\top \rangle \,. \tag{2.8}$$

Strong duality holds and the first order KKT condition gives for the optimal primal $\mathbf{P}^\star$ and dual $(\nu^\star, \mathbf{f}^\star, \mathbf{\Gamma}^\star)$ variables:

$$\nabla_{\mathbf{P}} \mathcal{L}(\mathbf{P}^\star, \mathbf{f}^\star, \nu^\star, \mathbf{\Gamma}^\star) = \mathbf{C} + \nu^\star \log \mathbf{P}^\star - 2\mathbf{f}^\star \mathbf{1}^\top + \mathbf{\Gamma}^\star - \mathbf{\Gamma}^{\star\top} = \mathbf{0} \,. \tag{2.9}$$

Since $\mathbf{P}^\star, \mathbf{C} \in \mathcal{S}$ we have $\mathbf{\Gamma}^\star - \mathbf{\Gamma}^{\star\top} = \mathbf{f}^\star \mathbf{1}^\top - \mathbf{1}\mathbf{f}^{\star\top}$. Hence $\mathbf{C} + \nu^\star \log \mathbf{P}^\star - \mathbf{f}^\star \oplus \mathbf{f}^\star = \mathbf{0}$. Suppose that $\nu^\star = 0$ then the previous reasoning implies that $\forall (i,j), C_{ij} = f_i^\star + f_j^\star$. Using that $\mathbf{C} \in \mathcal{D}$ we have $C_{ii} = C_{jj} = 0$ thus $\forall i, f_i^\star = 0$ and thus this would imply that $\mathbf{C} = 0$ which is not allowed by hypothesis. Therefore $\nu^\star \neq 0$ and the entropy constraint is saturated at the optimum by complementary slackness. Isolating $\mathbf{P}^\star$ then yields:

$$\mathbf{P}^\star = \exp\left((\mathbf{f}^\star \oplus \mathbf{f}^\star - \mathbf{C})/\nu^\star\right) . \tag{2.10}$$

$\mathbf{P}^\star$ must be primal feasible in particular $\mathbf{P}^\star \mathbf{1} = \mathbf{1}$. This constraint gives us the Sinkhorn fixed point relation for $\mathbf{f}^\star$:

$$\forall i \in [\![n]\!], \quad [\mathbf{f}^\star]_i = -\nu^\star \,\mathrm{LSE}\left((\mathbf{f}^\star - \mathbf{C}_{:i})/\nu^\star\right) , \tag{2.11}$$

where for a vector $\boldsymbol{\alpha}$, we use the notation $\mathrm{LSE}(\boldsymbol{\alpha}) = \log \sum_k \exp(\alpha_k)$. □

2.3.4 Optimal Transport Across Spaces : Gromov-Wasserstein Formulation

We review in this section the Gromov-Wasserstein formulation of OT aiming at comparing distributions "across spaces".

It is usually impossible to define a meaningful transportation cost from two spaces $\mathcal{X}$ and $\mathcal{Z}$ that are not part of a common ground metric space This occurs when considering ambient Euclidean spaces of different dimensions *i.e.* $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Z} \subseteq \mathbb{R}^d$ with $p \neq d$ which is precisely the context of dimensionality reduction. We first introduce the general GW problem before highlighting its utility to compare distributions in incomparable spaces.

Gromov-Wasserstein (GW). The GW framework Mémoli [2011], Sturm [2012] comprises a collection of OT methods designed to compare distributions by examining the pairwise relations *within each domain*. For two matrices $\mathbf{C} \in \mathbb{R}^{N\times N}$, $\overline{\mathbf{C}} \in \mathbb{R}^{n\times n}$, and weights $\mathbf{h} \in \Sigma_N, \overline{\mathbf{h}} \in \Sigma_n$, the GW discrepancy is defined as

$$\begin{aligned} \mathrm{GW}_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{h}, \overline{\mathbf{h}}) &\coloneqq \min_{\mathbf{T} \in \mathcal{U}(\mathbf{h}, \overline{\mathbf{h}})} E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) \,, \\ \text{where} \quad E_L(\mathbf{C}, \overline{\mathbf{C}}, \mathbf{T}) &\coloneqq \sum_{ijkl} L(C_{ij}, \overline{C}_{kl}) T_{ik} T_{jl} \,, \end{aligned} \tag{GW}$$

and $\mathcal{U}(\mathbf{h}, \overline{\mathbf{h}}) = \left\{\mathbf{T} \in \mathbb{R}_+^{N\times n} : \mathbf{T}\mathbf{1}_n = \mathbf{h}, \mathbf{T}^\top \mathbf{1}_N = \overline{\mathbf{h}}\right\}$ is the set of couplings between $\mathbf{h}$ and $\overline{\mathbf{h}}$. In this formulation, both pairs $(\mathbf{C}, \mathbf{h})$ and $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ can be interpreted as graphs with corresponding connectivity matrices $\mathbf{C}, \overline{\mathbf{C}}$, and where nodes are weighted by histograms $\mathbf{h}, \overline{\mathbf{h}}$. Equation (GW) is thus a *quadratic problem* (in $\mathbf{T}$) which consists in finding a soft-assignment matrix $\mathbf{T}$ that aligns the nodes of the two graphs in a way that preserves their pairwise connectivities.

From a distributional perspective, GW can also be viewed as a distance between distributions that do not belong to the same metric space. For two discrete probability distributions $\mu_X = \sum_{i=1}^N [\mathbf{h}_X]_i \delta_{\mathbf{x}_i} \in \mathcal{P}_N(\mathbb{R}^p), \mu_Z = \sum_{i=1}^n [\mathbf{h}_Z]_i \delta_{\mathbf{z}_i} \in \mathcal{P}_n(\mathbb{R}^d)$ and pairwise similarity matrices $\mathbf{C}_X(\mathbf{X})$ and $\mathbf{C}_Z(\mathbf{Z})$ associated with the supports $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ and $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^\top$, the quantity $\mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \mathbf{h}_X, \mathbf{h}_Z)$ is a measure of dissimilarity or discrepancy between μ_X, μ_Z. Specifically, when $L = L_2$, and $\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z})$ are pairwise distance matrices, GW defines a proper distance between μ_X and μ_Z with respect to measure preserving isometries[1] Mémoli [2011].

Due to its versatile properties, notably in comparing distributions over different domains, the GW problem has found many applications in machine learning, *e.g.,* for 3D meshes alignment Solomon et al. [2016], Ezuz et al. [2017], NLP Alvarez-Melis and Jaakkola [2018], (co-)clustering Peyré et al. [2016], Redko et al. [2020], single-cell analysis Demetci et al. [2020b], neuroimaging Thual et al. [2022], graph representation learning Xu [2020], Vincent-Cuaz et al. [2021], Liu et al. [2022b], Vincent-Cuaz et al. [2022b], Zeng et al. [2023] and partitioning Xu et al. [2019], Chowdhury and Needham [2021].

In this work, we leverage the GW discrepancy to extend classical DR approaches, framing them as the projection of a distribution onto a space of lower dimensionality.

2.4 Background DistR

2.5 Background on Dimensionality Reduction and Optimal Transport

We start by reviewing the most popular DR approaches and we introduce the Gromov-Wasserstein problem.

2.5.1 Unified View of Dimensionality Reduction

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^\top \in \mathbb{R}^{N\times p}$ be an input dataset. Dimensionality reduction focuses on constructing a low-dimensional representation or *embedding* $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)^\top \in \mathbb{R}^{N\times d}$, where $d < p$. The latter should preserve a prescribed geometry for the dataset encoded via a symmetric pairwise similarity matrix $\mathbf{C}_\mathbf{X} \in \mathbb{R}_+^{N\times N}$ obtained from $\mathbf{X}$. To this end,

[1] With weaker assumptions on $\mathbf{C}_X, \mathbf{C}_Z$, GW defines a pseudo-metric *w.r.t.* a different notion of isomorphism Chowdhury and Mémoli [2019]. See Appendix **??** for more details.

most popular DR methods optimize $\mathbf{Z}$ such that a certain pairwise similarity matrix in the output space matches $\mathbf{C}_\mathbf{X}$ according to some criteria. We subsequently introduce the functions

$$\mathbf{C}_X : \mathbb{R}^{N\times p} \to \mathbb{R}^{N\times N}, \mathbf{C}_Z : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times N}, \tag{2.12}$$

which define pairwise similarity matrices in the input and output space from embeddings $\mathbf{X}$ and $\mathbf{Z}$. The DR problem can be formulated quite generally as the optimization problem

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \sum_{(i,j)\in[\![N]\!]^2} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{ij}) \,. \tag{DR}$$

where $L : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a loss that quantifies how similar are two points in the input space $\mathbb{R}^p$ compared to two points in the output space $\mathbb{R}^d$. Various losses are used, such as the quadratic loss $L_2(x, y) \coloneqq (x - y)^2$ or the Kullback-Leibler divergence $L_{\mathrm{KL}}(x, y) \coloneqq x \log(x/y) - x + y$. Below, we recall several popular methods that can be placed within this framework.

Spectral methods. When $\mathbf{C}_X(\mathbf{X})$ is a positive semi-definite matrix, eq. (DR) recovers spectral methods by choosing the quadratic loss $L = L_2$ and $\mathbf{C}_Z(\mathbf{Z}) = (\langle \mathbf{z}_i, \mathbf{z}_j\rangle)_{(i,j)\in[\![N]\!]^2}$ the matrix of inner products in the embedding space. Indeed, in this case, the objective value of eq. (DR) reduces to

$$\sum_{(i,j)\in[\![N]\!]^2} L_2([\mathbf{C}_X(\mathbf{X})]_{ij}, \langle \mathbf{z}_i, \mathbf{z}_j\rangle) = \|\mathbf{C}_X(\mathbf{X}) - \mathbf{Z}\mathbf{Z}^\top\|_F^2$$

where $\|\cdot\|_F$ is the Frobenius norm. This problem is commonly known as kernel Principal Component Analysis (PCA) Schölkopf et al. [1997] and an optimal solution is given by $\mathbf{Z}^\star = (\sqrt{\lambda_1}\mathbf{v}_1, ..., \sqrt{\lambda_d}\mathbf{v}_d)^\top$ where λ_i is the i-th largest eigenvalue of $\mathbf{C}_X(\mathbf{X})$ with corresponding eigenvector $\mathbf{v}_i$ Eckart and Young [1936]. As shown by Ham et al. [2004], Ghojogh et al. [2021], numerous dimension reduction methods can be categorized in this manner. This includes PCA when $\mathbf{C}_X(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ is the matrix of inner products in the input space; (classical) multidimensional scaling Borg and Groenen [2005], when $\mathbf{C}_X(\mathbf{X}) = -\frac{1}{2}\mathbf{H}\mathbf{D}_\mathbf{X}\mathbf{H}$ with $\mathbf{D}_\mathbf{X}$ the matrix of squared euclidean distance between the points in $\mathbb{R}^p$ and $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ is the centering matrix; Isomap Tenenbaum et al. [2000], with $\mathbf{C}_X(\mathbf{X}) = -\frac{1}{2}\mathbf{H}\mathbf{D}_\mathbf{X}^{(g)}\mathbf{H}$ with $\mathbf{D}_\mathbf{X}^{(g)}$ the geodesic distance matrix; Laplacian Eigenmap Belkin and Niyogi [2003], with $\mathbf{C}_X(\mathbf{X}) = \mathbf{L}_\mathbf{X}^\dagger$ the pseudo-inverse of the Laplacian associated to some adjacency matrix $\mathbf{W}_\mathbf{X}$; but also Locally Linear Embedding Roweis and Saul [2000], and Diffusion Map Coifman and Lafon [2006] (for all of these examples we refer to Ghojogh et al. 2021, Table 1).

3

SNEkhorn: Dimension Reduction with Symmetric Entropic Affinities

Contents

3.1 Introduction

Considering symmetric similarities is appealing since the proximity between two points is inherently symmetric. Nonetheless, the Euclidean projection in equation Symmetric-SNE *does not preserve the construction of entropic affinities*. In particular, $\overline{\mathbf{P}^{\mathrm{e}}}$ is not stochastic in general and $\mathrm{H}(\overline{\mathbf{P}^{\mathrm{e}}_{i:}}) \neq (\log \xi + 1)$ thus the entropy associated with each point is no longer controlled after symmetrization (see the bottom left plot of fig. 3.1). This is arguably one of the main drawbacks of the approach. By contrast, the $\mathbf{P}^{\mathrm{se}}$ affinity that will be introduced in section 3.2 can accurately set the entropy in each point to the desired value $\log \xi + 1$. As shown in fig. 3.1 this leads to more faithful embeddings with higher silhouette scores when combined with the SNEkhorn algorithm (section 3.3).

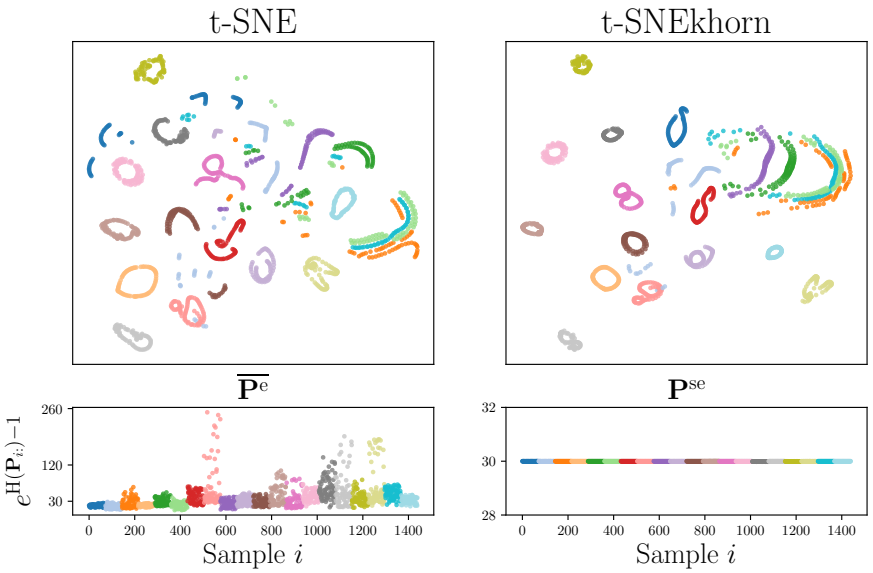

Figure 3.1: Top: COIL Nene et al. [1996] embeddings with silhouette scores produced by Symmetric-SNE and SNEkhorn (our method introduced in section 3.3) for $\xi = 30$. Bottom: $e^{\mathrm{H}(\mathbf{P}_{i:})-1}$ (*perplexity*) for each point i.

Contributions. In this work, we study the missing link between EAs, which are easy to tune and adaptable to data with heterogeneous density, and DS affinities which have interesting properties in practical applications as aforementioned. Our main contributions are as follows. We uncover the convex optimization problem that underpins classical entropic affinities, exhibiting novel links with entropy-regularized Optimal Transport (OT) (section 3.2.1). We then propose in section 3.2.2 a principled symmetrization of entropic affinities. The latter enables controlling the entropy in each point, unlike t-SNE's post-processing symmetrization, and producing a genuinely doubly stochastic affinity. We show how to compute this new affinity efficiently using a dual ascent algorithm. In section 3.3, we introduce SNEkhorn: a DR algorithm that couples this new symmetric entropic affinity with a doubly stochastic kernel in the low-dimensional embedding space, without sphere concentration issue Lu et al. [2019]. We finally showcase the benefits of symmetric entropic affinities on a variety of applications in Section 3.4 including spectral clustering and DR experiments on datasets ranging from images to genomics data.

3.2 Symmetric Entropic Affinities

In this section, we present our first major contribution: symmetric entropic affinities. We begin by providing a new perspective on EAs through the introduction of an equivalent

convex problem.

3.2.1 Entropic Affinities as Entropic Optimal Transport

We introduce the following set of matrices with row-wise stochasticity and entropy constraints:

$$\mathcal{H}_\xi = \{\mathbf{P} \in \mathbb{R}_+^{n\times n} \text{ s.t. } \mathbf{P}\mathbf{1} = \mathbf{1} \text{ and } \forall i,\ \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1\}\,. \tag{3.1}$$

This space is convex since $\mathbf{p} \in \mathbb{R}_+^n \mapsto \mathrm{H}(\mathbf{p})$ is concave, thus its superlevel set is convex. In contrast to the entropic constraints utilized in standard entropic optimal transport which set a lower-bound on the *global* entropy, as demonstrated in the formulation equation EOT, $\mathcal{H}_\xi$ imposes a constraint on the entropy of *each row* of the matrix $\mathbf{P}$. Our first contribution is to prove that EAs can be computed by solving a specific problem involving $\mathcal{H}_\xi$ (see Appendix **??** for the proof).

Proposition 3.1. Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ without constant rows. Then $\mathbf{P}^{\mathrm{e}}$ solves the entropic affinity problem (EA) with cost $\mathbf{C}$ if and only if $\mathbf{P}^{\mathrm{e}}$ is the unique solution of the convex problem

$$\min_{\mathbf{P}\in\mathcal{H}_\xi} \langle \mathbf{P}, \mathbf{C}\rangle. \tag{EA as OT}$$

Interestingly, this result shows that EAs boil down to minimizing a transport objective with cost $\mathbf{C}$ and row-wise entropy constraints $\mathcal{H}_\xi$ where ξ is the desired perplexity. As such, equation EA as OT can be seen as a specific *semi-relaxed* OT problem [Rabin et al., 2014, Flamary et al., 2016] (*i.e.* without the second constraint on the marginal $\mathbf{P}^\top\mathbf{1} = \mathbf{1}$) but with entropic constraints on the rows of $\mathbf{P}$. We also show that the optimal solution $\mathbf{P}^\star$ of equation EA as OT has *saturated entropy i.e.* $\forall i,\ \mathrm{H}(\mathbf{P}^\star_{i:}) = \log\xi + 1$. In other words, relaxing the equality constraint in equation EA as a inequality constraint in $\mathbf{P} \in \mathcal{H}_\xi$ does not affect the solution while it allows reformulating entropic affinity as a convex optimization problem. To the best of our knowledge, this connection between OT and entropic affinities is novel and is an essential key to the method proposed in the next section.

Remark 3.1. The kernel bandwidth parameter $\boldsymbol{\varepsilon}$ from the original formulation of entropic affinities (EA) is the Lagrange dual variable associated with the entropy constraint in (EA as OT). Hence computing $\boldsymbol{\varepsilon}^\star$ in (EA) exactly corresponds to solving the dual problem of (EA as OT).

Remark 3.2. Let $\mathbf{K}_\sigma = \exp(-\mathbf{C}/\sigma)$. As shown in **??**, if $\boldsymbol{\varepsilon}^\star$ solves (EA) and $\sigma \leq \min(\boldsymbol{\varepsilon}^\star)$, then $\mathbf{P}^{\mathrm{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K}_\sigma) = \arg\min_{\mathbf{P}\in\mathcal{H}_\xi} \mathrm{KL}(\mathbf{P}|\mathbf{K}_\sigma)$. Therefore $\mathbf{P}^{\mathrm{e}}$ can be seen as a KL Bregman projection [Benamou et al., 2015] of a Gaussian kernel onto $\mathcal{H}_\xi$. Hence the input matrix in equation Symmetric-SNE is $\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(\mathbf{K}_\sigma))$ which corresponds to a surprising mixture of KL and orthogonal projections.

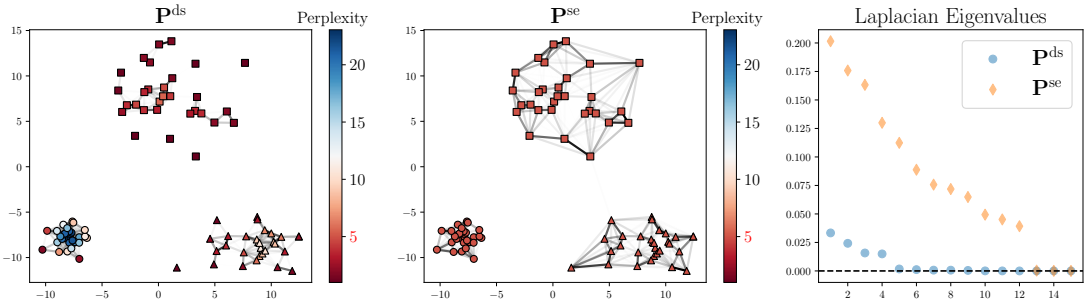

Figure 3.2: Samples from a mixture of three Gaussians with varying standard deviations. The edges' strength is proportional to the weights in the affinities $\mathbf{P}^{\mathrm{ds}}$ equation DS and $\mathbf{P}^{\mathrm{se}}$ equation SEA computed with $\xi = 5$ (for $\mathbf{P}^{\mathrm{ds}}$, ξ is the average perplexity such that $\sum_i \mathrm{H}(\mathbf{P}^{\mathrm{ds}}_{i:}) = \sum_i \mathrm{H}(\mathbf{P}^{\mathrm{se}}_{i:})$). Points' color represents the perplexity $e^{\mathrm{H}(\mathbf{P}_{i:})-1}$. Right plot: smallest eigenvalues of the Laplacian for the two affinities.

3.2.2 Symmetric Entropic Affinity Formulation

Based on the previous formulation we now propose symmetric entropic affinities: a symmetric version of EAs that enables keeping the entropy associated with each row (or equivalently column) to the desired value of $\log \xi + 1$ while producing a symmetric doubly stochastic affinity matrix. Our strategy is to enforce symmetry through an additional constraint in (EA as OT), in a similar fashion as equation EOT. More precisely we consider the convex optimization problem

$$\min_{\mathbf{P} \in \mathcal{H}_\xi \cap \mathcal{S}} \langle \mathbf{P}, \mathbf{C} \rangle \,. \tag{SEA}$$

where we recall that $\mathcal{S}$ is the set of $n \times n$ symmetric matrices. Note that for any $\xi \leq n-1$, $\frac{1}{n}\mathbf{1}\mathbf{1}^\top \in \mathcal{H}_\xi \cap \mathcal{S}$ hence the set $\mathcal{H}_\xi \cap \mathcal{S}$ is a non-empty and convex set. We first detail some important properties of problem equation SEA (the proofs of the following results can be found in Appendix **??**).

Proposition 3.2 (Saturation of the entropies)**.** Let $\mathbf{C} \in \mathcal{S}$ with zero diagonal, then equation SEA with cost $\mathbf{C}$ has a *unique solution* that we denote by $\mathbf{P}^{\mathrm{se}}$. If moreover $\mathbf{C} \in \mathcal{D}$, then for at least $n-1$ indices $i \in [\![n]\!]$ the solution satisfies $\mathrm{H}(\mathbf{P}^{\mathrm{se}}_{i:}) = \log \xi + 1$.

In other words, the unique solution $\mathbf{P}^{\mathrm{se}}$ has at least $n-1$ saturated entropies *i.e.* the corresponding $n-1$ points have exactly a perplexity of ξ. In practice, with the algorithmic solution detailed below, we have observed that all n entropies are saturated. Therefore, we believe that this proposition can be extended with a few more assumptions on $\mathbf{C}$. Accordingly, problem equation SEA allows accurate control over the point-wise entropies while providing a symmetric doubly stochastic matrix, unlike $\overline{\mathbf{P}^{\mathrm{e}}}$ defined in equation Symmetric-SNE, as summarized in **??**. In the sequel, we denote by

Table 3.1: Properties of $\mathbf{P}^{\mathrm{e}}$, $\overline{\mathbf{P}^{\mathrm{e}}}$, $\mathbf{P}^{\mathrm{ds}}$ and $\mathbf{P}^{\mathrm{se}}$.

Affinity matrix	$\mathbf{P}^{\mathrm{e}}$	$\overline{\mathbf{P}^{\mathrm{e}}}$	$\mathbf{P}^{\mathrm{ds}}$	$\mathbf{P}^{\mathrm{se}}$
$\mathbf{P} = \mathbf{P}^\top$	×	✓	✓	✓
$\mathbf{P}\mathbf{1} = \mathbf{P}^\top\mathbf{1} = \mathbf{1}$	×	×	✓	✓
$\mathrm{H_r}(\mathbf{P}) = (\log\xi + 1)\mathbf{1}$	✓	×	×	✓

$\mathrm{H_r}(\mathbf{P}) = (\mathrm{H}(\mathbf{P}_{i:}))_i$ the vector of row-wise entropies of $\mathbf{P}$. We rely on the following result to compute $\mathbf{P}^{\mathrm{se}}$.

Proposition 3.3 (Solving for SEA)**.** Let $\mathbf{C} \in \mathcal{D}, \mathcal{L}(\mathbf{P}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \langle \mathbf{P}, \mathbf{C}\rangle + \langle \boldsymbol{\gamma}, (\log\xi + 1)\mathbf{1} - \mathrm{H_r}(\mathbf{P})\rangle + \langle \boldsymbol{\lambda}, \mathbf{1} - \mathbf{P}\mathbf{1}\rangle$ and $q(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}\cap\mathcal{S}} \mathcal{L}(\mathbf{P}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$. Strong duality holds for equation SEA. Moreover, let $\boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star \in \operatorname{argmax}_{\gamma\geq 0, \boldsymbol{\lambda}} q(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ be the optimal dual variables respectively associated with the entropy and marginal constraints. Then, for at least $n-1$ indices $i \in [\![n]\!], \gamma_i^\star > 0$. When $\forall i \in [\![n]\!]$, $\gamma_i^\star > 0$ then $\mathrm{H_r}(\mathbf{P}^{\mathrm{se}}) = (\log\xi + 1)\mathbf{1}$ and $\mathbf{P}^{\mathrm{se}}$ has the form

$$\mathbf{P}^{\mathrm{se}} = \exp\left((\boldsymbol{\lambda}^\star \oplus \boldsymbol{\lambda}^\star - 2\mathbf{C}) \oslash (\boldsymbol{\gamma}^\star \oplus \boldsymbol{\gamma}^\star)\right). \tag{3.2}$$

By defining the symmetric matrix $\mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \exp\left((\boldsymbol{\lambda} \oplus \boldsymbol{\lambda} - 2\mathbf{C}) \oslash (\boldsymbol{\gamma} \oplus \boldsymbol{\gamma})\right)$, we prove that, when $\boldsymbol{\gamma} > 0, \min_{\mathbf{P}\in\mathcal{S}} \mathcal{L}(\mathbf{P}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ has a unique solution given by $\mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ which implies $q(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \mathcal{L}(\mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda}), \boldsymbol{\gamma}, \boldsymbol{\lambda})$. Thus the proposition shows that when $\boldsymbol{\gamma}^\star > 0$, $\mathbf{P}^{\mathrm{se}} = \mathbf{P}(\boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star)$ where $\boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star$ solve the *concave* dual problem

$$\max_{\boldsymbol{\gamma}>0, \boldsymbol{\lambda}} \mathcal{L}(\mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda}), \boldsymbol{\gamma}, \boldsymbol{\lambda}). \tag{Dual-SEA}$$

Consequently, to find $\mathbf{P}^{\mathrm{se}}$ we solve the problem equation Dual-SEA. Although the form of $\mathbf{P}^{\mathrm{se}}$ presented in Proposition 3.3 is only valid when $\boldsymbol{\gamma}^\star$ is positive and we have only proved it for $n-1$ indices, we emphasize that if equation Dual-SEA has a finite solution, then it is equal to $\mathbf{P}^{\mathrm{se}}$. Indeed in this case the solution satisfies the KKT system associated with equation SEA.

Numerical optimization. The dual problem equation Dual-SEA is concave and can be solved with guarantees through a dual ascent approach with closed-form gradients (using *e.g.* `SGD`, `BFGS` Liu and Nocedal [1989] or `ADAM` Kingma and Ba [2014]). At each gradient step, one can compute the current estimate $\mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ while the gradients of the loss *w.r.t.* $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are given respectively by the constraints $(\log\xi + 1)\mathbf{1} - \mathrm{H_r}(\mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda}))$ and $\mathbf{1} - \mathbf{P}(\boldsymbol{\gamma}, \boldsymbol{\lambda})\mathbf{1}$ (see *e.g.* [Bertsekas, 1997, Proposition 6.1.1]). Concerning time complexity, each step can be performed with $\mathcal{O}(n^2)$ algebraic operations. From a practical perspective, we found that using a change of variable $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}^2$ and optimize $\boldsymbol{\gamma} \in \mathbb{R}^n$ leads to enhanced numerical stability.

Remark 3.3. In the same spirit as remark 3.2, one can express $\mathbf{P}^{\text{se}}$ as a KL projection of $\mathbf{K}_\sigma = \exp(-\mathbf{C}/\sigma)$. Indeed, we show in **??** that if $0 < \sigma \leq \min_i \gamma_i^\star$, then $\mathbf{P}^{\text{se}} = \text{Proj}^{\text{KL}}_{\mathcal{H}_\xi \cap \mathcal{S}}(\mathbf{K}_\sigma)$. This characterization opens the door for alternating Bregman projection methods (described in **??**) which were not found to be more efficient than dual ascent.

Comparison between $\mathbf{P}^{\text{ds}}$ and $\mathbf{P}^{\text{se}}$. In fig. 3.2 we illustrate the ability of our proposed affinity $\mathbf{P}^{\text{se}}$ to adapt to varying noise levels. In the OT problem that we consider, each sample is given a mass of one that is distributed over its neighbors (including itself since self-loops are allowed). For each sample, we refer to the entropy of the distribution over its neighbors as the *spreading* of its mass. One can notice that for $\mathbf{P}^{\text{ds}}$ equation DS (OT problem with global entropy constraint equation EOT) , the samples do not spread their mass evenly depending on the density around them. On the contrary, the per-row entropy constraints of $\mathbf{P}^{\text{se}}$ force equal spreading among samples. This can have benefits, particularly for clustering, as illustrated in the rightmost plot, which shows the eigenvalues of the associated Laplacian matrices (recall that the number of connected components equals the dimension of the null space of its Laplacian Chung [1997b]). As can be seen, $\mathbf{P}^{\text{ds}}$ results in many unwanted clusters, unlike $\mathbf{P}^{\text{se}}$, which is robust to varying noise levels (its Laplacian matrix has only 3 vanishing eigenvalues).

3.3 Optimal Transport for Dimension Reduction with SNEkhorn

In this section, we build upon symmetric entropic affinities to introduce SNEkhorn, a new DR algorithm that fully benefits from the advantages of doubly stochastic affinities.

SNEkhorn's objective. Our proposed method relies on doubly stochastic affinity matrices to capture the dependencies among the samples in both input *and* latent spaces. The KL divergence, which is the central criterion in most popular DR methods Van Assel et al. [2022], is used to measure the discrepancy between the two affinities. As detailed in sections **??** and 3.2, $\mathbf{P}^{\text{se}}$ corrects for heterogeneity in the data density by imposing point-wise entropy constraints. As we do not need such correction for embedding coordinates $\mathbf{Z}$ since they must be optimized, we opt for the standard affinity equation DS built as an OT transport plan with global entropy constraint equation EOT. This OT plan can be efficiently computed using Sinkhorn's algorithm. More precisely, we propose the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times q}} \text{KL}(\mathbf{P}^{\text{se}} | \mathbf{Q}^{\text{ds}}_{\mathbf{Z}}) \,, \tag{SNEkhorn}$$

where $\mathbf{Q}^{\text{ds}}_{\mathbf{Z}} = \exp\left(\mathbf{f}_{\mathbf{Z}} \oplus \mathbf{f}_{\mathbf{Z}} - \mathbf{C}_{\mathbf{Z}}\right)$ stands for the *equation DS* affinity computed with cost $\mathbf{C}_{\mathbf{Z}}$ and $\mathbf{f}_{\mathbf{Z}}$ is the optimal dual variable found by Sinkhorn's algorithm. We set the bandwidth to $\nu = 1$ in $\mathbf{Q}^{\text{ds}}_{\mathbf{Z}}$ similarly to Van der Maaten and Hinton [2008] as the bandwidth in the low dimensional space only affects the scales of the embeddings and

not their shape. Keeping only the terms that depend on $\mathbf{Z}$ and relying on the double stochasticity of $\mathbf{P}^{\text{se}}$, the objective in (SNEkhorn) can be expressed as $\langle \mathbf{P}^{\text{se}}, \mathbf{C}_{\mathbf{Z}} \rangle - 2\langle \mathbf{f}_{\mathbf{Z}}, \mathbf{1} \rangle$.

Heavy-tailed kernel in latent space. Since it is well known that heavy-tailed kernels can be beneficial in DR Kobak et al. [2020], we propose an extension called t-SNEkhorn that simply amounts to computing a doubly stochastic student-t kernel in the low-dimensional space. With our construction, it corresponds to choosing the cost $[\mathbf{C}_{\mathbf{Z}}]_{ij} = \left(\log(1 + \|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2)\right)_{ij}$ instead of $(\|\mathbf{Z}_{i:} - \mathbf{Z}_{j:}\|_2^2)_{ij}$.

Inference. This new DR objective involves computing a doubly stochastic normalization for each update of $\mathbf{Z}$. Interestingly, to compute the optimal dual variable $\mathbf{f}_{\mathbf{Z}}$ in $\mathbf{Q}_{\mathbf{Z}}^{\text{ds}}$, we leverage a well-conditioned Sinkhorn fixed point iteration [Knight et al., 2014, Feydy et al., 2019], which converges extremely fast in the symmetric setting:

$$\forall i,\ [\mathbf{f}_{\mathbf{Z}}]_i \leftarrow \frac{1}{2}\left([\mathbf{f}_{\mathbf{Z}}]_i - \log \sum_k \exp\left([\mathbf{f}_{\mathbf{Z}}]_k - [\mathbf{C}_{\mathbf{Z}}]_{ki}\right)\right) . \tag{Sinkhorn}$$

On the right side of fig. 3.3, we plot $\|\mathbf{Q}_{\mathbf{Z}}^{\text{ds}}\mathbf{1} - \mathbf{1}\|_\infty$ as a function of equation Sinkhorn iterations for a toy example presented in section 3.4. In most practical cases, we found that about 10 iterations were enough to reach a sufficiently small error. $\mathbf{Z}$ is updated through gradient descent with gradients obtained by performing backpropagation through the Sinkhorn iterations. These iterations can be further accelerated with a *warm start* strategy by plugging the $\mathbf{f}_{\mathbf{Z}}$ of the last Sinkhorn to initialize the current one.

Related work. Using doubly stochastic affinities for SNE has been proposed in Lu et al. [2019], with two key differences from our work. First, they do not consider EAs and resort to $\mathbf{P}^{\text{ds}}$ equation DS. This affinity, unlike $\mathbf{P}^{\text{se}}$, is not adaptive to the data heterogeneous density (as illusrated in fig. 3.2). Second, they use the affinity $\widetilde{\mathbf{Q}}_{\mathbf{Z}}$ in the low-dimensional space and demonstrate both empirically and theoretically that matching the latter with a doubly stochastic matrix (*e.g.* $\mathbf{P}^{\text{ds}}$ or $\mathbf{P}^{\text{se}}$) imposes spherical constraints on the embedding $\mathbf{Z}$. This is detrimental for projections onto a $2D$ flat space (typical use case of DR) where embeddings tend to form circles. This can be verified on the left side of fig. 3.3. In contrast, in SNEkhorn, the latent affinity *is also doubly stochastic* so that latent coordinates $\mathbf{Z}$ are not subject to spherical constraints anymore. The corresponding SNEkhorn embedding is shown in fig. 3.4 (bottom right).

3.4 Numerical experiments

This section aims at illustrating the performances of the proposed affinity matrix $\mathbf{P}^{\text{se}}$ equation SEA and DR method SNEkhorn at faithfully representing dependencies and

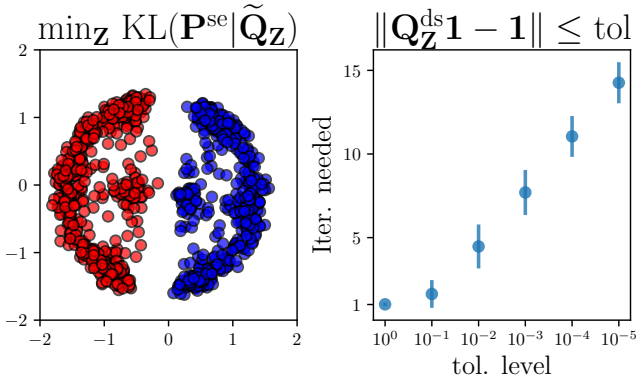

Figure 3.3: Left: SNEkhorn embedding on the simulated data of section 3.4 using $\widetilde{\mathbf{Q}}_{\mathbf{Z}}$ instead of $\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}$ with $\xi = 30$. Right: number of iterations needed to achieve $\|\mathbf{Q}_{\mathbf{Z}}^{\mathrm{ds}}\mathbf{1} - \mathbf{1}\|_\infty \leq \mathrm{tol}$ with equation Sinkhorn.

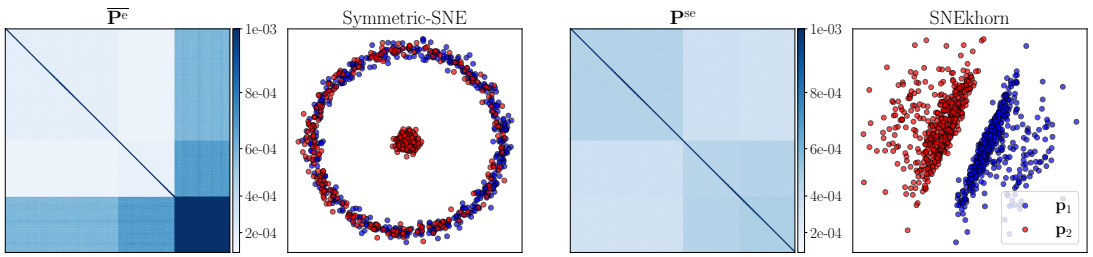

Figure 3.4: From left to right: entries of $\overline{\mathbf{P}^{\mathrm{e}}}$ equation Symmetric-SNE and associated embeddings generated using $\overline{\mathbf{P}^{\mathrm{e}}}$. Then $\mathbf{P}^{\mathrm{se}}$ equation SEA matrix and associated SNEkhorn embeddings. Perplexity $\xi = 30$.

clusters in low dimensions. First, we showcase the relevance of our approach on a simple synthetic dataset with heteroscedastic noise. Then, we evaluate the spectral clustering performances of symmetric entropic affinities before benchmarking t-SNEkhorn with t-SNE and UMAP McInnes et al. [2018] on real-world images and genomics datasets.

Simulated data. We take inspiration from Landa et al. [2021] and consider the task of discriminating between samples from two multinomial distributions. We first sample uniformly two vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ in the 10^4-dimensional probability simplex. We then generate $n = 10^3$ samples as $\mathbf{x}_i = \tilde{\mathbf{x}}_i/(\sum_j \tilde{x}_{ij})$ such that:

$$\tilde{\mathbf{x}}_i \sim \begin{cases} \mathcal{M}(10^3, \mathbf{p}_1), & 1 \leq i \leq 500 \\ \mathcal{M}(10^3, \mathbf{p}_2), & 501 \leq i \leq 750 \\ \mathcal{M}(10^4, \mathbf{p}_2), & 751 \leq i \leq 1000\,. \end{cases}$$

where $\mathcal{M}$ stands for the multinomial distribution. The goal of the task is to test the robustness to heteroscedastic noise. Indeed, points generated using $\mathbf{p}_2$ exhibit different levels of noise due to various numbers of multinomial trials (10^3 and 10^4) to form an estimation of $\mathbf{p}_2$. This typically occurs in real-world scenarios when the same entity is measured using different experimental setups thus creating heterogeneous technical noise levels (*e.g.* in single-cell sequencing Kobak and Berens [2019]). This phenomenon is known as *batch effect* Tran et al. [2020]. In fig. 3.4, we show that, unlike $\overline{\mathbf{P}^{\mathrm{e}}}$ equation Symmetric-SNE, $\mathbf{P}^{\mathrm{se}}$ equation SEA manages to properly filter the noise (top row) to discriminate between samples generated by $\mathbf{p}_1$ and $\mathbf{p}_2$, and represent these two clusters separately in the embedding space (bottom row). In contrast, $\overline{\mathbf{P}^{\mathrm{e}}}$ and SNE are misled by the batch effect. This shows that $\overline{\mathbf{P}^{\mathrm{e}}}$ doesn't fully benefit from the adaptivity of EAs due to poor normalization and symmetrization. This phenomenon partly explains the superiority of SNEkhorn and t-SNEkhorn over current approaches on real-world datasets as illustrated below.

Real-world datasets. We then experiment with various labeled classification datasets including images and genomic data. For images, we use COIL 20 Nene et al. [1996], OLIVETTI faces Ferdinando and Andy [1994], UMNIST Graham and Allinson [1998] and CIFAR 10 Krizhevsky et al. [2009]. For CIFAR, we experiment with features obtained from the last hidden layer of a pre-trained ResNet Phan [2021] while for the other three datasets, we take as input the raw pixel data. Regarding genomics data, we consider the Curated Microarray Database (CuMiDa) Feltes et al. [2019] made of microarray datasets for various types of cancer, as well as the pre-processed SNAREseq (chromatin accessibility) and scGEM (gene expression) datasets used in Demetci et al. [2020a]. For CuMiDa, we retain the datasets with most samples. For all the datasets, when the data dimension exceeds 50 we apply a pre-processing step of PCA in dimension 50, as usually done in practice Van der Maaten and Hinton [2008]. In the following experiments, when not specified the hyperparameters are set to the value leading to the

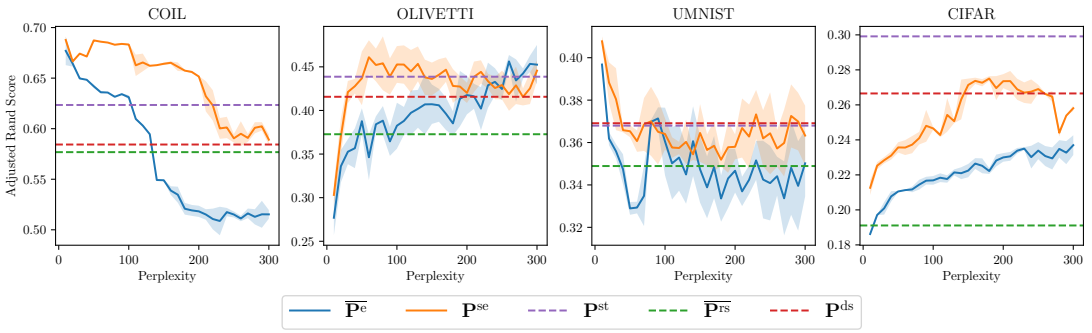

Figure 3.5: ARI spectral clustering score as a function of the perplexity parameter for image datasets.

best average score on five different seeds with grid-search. For perplexity parameters, we test all multiples of 10 in the interval $[10, \min(n, 300)]$ where n is the number of samples in the dataset. We use the same grid for the k of the self-tuning affinity $\mathbf{P}^{\mathrm{st}}$ Zelnik-Manor and Perona [2004] and for the `n_neighbors` parameter of UMAP. For scalar bandwidths, we consider powers of 10 such that the corresponding affinities' average perplexity belongs to the perplexity range.

Spectral Clustering. Building on the strong connections between spectral clustering mechanisms and t-SNE Van Assel et al. [2022], Linderman and Steinerberger [2019] we first consider spectral clustering tasks to evaluate the affinity matrix $\mathbf{P}^{\mathrm{se}}$ equation SEA and compare it against $\overline{\mathbf{P}^{\mathrm{e}}}$ equation Symmetric-SNE. We also consider two versions of the Gaussian affinity with scalar bandwidth $\mathbf{K} = \exp(-\mathbf{C}/\nu)$: the symmetrized row-stochastic $\overline{\mathbf{P}^{\mathrm{rs}}} = \mathrm{Proj}_{\mathcal{S}}^{\ell_2}(\mathbf{P}^{\mathrm{rs}})$ where $\mathbf{P}^{\mathrm{rs}}$ is $\mathbf{K}$ normalized by row and $\mathbf{P}^{\mathrm{ds}}$ equation DS. We also consider the adaptive Self-Tuning $\mathbf{P}^{\mathrm{st}}$ affinity from Zelnik-Manor and Perona [2004] which relies on an adaptive bandwidth corresponding to the distance from the k-th nearest neighbor of each point. We use the spectral clustering implementation of `scikit-learn` Pedregosa et al. [2011] with default parameters which uses the unnormalized graph Laplacian. We measure the quality of clustering using the Adjusted Rand Index (ARI). Looking at both table 3.2 and fig. 3.5, one can notice

Table 3.2: ARI (×100) clustering scores on genomics data.

Data set	$\overline{\mathbf{P}^{\mathrm{rs}}}$	$\mathbf{P}^{\mathrm{ds}}$	$\mathbf{P}^{\mathrm{st}}$	$\overline{\mathbf{P}^{\mathrm{e}}}$	$\mathbf{P}^{\mathrm{se}}$
Liver (14520)	75.8	75.8	84.9	80.8	**85.9**
Breast (70947)	**30.0**	**30.0**	26.5	23.5	28.5
Leukemia (28497)	43.7	44.1	49.7	42.5	**50.6**
Colorectal (44076)	**95.9**	**95.9**	93.9	**95.9**	**95.9**
Liver (76427)	76.7	76.7	**83.3**	81.1	81.1
Breast (45827)	43.6	53.8	74.7	71.5	**77.0**
Colorectal (21510)	57.6	57.6	54.7	**94.0**	79.3
Renal (53757)	47.6	47.6	**49.5**	**49.5**	**49.5**
Prostate (6919)	12.0	13.0	13.2	16.3	**17.4**
Throat (42743)	9.29	9.29	11.4	11.8	**44.2**
scGEM	57.3	58.5	**74.8**	69.9	71.6
SNAREseq	8.89	9.95	46.3	55.4	**96.6**

that, in general, symmetric entropic affinities yield better results than usual entropic affinities with significant improvements in some datasets (*e.g.* throat microarray and SNAREseq). Overall $\mathbf{P}^{\text{se}}$ outperforms all the other affinities in 8 out of 12 datasets. This shows that the adaptivity of EAs is crucial. fig. 3.5 also shows that this superiority is verified for the whole range of perplexities. This can be attributed to the fact that symmetric entropic affinities combine the advantages of doubly stochastic normalization in terms of clustering and of EAs in terms of adaptivity. In the next experiment, we show that these advantages translate into better clustering and neighborhood retrieval at the embedding level when running SNEkhorn.

Dimension Reduction. To guarantee a fair comparison, we implemented not only SNEkhorn, but also t-SNE and UMAP in `PyTorch` Paszke et al. [2017]. All models were optimized using `ADAM` Kingma and Ba [2014] with default parameters and the same stopping criterion: the algorithm stops whenever the relative variation of the loss becomes smaller than 10^{-5}. For each run, we draw independent $\mathcal{N}(0, 1)$ coordinates and use this same matrix to initialize all the methods that we wish to compare. To evaluate the embeddings' quality, we make use of the silhouette Rousseeuw [1987] and trustworthiness Venna and Kaski [2001] scores from `scikit-learn` Pedregosa et al. [2011] with default parameters. While the former relies on class labels, the latter measures the agreement between the neighborhoods in input and output spaces, thus giving two complementary metrics to properly evaluate the embeddings. The results, presented in table 3.3, demonstrate the notable superiority of t-SNEkhorn compared to the commonly used t-SNE and UMAP algorithms. Across the 16 datasets examined, t-SNEkhorn almost consistently outperformed the others, achieving the highest silhouette score on 15 datasets and the highest trustworthiness score on 12 datasets. To visually assess the quality of the embeddings, we provide SNAREseq embeddings in fig. 3.6. Notably, one can notice that the use of t-SNEkhorn results in improved class separation compared to t-SNE.

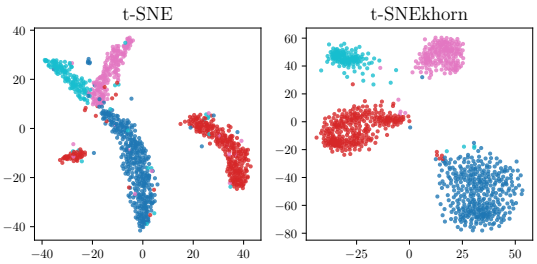

Figure 3.6: SNAREseq embeddings produced by t-SNE and t-SNEkhorn with $\xi = 50$.

3.5 Conclusion

We have introduced a new principled and efficient method for constructing symmetric entropic affinities. Unlike the current formulation that enforces symmetry through an

Table 3.3: Scores for the UMAP, t-SNE and t-SNEkhorn embeddings.

	Silhouette (×100)			Trustworthiness (×100)		
	UMAP	t-SNE	t-SNEkhorn	UMAP	t-SNE	t-SNEkhorn
COIL	20.4 ± 3.3	30.7 ± 6.9	$\mathbf{52.3 \pm 1.1}$	99.6 ± 0.1	99.6 ± 0.1	$\mathbf{99.9 \pm 0.1}$
OLIVETTI	6.4 ± 4.2	4.5 ± 3.1	$\mathbf{15.7 \pm 2.2}$	96.5 ± 1.3	96.2 ± 0.6	$\mathbf{98.0 \pm 0.4}$
UMNIST	-1.4 ± 2.7	-0.2 ± 1.5	$\mathbf{25.4 \pm 4.9}$	93.0 ± 0.4	99.6 ± 0.2	$\mathbf{99.8 \pm 0.1}$
CIFAR	13.6 ± 2.4	18.3 ± 0.8	$\mathbf{31.5 \pm 1.3}$	90.2 ± 0.8	90.1 ± 0.4	$\mathbf{92.4 \pm 0.3}$
Liver (14520)	49.7 ± 1.3	50.9 ± 0.7	$\mathbf{61.1 \pm 0.3}$	89.2 ± 0.7	90.4 ± 0.4	$\mathbf{92.3 \pm 0.3}$
Breast (70947)	28.6 ± 0.8	29.0 ± 0.2	$\mathbf{31.2 \pm 0.2}$	90.9 ± 0.5	91.3 ± 0.3	$\mathbf{93.2 \pm 0.4}$
Leukemia (28497)	22.3 ± 0.7	20.6 ± 0.7	$\mathbf{26.2 \pm 2.3}$	90.4 ± 1.1	92.3 ± 0.8	$\mathbf{94.3 \pm 0.5}$
Colorectal (44076)	67.6 ± 2.2	69.5 ± 0.5	$\mathbf{74.8 \pm 0.4}$	93.2 ± 0.7	93.7 ± 0.5	$\mathbf{94.3 \pm 0.6}$
Liver (76427)	39.4 ± 4.3	38.3 ± 0.9	$\mathbf{51.2 \pm 2.5}$	85.9 ± 0.4	89.4 ± 1.0	$\mathbf{92.0 \pm 1.0}$
Breast (45827)	35.4 ± 3.3	39.5 ± 1.9	$\mathbf{44.4 \pm 0.5}$	93.2 ± 0.4	94.3 ± 0.2	$\mathbf{94.7 \pm 0.3}$
Colorectal (21510)	38.0 ± 1.3	$\mathbf{42.3 \pm 0.6}$	35.1 ± 2.1	85.6 ± 0.7	$\mathbf{88.3 \pm 0.9}$	88.2 ± 0.7
Renal (53757)	44.4 ± 1.5	45.9 ± 0.3	$\mathbf{47.8 \pm 0.1}$	93.9 ± 0.2	$\mathbf{94.6 \pm 0.2}$	94.0 ± 0.2
Prostate (6919)	5.4 ± 2.7	8.1 ± 0.2	$\mathbf{9.1 \pm 0.1}$	77.6 ± 1.8	$\mathbf{80.6 \pm 0.2}$	73.1 ± 0.5
Throat (42743)	26.7 ± 2.4	28.0 ± 0.3	$\mathbf{32.3 \pm 0.1}$	$\mathbf{91.5 \pm 1.3}$	88.6 ± 0.8	86.8 ± 1.0
scGEM	26.9 ± 3.7	33.0 ± 1.1	$\mathbf{39.3 \pm 0.7}$	95.0 ± 1.3	96.2 ± 0.6	$\mathbf{96.8 \pm 0.3}$
SNAREseq	6.8 ± 6.0	35.8 ± 5.2	$\mathbf{67.9 \pm 1.2}$	93.1 ± 2.8	99.1 ± 0.1	$\mathbf{99.2 \pm 0.1}$

orthogonal projection, our approach allows control over the entropy in each point thus achieving entropic affinities' primary goal. Additionally, it produces a DS-normalized affinity and thus benefits from the well-known advantages of this normalization. Our affinity takes as input the same perplexity parameter as EAs and can thus be used with little hassle for practitioners. We demonstrate experimentally that both our affinity and DR algorithm (SNEkhorn), leveraging a doubly stochastic kernel in the latent space, achieve substantial improvements over state-of-the-art approaches.

Note that in the present work we do not address the issue of large-scale dependencies that are not faithfully represented in the low-dimensional space Van Assel et al. [2022]. The latter shall be treated in future works. Among other promising research directions, one could focus on building multi-scale versions of symmetric entropic affinities Lee et al. [2015] as well as fast approximations for SNEkhorn forces by adapting *e.g.* Barnes-Hut Van Der Maaten [2013] or interpolation-based methods Linderman et al. [2019] to the doubly stochastic setting. It could also be interesting to use SEAs in order to study the training dynamics of transformers Zhai et al. [2023].

4

A Probabilistic Graph Coupling View of Dimension Reduction

Contents

This chapter is based on the following publication: Van Assel et al. [2022].

4.1 Introduction

Due to a lack of clear probabilistic foundations, these properties remain mostly empirical. This gap between theory and practice is detrimental as practitioners may rely on strategies that are not optimal for their use case. While recent software developments are making these methods more scalable Chan et al. [2018], Pezzotti et al. [2019], Linderman et al. [2019] and further expanding their use, the need for a well-established probabilistic framework is becoming more prominent. In this work, we define the

generative probabilistic model that encompasses current embedding methods, while establishing new links with the well-established PCA model.

Outline. Consider $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$, an input dataset that consists of n vectors of dimension p. Our task is to embed $\mathbf{X}$ in a lower dimensional space of dimension $q < p$ (typically $q = 2$ for visualization), and we denote by $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)^\top \in \mathbb{R}^{n \times q}$ the unknown embeddings. The rationale of our framework is to suppose that the observations $\mathbf{X}$ and $\mathbf{Z}$ are structured by two latent graphs with $\mathbf{W}_X$ and $\mathbf{W}_Z$ standing for their n-square weight matrices. As the goal of DR is to preserve the input's structure in the latent space, we propose to find the best low-dimensional representation $\mathbf{Z}$ of $\mathbf{X}$ such that $\mathbf{W}_X$ and $\mathbf{W}_Z$ are close. To build a flexible and robust probabilistic framework, we consider random graphs distributed according to some predefined prior distributions. Our objective is to match the posterior distributions of $\mathbf{W}_X$ and $\mathbf{W}_Z$. Note that as they share the same dimensionality the latter graphs can be easily compared unlike $\mathbf{X}$ and $\mathbf{Z}$. The coupling is done with a cross-entropy criterion, the minimization of which will be referred to as graph coupling.

In this work, our main contributions are as follows.

- We show that SNE, t-SNE, LargeVis and UMAP are all instances of graph coupling and characterized by different choices of prior for discrete latent structuring graphs (section 4.4). We demonstrate that such graphs essentially capture conditional independencies among rows through a pairwise Markov Random Field (MRF) model whose construction can be found in section 4.3.

- We uncover the intrinsic probabilistic property explaining why such methods perform poorly on conserving the large-scale structure of the data as a consequence of the degeneracy of the MRF when shift-invariant kernels are used (theorem 4.2). Such degeneracy induces the loss of the relative positions of clusters corresponding to the connected components of the posterior latent graphs whose distributions are identified (proposition 6.1). These findings are highlighted by a new initialization of the embeddings (section 6.1.4).

- We show that for Gaussian MRFs, when adapting graph coupling to precision matrices with suitable priors, PCA appears as a natural extension of the coupling problem in its continuous version (theorem 4.1). Such a model does not suffer from the aforementioned degeneracy hence preserves the large-scale structure.

4.2 PCA as Graph Coupling

As we argue that the inability of SNE-like methods to reproduce the coarse-grain dependencies of the input in the latent space is due to the degeneracy of the conditional

(4.5), a natural solution would be to consider graphical models that are well defined and integrable on the entire definition spaces of $\mathbf{X}$ and $\mathbf{Z}$. For simplicity, we consider the Gaussian model and leave the extension to other kernels for future works. Note that in this case, integrability translates into the precision matrix being full-rank. As we see with the following, the natural extension of our framework to such models leads to a well-established PCA algorithm. In the following, for a continuous variable $\mathbf{\Theta}_Z$, $\mathbb{P}(\mathbf{\Theta}_Z = \cdot)$ denotes its density.

Theorem 4.1. Let $\nu \geq n$, $\mathbf{\Theta}_X \sim \mathcal{W}(\nu, \boldsymbol{I}_n)$ and $\mathbf{\Theta}_Z \sim \mathcal{W}(\nu + p - q, \boldsymbol{I}_n)$. Assume that $\mathbf{\Theta}_X$ and $\mathbf{\Theta}_Z$ structure the rows of respectively $\mathbf{X}$ and $\mathbf{Z}$ such that:

$$\text{vec}(\mathbf{X})|\mathbf{\Theta}_X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_X^{-1} \otimes \boldsymbol{I}_p), \tag{4.1}$$

$$\text{vec}(\mathbf{Z})|\mathbf{\Theta}_Z \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q)\,. \tag{4.2}$$

Then the solution to the precision coupling problem:

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times q}} -\mathbb{E}_{\mathbf{\Theta}_X|\mathbf{X}} \left[\log \mathbb{P}(\mathbf{\Theta}_Z = \mathbf{\Theta}_X|\mathbf{Z})\right]$$

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

We now highlight the parallels with the previous construction done for neighbor embedding methods. First note that the multivariate Gaussian with full-rank precision is inherently a pairwise MRF [Rue and Held, 2005]. When choosing the Gaussian kernel for neighbor embedding methods, we saw that the graph Laplacian $\boldsymbol{L}_X$ of $\mathbf{W}_X$ was playing the role of the among-row precision matrix, as we had $\mathbf{X}|\mathbf{W}_X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{L}_X^{-1} \otimes \boldsymbol{I}_p)$ (equation 4.4). Recall that the latter always has a null space which is spanned by the CC indicator vectors of $\mathbf{W}$ (section 4.3.1). Here, the key difference is that we impose a full-rank constraint on the precision $\mathbf{\Theta}$. Concerning the priors, we choose the ones that are conjugate to the conditionals (4.1) and (4.2), as previously done when constructing the prior for neighbor embedding methods (definition 4.2). Hence in the full-rank setting, the prior simply amounts to a Wishart distribution denoted by $\mathcal{W}$.

The above theorem further highlights the flexibility and generality of the graph coupling framework. Unlike usual constructions of PCA or probabilistic PCA [Tipping and Bishop, 1999], in the above the linear relation between $\mathbf{X}$ and $\mathbf{Z}$ is recovered by solving the graph coupling problem and not explicitly stated beforehand. To the best of our knowledge, it is the first time such a link has been uncovered between PCA and SNE-like methods. In contrast with the latter, PCA is well-known for its ability to preserve global structure while being significantly less efficient at identifying clusters [Anowar et al., 2021]. Therefore, as suspected in section 4.4.2, the degeneracy of the conditional distribution given the graph is key to determining the distance preservation properties of the embeddings. We propose in section 6.1.4 to combine both graph coupling approaches to strike a balance between global and local structure preservation.

4.3 Shift-Invariant Pairwise MRF to Model Row Dependencies

We start by defining the distribution of the observations given a graph. The latter takes the form of a pairwise MRF model which as we show is improper (*i.e.* not integrable on $\mathbb{R}^{n\times p}$) when shift-invariant kernels are used. We consider a fixed directed graph $\mathbf{W} \in \mathcal{S}_W$ where:

$$\mathcal{S}_W = \left\{ \mathbf{W} \in \mathbb{N}^{n\times n} \mid \forall (i,j) \in [\![n]\!]^2, W_{ii} = 0, W_{ij} \leq n \right\}$$

Throughout, $(E, \mathcal{B}(E), \lambda_E)$ denotes a measure space where $\mathcal{B}(E)$ is the Borel σ-algebra on E and λ_E is the Lebesgue measure on E.

4.3.1 Graph Laplacian Null Space

A central element in our construction is the graph Laplacian linear map, defined as follows, where $\mathcal{S}^n_+(\mathbb{R})$ is the set of positive semidefinite matrices.

Definition 4.1. The graph Laplacian operator is the map $L\colon \mathbb{R}^{n\times n}_+ \cap \mathcal{S}^n(\mathbb{R}) \to \mathcal{S}^n_+(\mathbb{R})$ such that

$$\text{for } (i,j) \in [\![n]\!]^2, \quad L(\mathbf{W})_{ij} = \begin{cases} -W_{ij} & \text{if } i \neq j \\ \sum_{k\in[\![n]\!]} W_{ik} & \text{otherwise .} \end{cases}$$

With an abuse of notation, let $\mathbf{L} = L(\overline{\mathbf{W}})$ where $\overline{\mathbf{W}} = \mathbf{W} + \mathbf{W}^\top$. Let $(C_1, ..., C_R)$ be a partition of $[\![n]\!]$ (*i.e.* the set $\{1, 2, ..., n\}$) corresponding to the connected components (CCs) of $\overline{\mathbf{W}}$. As well known in spectral graph theory [Chung, 1997a], the null space of $\mathbf{L}$ is spanned by the orthonormal vectors $\{\mathbf{U}_r\}_{r\in[R]}$ such that for $r \in [R]$, $\mathbf{U}_r = \left(n_r^{-1/2} \mathbb{1}_{i\in C_r}\right)_{i\in[\![n]\!]}$ with $n_r = \mathrm{Card}(C_r)$. By the spectral theorem, $\mathbf{U}_{[R]}$ can be completed such that $\mathbf{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\mathbf{U} = (\mathbf{U}_1, ..., \mathbf{U}_n)$ is orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}((\lambda_i)_{i\in[\![n]\!]})$ with $0 = \lambda_1 = ... = \lambda_R < \lambda_{R+1} \leq ... \leq \lambda_n$.

In what follows, the data is split into two parts: $\mathbf{X}_M$, the orthogonal projection of $\mathbf{X}$ on $\mathcal{S}_M = (\ker \mathbf{L}) \otimes \mathbb{R}^p$, and $\mathbf{X}_C$, the projection on $\mathcal{S}_C = (\ker \mathbf{L})^\perp \otimes \mathbb{R}^p$. For $i \in [\![n]\!]$, $\mathbf{X}_{M,i} = \sum_{r\in[R]} n_r^{-1} \mathbb{1}_{i\in C_r} \sum_{\ell\in C_r} \mathbf{X}_\ell$ hence $\mathbf{X}_M$ stands for the empirical means of $\mathbf{X}$ on CCs, thus modelling the CC positions, while $\mathbf{X}_C = \mathbf{X} - \mathbf{X}_M$ is CC-wise centered, thus modeling the relative positions of the nodes within CCs. We now introduce the probability distribution of these variables.

4.3.2 Pairwise MRF and Shift-Invariances

In this work, the dependency structure among rows of the data is governed by a graph. The strength of the connection between two nodes is given by a symmetric function $k : \mathbb{R}^p \to \mathbb{R}_+$ reformuler symmetric. We consider the following pairwise MRF

unnormalized density function:

$$f_k \colon (\mathbf{X}, \mathbf{W}) \mapsto \prod_{(i,j)\in[\![n]\!]^2} k(\mathbf{x}_i - \mathbf{x}_j)^{W_{ij}} \,. \tag{4.3}$$

As we will see shortly, the above is at the heart of DR methods based on pairwise similarities. Note that as k measures the similarity between couples of samples, f_k will take high values if the rows of $\mathbf{X}$ vary smoothly on the graph $\mathbf{W}$. Thus we can expect $\mathbf{x}_i$ and $\mathbf{x}_j$ to be close if there is an edge between node i and node j in $\mathbf{W}$. A key remark is that f_k is kept invariant by translating $\mathbf{X}_M$. Namely for all $\mathbf{X} \in \mathbb{R}^{n\times p}$, $f_k(\mathbf{X}, \mathbf{W}) = f_k(\mathbf{X}_C, \mathbf{W})$. This invariance results in $f_k(\cdot, \mathbf{W})$ being non integrable on $\mathbb{R}^{n\times p}$, as we see with the following example.

Gaussian kernel. For a positive definite matrix $\mathbf{\Sigma} \in \mathcal{S}^n_{++}(\mathbb{R})$, consider the Gaussian kernel $k : \mathbf{X} \mapsto e^{-\frac{1}{2}\|\mathbf{X}\|^2_{\mathbf{\Sigma}}}$ where $\mathbf{\Sigma}$ stands for the covariance among columns. One has:

$$\log f_k(\mathbf{X}, \mathbf{W}) = -\sum_{(i,j)\in[\![n]\!]^2} W_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbf{\Sigma}} = -\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{X}^T\mathbf{L}\mathbf{X}\right) \tag{4.4}$$

by property of the graph Laplacian (definition 4.1). In this case, it is clear that due to the rank deficiency of $\mathbf{L}$, $f_k(\cdot, \mathbf{W})$ is only $\lambda_{\mathcal{S}_C}$-integrable. In general DR settings one does not want to rely on Gaussian kernels only. A striking example is the use of the Student kernel in t-SNE [van der Maaten and Hinton, 2008]. Heavy-tailed kernels appear useful when the dimension of the embeddings is smaller than the intrinsic dimension of the data [Kobak et al., 2019]. Our contribution provides flexibility by extending the previous result to a large class of kernels, as stated in the following theorem.

Theorem 4.2. If k is $\lambda_{\mathbb{R}^p}$-integrable and bounded above $\lambda_{\mathbb{R}^p}$-almost everywhere then $f_k(\cdot, \mathbf{W})$ is $\lambda_{\mathcal{S}_C}$-integrable.

We refer to section 6.1.2 for the proof. We can now define a distribution on $(\mathcal{S}_C, \mathcal{B}(\mathcal{S}_C))$, where $\mathcal{C}_k(\mathbf{W}) = \int f_k(\cdot, \mathbf{W}) d\lambda_{\mathcal{S}_C}$:

$$\mathbb{P}_k(d\mathbf{X}_C|\mathbf{W}) = \mathcal{C}_k(\mathbf{W})^{-1} f_k(\mathbf{X}_C, \mathbf{W}) \lambda_{\mathcal{S}_C}(d\mathbf{X}_C) \,. \tag{4.5}$$

Remark 4.1. Kernels may have node-specific bandwidths $\boldsymbol{\tau}$, set during a pre-processing step, giving $f_k(\mathbf{X}, \mathbf{W}) = \prod_{(i,j)} k((\mathbf{x}_i - \mathbf{x}_j)/\tau_i)^{W_{ij}}$. Note that such bandwidth does not affect the degeneracy of the distribution and theorem 4.2 still holds.

Between-Rows Dependency Structure. By symmetry of k, reindexing gives: $f_k(\mathbf{X}, \mathbf{W}) = \prod_{j\in[\![n]\!]} \prod_{i\in[\![j]\!]} k(\mathbf{x}_i - \mathbf{x}_j)^{\overline{W}_{ij}}$. Hence distribution equation 4.5 boils down to a pairwise MRF model [Clifford, 1990] with respect to the undirected graph $\overline{\mathbf{W}}$, $\mathcal{C}_k$ playing the role of the partition function. Note that since f_k (Equation 4.3) trivially factorize according to the cliques of $\overline{\mathbf{W}}$, the Hammersley-Clifford theorem ensures that the rows of $\mathbf{X}_C$ satisfy the local and global Markov properties with respect to $\overline{\mathbf{W}}$.

4.3.3 Uninformative Model for CC-wise Means

We showed that the MRF (4.3) is only integrable on $\mathcal{S}_C$, the definition of which depends on the connectivity structure of $\mathbf{W}$. As we now demonstrate, the latter MRF can be seen as a limit of proper distributions on $\mathbb{R}^{n\times p}$, see *e.g.* Rue and Held [2005] for a similar construction in the Gaussian case. We introduce the Borel function $f^\varepsilon(\cdot,\mathbf{W})\colon \mathbb{R}^{n\times p}\to\mathbb{R}_+$ for $\varepsilon>0$ such that for all $\mathbf{X}\in\mathbb{R}^{n\times p}$, $f^\varepsilon(\mathbf{X},\mathbf{W})=f^\varepsilon(\mathbf{X}_M,\mathbf{W})$. To allow f^ε to become arbitrarily non-informative, we assume that for all $\mathbf{W}\in\mathcal{S}_W$, $f^\varepsilon(\cdot,\mathbf{W})$ is $\lambda_{\mathcal{S}_M}$-integrable for all $\varepsilon\in\mathbb{R}_+^*$ and $f^\varepsilon(\cdot,\mathbf{W})\xrightarrow[\varepsilon\to 0]{}1$ almost everywhere. We now define the conditional distribution on $(\mathcal{S}_M,\mathcal{B}(\mathcal{S}_M))$ as follows:

$$\mathbb{P}^\varepsilon(d\mathbf{X}_M|\mathbf{W})=\mathcal{C}^\varepsilon(\mathbf{W})^{-1}f^\varepsilon(\mathbf{X}_M,\mathbf{W})\lambda_{\mathcal{S}_M}(d\mathbf{X}_M) \tag{4.6}$$

where $\mathcal{C}^\varepsilon(\mathbf{W})=\int f^\varepsilon(\cdot,\mathbf{W})d\lambda_{\mathcal{S}_M}$. With this at hand, the joint conditional is defined as the product measure of (4.5) and (4.6) over the row axis, the integrability of which is ensured by the Fubini-Tonelli theorem. In the following we will use the compact notation $\mathcal{C}_k^\varepsilon(\mathbf{W})=\mathcal{C}_k(\mathbf{W})\mathcal{C}^\varepsilon(\mathbf{W})$ for the joint normalizing constant.

Remark 4.2. At the limit $\varepsilon\to 0$ the above construction amounts to setting an infinite variance on the distribution of the empirical means of $\mathbf{X}$ on CCs, thus losing the inter-CC structure.

As an illustration, one can structure the CCs' relative positions according to a Gaussian model with positive definite precision $\varepsilon\boldsymbol{\Theta}\in\mathcal{S}_{++}^R(\mathbb{R})$, as it amounts to choosing $f^\varepsilon:\mathbf{X}\to\exp\left(-\frac{\varepsilon}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{X}^\top\mathbf{U}_{[:R]}\boldsymbol{\Theta}\mathbf{U}_{[:R]}^\top\mathbf{X}\right)\right)$ such that: $\operatorname{vec}(\mathbf{X}_M)|\boldsymbol{\Theta}\sim\mathcal{N}\left(\mathbf{0},\left(\varepsilon\mathbf{U}_{[:R]}\boldsymbol{\Theta}\mathbf{U}_{[:R]}^\top\right)^{-1}\otimes\boldsymbol{\Sigma}\right)$ where $\otimes$ denotes the Kronecker product.

4.4 Graph Coupling as a Unified Objective for Pairwise Similarity Methods

In this section, we show that neighbor embedding methods can be recovered in the presented framework. They are obtained, for particular choices of graph priors, at the limit $\varepsilon\to 0$ when f^ε becomes noninformative and the CCs' relative positions are lost.

We now turn to the priors for $\mathbf{W}$. Our methodology is similar to that of constructing conjugate priors for distributions in the exponential family [Wainwright and Jordan, 2008], notably we insert the cumulant function $\mathcal{C}_k^\varepsilon$ (*i.e.* normalizing constant of the conditional) as a multivariate term of the prior.

We consider different forms: binary (B), unitary out-degree (D) and n-edges (E), relying on an additional term (Ω) to constrain the topology of the graph. For a matrix $\mathbf{A}$, A_{i+} denotes $\sum_j A_{ij}$ and A_{++} denotes $\sum_{ij}A_{ij}$. In the following, $\boldsymbol{\pi}$ plays the role of the edge's prior. The latter can be leveraged to incorporate some additional information

about the dependency structure, for instance when a network is observed as in Li et al. [2020].

Definition 4.2. Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$, $\varepsilon \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$, k satisfies the assumptions of theorem 4.2 and $\mathcal{P} \in \{B, D, E\}$. For $\mathbf{W} \in \mathcal{S}_W$ we introduce:

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

where $\Omega_B(\mathbf{W}) = \prod_{ij} \mathbb{1}_{W_{ij}\leq 1}$, $\Omega_D(\mathbf{W}) = \prod_i \mathbb{1}_{W_{i+}=1}$ and $\Omega_E(\mathbf{W}) = \mathbb{1}_{W_{++}=n} \prod_{ij}(W_{ij}!)^{-1}$.

When $\alpha = 0$, the above no longer depends on ε and k. We will use the compact notation $\mathbb{P}_{\mathcal{P}}(\mathbf{W}; \boldsymbol{\pi}) = \mathbb{P}^{\varepsilon}_{\mathcal{P},k}(\mathbf{W}; \boldsymbol{\pi}, 0)$. Note that by $\mathbf{W} \sim \mathbb{P}_{\mathcal{P}}(\cdot\,; \boldsymbol{\pi})$ we have the following simple Bernoulli ($\mathcal{B}$) and multinomial ($\mathcal{M}$) distributions, where matrix or vector division is to be understood as element-wise.

- If $\mathcal{P} = B$, $\forall (i,j) \in [\![n]\!]^2$, $W_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathcal{B}\left(\pi_{ij}/(1+\pi_{ij})\right)$.
- If $\mathcal{P} = D$, $\forall i \in [\![n]\!]$, $\mathbf{W}_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{M}\left(1, \boldsymbol{\pi}_i/\pi_{i+}\right)$.
- If $\mathcal{P} = E$, $\mathbf{W} \sim \mathcal{M}\left(n, \boldsymbol{\pi}/\pi_{++}\right)$.

We now show that the posterior distribution of the graph given the observations takes a simple form when the distribution of CC empirical means $\boldsymbol{X}_M$ diffuses *i.e.* when $\varepsilon \to 0$ (a proof of the following result can be found in section 6.1.3). In the following, $\odot$ stands for the Hadamard product and $\mathcal{D}$ for the convergence in distribution.

Proposition 4.1. Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$, k satisfies the assumptions of theorem 4.2 with $\mathbf{K}_X = (k(\mathbf{X}_i - \mathbf{X}_j))_{(i,j)\in[\![n]\!]^2}$ and $\mathcal{P} \in \{B, D, E\}$. If $\mathbf{W}^{\varepsilon} \sim \mathbb{P}^{\varepsilon}_{\mathcal{P},k}(\cdot\,; \boldsymbol{\pi}, 1)$ then

$$\mathbf{W}^{\varepsilon} | \mathbf{X} \xrightarrow[\varepsilon\to 0]{\mathcal{D}} \mathbb{P}_{\mathcal{P}}(\cdot\,; \boldsymbol{\pi} \odot \mathbf{K}_X)\,.$$

Remark 4.3. For all $\mathbf{W} \in \mathcal{S}_W$, $\mathcal{C}^{\varepsilon}(\mathbf{W})$ diverges as $\varepsilon \to 0$, hence the graph prior (definition 4.2) is improper at the limit. This compensates for the uninformative diffuse conditional and allows to retrieve a well-defined tractable posterior limit.

4.4.1 Retrieving Popular Dimension Reduction Methods

We now provide a unified view of neighbor embedding objectives as a coupling between graph posterior distributions. To that extent, we derive the cross entropy associated with the various graph priors at hand. In what follows, k_x and k_z satisfy the assumptions of theorem 4.2 and we denote by $\mathbf{K}_X$ and $\mathbf{K}_Z$ the associated kernel matrices on $\mathbf{X}$ and $\mathbf{Z}$ respectively. For both graph priors we consider the parameters $\boldsymbol{\pi} = \mathbf{1}$ and $\alpha = 1$. For

$(\mathcal{P}_X, \mathcal{P}_Z) \in \{B, D, E\}^2$, we introduce the cross entropy between the limit posteriors at $\varepsilon \to 0$,

$$\mathcal{H}_{\mathcal{P}_X,\mathcal{P}_Z} = -\mathbb{E}_{\mathbf{W}_X \sim \mathbb{P}_{\mathcal{P}_X}(\cdot;\mathbf{K}_X)}[\log \mathbb{P}_{\mathcal{P}_Z}(\mathbf{W}_Z = \mathbf{W}_X; \mathbf{K}_Z)]$$

defining a coupling criterion to be optimized with respect to embedding coordinates $\mathbf{Z}$. We now go through each couple $(\mathcal{P}_X, \mathcal{P}_Z)$ such that $\operatorname{supp}(\mathbb{P}_{\mathcal{P}_X}) \subset \operatorname{supp}(\mathbb{P}_{\mathcal{P}_Z})$ for the cross-entropy to be defined.

SNE. When $\mathcal{P}_X = \mathcal{P}_Z = D$, the probability of the limit posterior graphs factorizes over the nodes and the cross-entropy between limit posteriors takes the form of the objective of SNE [Hinton and Roweis, 2002], where for $i \in [\![n]\!]$, $\mathbf{P}^D_{i:} = [\mathbf{K}_X]_{i:} / \sum_j [\mathbf{K}_X]_{ij}$ and $\mathbf{Q}^D_{i:} = [\mathbf{K}_Z]_{i:} / \sum_j [\mathbf{K}_Z]_{ij}$,

$$\mathcal{H}_{D,D} = -\sum_{i \neq j} P^D_{ij} \log Q^D_{ij} \,.$$

Symmetric-SNE. Choosing $\mathcal{P}_X = D$ and $\mathcal{P}_Z = E$, we define for $(i,j) \in [\![n]\!]^2$, $Q^E_{ij} = [\mathbf{K}_Z]_{ij} / \sum_{k,\ell} [\mathbf{K}_Z]_{k\ell}$ and $\overline{P}^D_{ij} = P^D_{ij} + P^D_{ji}$. The symmetry of $\mathbf{Q}^E$ yields:

$$\mathcal{H}_{D,E} = -\sum_{i \neq j} P^D_{ij} \log Q^E_{ij} = -\sum_{i<j} \overline{P}^D_{ij} \log Q^E_{ij}$$

and the symmetrized objective of t-SNE [van der Maaten and Hinton, 2008] is recovered.

LargeVis. Now choosing $\mathcal{P}_X = D$ and $\mathcal{P}_Z = B$, one can also notice that $\mathbf{Q}^B = ([\mathbf{K}_Z]_{ij}/(1 + [\mathbf{K}_Z]_{ij}))_{(i,j)\in[\![n]\!]^2}$ is symmetric. With this at hand the limit cross-entropy reads

$$\begin{aligned}
\mathcal{H}_{D,B} &= -\sum_{i \neq j} P^D_{ij} \log Q^B_{ij} + \left(1 - P^D_{ij}\right) \log\left(1 - Q^B_{ij}\right) \\
&= -\sum_{i<j} \overline{P}^D_{ij} \log Q^B_{ij} + \left(2 - \overline{P}^D_{ij}\right) \log(1 - Q^B_{ij})
\end{aligned}$$

which is the objective of LargeVis Tang et al. [2016].

UMAP. Let us take $\mathcal{P}_X = \mathcal{P}_Z = B$ and consider the symmetric thresholded graph $\widetilde{\mathbf{W}}_X = \mathbb{1}_{\mathbf{W}_X + \mathbf{W}_X^\top \geq 1}$. By independence of the edges, $[\widetilde{\mathbf{W}}_X]_{ij} \sim \mathcal{B}\left(P^B_{ij}\right)$ where $\widetilde{P}^B_{ij} = P^B_{ij} + P^B_{ji} - P^B_{ij} P^B_{ji}$ and $\mathbf{P}^B = ([\mathbf{K}_X]_{ij}/(1 + [\mathbf{K}_X]_{ij}))_{(i,j)\in[\![n]\!]^2}$. Coupling $\widetilde{\mathbf{W}}_X$ and $\mathbf{W}_Z$ gives:

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

Table 4.1: Prior distributions for $\mathbf{W}_X$ and $\mathbf{W}_Z$ associated with the pairwise similarity coupling DR algorithms. Grey-colored boxes are such that the cross-entropy is undefined.

$\mathcal{P}_X$ \ $\mathcal{P}_Z$	B	D	E
$\widetilde{B}$	UMAP		
D	LargeVis	SNE	T-SNE

which is the loss function considered in UMAP McInnes et al. [2018], the construction of $\widetilde{\mathbf{W}}_X$ being borrowed from section 3.1 of the paper.

Remark 4.4. One can also consider $\mathcal{H}_{E,E}$ but as detailed in van der Maaten and Hinton [2008], this criterion fails at positioning outliers and is therefore not considered. Interestingly, any other feasible combination of the presented priors relates to an existing method.

4.4.2 Interpretations

As we have seen in section 4.4.1, SNE-like methods can all be derived from the graph coupling framework. What characterizes each of them is the choice of priors considered for the latent structuring graphs. To the best of our knowledge, the presented framework is the first that manages to unify all these DR algorithms. Such a framework opens many perspectives for improving upon current practices as we discuss in section 6.1.4 and section 4.5. We now focus on a few insights that our work provides about the empirical performances of these methods.

Repulsion & Attraction. Decomposing $\mathcal{H}_{\mathcal{P}_X,\mathcal{P}_Z}$ with Bayes' rule and simplifying constant terms one has the following optimization problem:

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} -\sum_{(i,j)\in[\![n]\!]^2} \mathbf{P}_{ij}^{\mathcal{P}_X} \log k_z(\mathbf{z}_i - \mathbf{z}_j) + \log \mathbb{P}(\mathbf{Z}). \tag{4.7}$$

The first and second terms in eq. (4.7) respectively summarize the attractive and repulsive forces of the objective. Recall from proposition 6.1 that $\mathbf{P}^{\mathcal{P}_X}$ is the posterior expectation of $\mathbf{W}_X$. Hence in SNE-like methods, the attractive forces resume to a pairwise MRF log likelihood with respect to a graph posterior expectation given $\mathbf{X}$. For instance if k_z is the Gaussian kernel, this attractive term reads $\mathrm{tr}\left(\mathbf{Z}^\top \boldsymbol{L}^\star \mathbf{Z}\right)$ where $\boldsymbol{L}^\star = \mathbb{E}_{\mathbf{W}\sim\mathbb{P}_{\mathcal{P}_X}(\cdot;\mathbf{K}_X)}[L(\mathbf{W})]$, boiling down to the objective of Laplacian eigenmaps Belkin and Niyogi [2003]. Therefore, for Gaussian MRFs, the attractive forces resume to

an unconstrained Laplacian eigenmaps objective. Such link, already noted in Carreira-Perpinán [2010], is easily unveiled in our framework. Moreover, one can notice that only this attractive term depends on $\mathbf{X}$ as the repulsion is given by the marginal term in (4.7). The latter reads $\mathbb{P}(\mathbf{Z}) = \sum_{\mathbf{W}\in\mathcal{S}_W} \mathbb{P}(\mathbf{Z}, \mathbf{W})$ with $\mathbb{P}(\mathbf{Z}, \mathbf{W}) \propto f_k(\mathbf{Z}, \mathbf{W})\Omega_{\mathcal{P}_Z}(\mathbf{W})$. Such penalty notably prevents a trivial solution, as $\mathbf{0}$, like any constant vector, is a mode of $f_k(\cdot, \mathbf{W})$ for all $\mathbf{W}$. Also note that the prior for $\mathbf{W}_X$ only conditions attraction while the prior for $\mathbf{W}_Z$ only affects repulsion. In the present work we focus solely on deciphering the probabilistic model that accounts for neighbor embedding loss functions and refer to Böhm et al. [2020] for a quantitative study of attraction and repulsion in these methods.

Global Structure Preservation. To gain intuition, consider that $\mathbf{W}_X$ is observed. As we showed in section 4.3.2, when one relies on shift-invariant kernels, the positions of the CC means are taken from a diffuse distribution. Since the above methods are all derived from the limit posteriors at $\varepsilon \to 0$, $\mathbf{X}_M$ and $\mathbf{Z}_M$ have no influence on the coupling objective. Hence if two nodes belong to different CCs, their low dimensional pairwise distance will likely not be faithful. We can expect this phenomenon to persist when the expectation on $\mathbf{W}_X$ is considered, especially when clusters are well distinguishable in $\mathbf{X}$. This observation is central to understand the large scale deficiency of these methods. Note that this happens at the benefit of the local structure which is faithfully represented in low dimension, as discussed in section 4.1. In the following section we propose to mitigate the global structure deficiency with non-degenerate MRF models.

4.5 Conclusion and Perspectives

In this work, we shed new light on the most popular DR methods by showing that they can be unified within a common probabilistic model in the form of latent Markov Random Fields Graphs coupled by a cross-entropy. The definition of such a model constitutes a major step towards the understanding of common dimension reduction methods, in particular their structure preservation properties as discussed in this article.

Our work offers many perspectives, among which the possibility to enrich the probabilistic model with more suited graph priors. Currently considered priors are simply the ones that are conjugate to the MRFs thus they are mostly designed to yield a tractable coupling objective. However they may not be optimal and could be modified to capture targeted features, *e.g.* communities, in the input data, and give adapted representations in the latent space. The graph coupling approach could also be extended to more general latent structures governing the joint distribution of observations. Finally, the probabilistic model could be leveraged to tackle hyper-parameter calibration, especially kernel bandwidths that have a great influence on the

quality of the representations and are currently tuned using heuristics with unclear motivations.

5

Distributional Reduction

Contents

This chapter is based on the following publication: .

5.1 Introduction

Two sides of the same coin. As a matter of fact, methods from both families share many similitudes, including the construction of a similarity graph between input samples. In clustering, many popular approaches design a reduced or coarsened version of the initial similarity graph while preserving some of its spectral properties Von Luxburg [2007], Schaeffer [2007]. In DR, the goal is to solve the inverse problem of finding low-dimensional embeddings that generate a similarity graph close to the one computed from input data points Ham et al. [2004], Hinton and Roweis [2002]. Our work builds on these converging viewpoints and addresses the following question: *can DR and clustering be expressed in a common and unified framework ?*

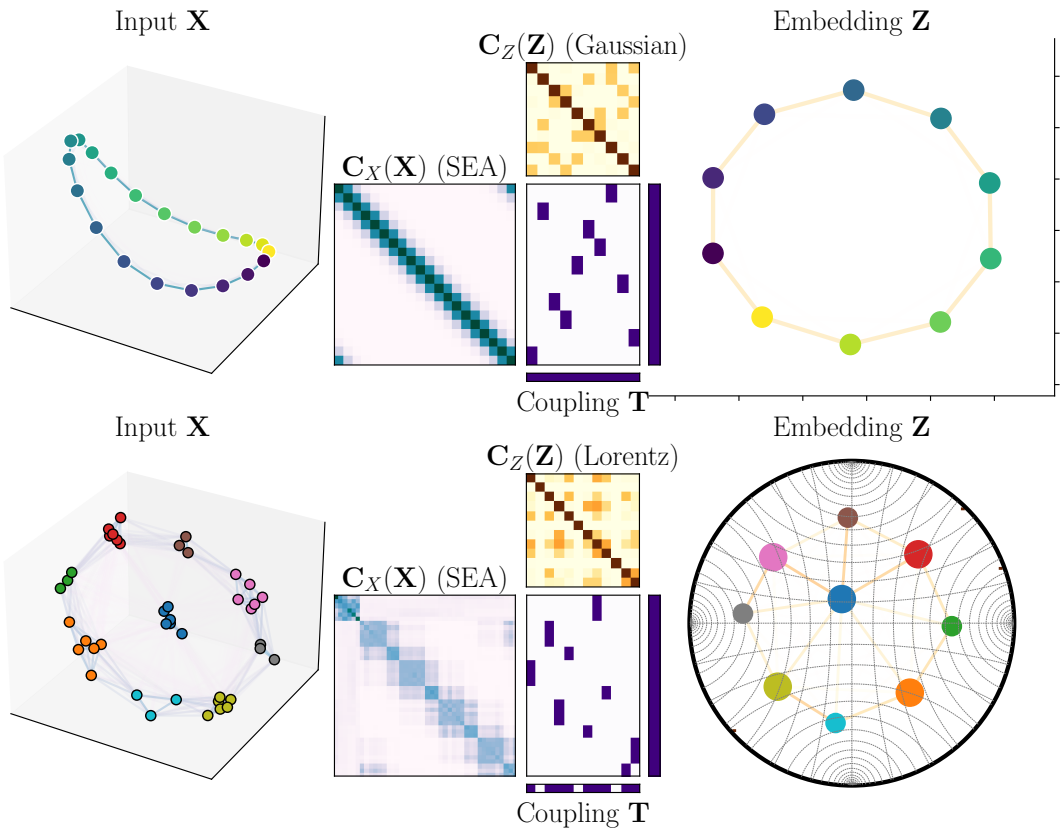

Figure 5.1: Illustration of our DistR method on two toys examples: points arranged on a circle (*top row*) and clusters with varying sizes (*bottom*) both in 3 dimensions. *Middle column*: input similarity matrix $\mathbf{C}_X(\mathbf{X})$ and final resulting embedding similarity $\mathbf{C}_Z(\mathbf{Z})$. Both are coupled through the coupling matrix $\mathbf{T}$ (depicted in purple, with its marginals)

A distributional perspective. To answer this question, we propose to look at both problems from a distributional point of view, treating the data as an empirical probability distribution $\mu = \frac{1}{N}\sum_i \delta_{\mathbf{x}_i}$. This enables us to consider statistical measures of similarity such as Optimal Transport (OT), which is at the core of our work. On the one hand, OT and clustering are strongly related. The celebrated K-means approach can be seen as a particular case of minimal Wasserstein estimator where a distribution of n Diracs is optimized *w.r.t* their weights and positions Canas and Rosasco [2012]. Other connections between spectral clustering and the OT-based Gromov-Wasserstein (GW) distance have been recently developed in Chowdhury and Needham [2021], Chen et al. [2023], Vincent-Cuaz et al. [2022a]. On the other hand, the link between DR and OT has not been explored yet. DR methods, when modeling data as distributions, usually focus on joint distribution between samples within each space separately, see *e.g.* Van Assel et al.

[2023] or Lu et al. [2019]. Consequently, they do not consider couplings to transport samples across spaces of varying dimensions.

Our contributions. In this paper, we propose to bridge this gap by proposing a novel and general distributional framework that encompasses both DR and clustering as special cases. We notably cast those problems as finding a reduced distribution that minimizes the GW divergence from the original empirical data distribution. Our method proceeds by first constructing an input similarity matrix $\mathbf{C}_X(\mathbf{X})$ that is matched with the embedding similarity $\mathbf{C}_Z(\mathbf{Z})$ through an OT coupling matrix $\mathbf{T}$. The latter establishes correspondences between input and embedding samples. We illustrate this principle in Figure 5.1 where one can notice that $\mathbf{C}_Z(\mathbf{Z})$ preserves the topology of $\mathbf{C}_X(\mathbf{X})$ with a reduced number of nodes. The adaptivity of our model that can select an effective number of cluster $< n$, is visible in the bottom plot, where only the exact number of clusters in the original data (9 out of the 12 initially proposed) is automatically recovered. Our method can operate in any embedding space, which is illustrated by projecting in either a 2D Euclidean plane or a Poincaré ball as embedding spaces.

We show that this framework is versatile and allows to recover many popular DR methods such as the kernel PCA and neighbor embedding algorithms, but also clustering algorithms such as K-means and spectral clustering. We first prove in Section 5.2 that DR can be formulated as a GW projection problem under some conditions on the loss and similarity functions. We then propose in Section 5.3 a novel formulation of data summarization as a minimal GW estimator that allows to select both the dimensionality of the embedding d (DR) but also the number of Diracs n (Clustering). Finally, we show in section **??** the practical interest of our approach, which regularly outperforms its competitors for various joint DR/Clustering tasks.

5.2 Dimensionality Reduction as OT

In this section, we present the strong connections between the classical eq. (DR) and the GW problem.

Gromov-Monge interpretation of DR. As suggested by eq. (DR), dimension reduction seeks to find embeddings $\mathbf{Z}$ so that the similarity between the (i, j) samples of the input data is as close as possible to the similarity between the (i, j) samples of the embeddings. Under reasonable assumptions about $\mathbf{C}_Z$, this also amounts to identifying the embedding $\mathbf{Z}$ *and* the best permutation that realigns the two similarity matrices. Recall that the function $\mathbf{C}_Z$ is equivariant by permutation, if, for any $N \times N$ permutation matrix $\mathbf{P}$ and any $\mathbf{Z}$, $\mathbf{C}_Z(\mathbf{PZ}) = \mathbf{P}\mathbf{C}_Z(\mathbf{Z})\mathbf{P}^\top$ Bronstein et al. [2021]. This type of assumption is natural for $\mathbf{C}_Z$: if we rearrange the order of samples (*i.e.* , the rows of $\mathbf{Z}$), we expect

the similarity matrix between the samples to undergo the same rearrangement.

Lemma 5.1. Let $\mathbf{C}_Z$ be a permutation equivariant function and L any loss. The minimum eq. (DR) is equal to

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times d}} \min_{\sigma\in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)}) \,. \tag{5.1}$$

Also, any sol. $\mathbf{Z}$ of eq. (DR) is such that $(\mathbf{Z}, \mathrm{id})$ is sol. of eq. (5.1) and conversely any $(\mathbf{Z}, \sigma)$ sol. of eq. (5.1) is such that $\mathbf{Z}$ is a sol of eq. (DR) up to σ.

See proof in **??**. The correspondence established between eq. (DR) and eq. (5.1) unveils a *quadratic problem* similar to GW. Specifically, eq. (5.1) relates to the Gromov-Monge problem[1] Mémoli and Needham [2018] which seeks to identify, by solving a quadratic assignment problem Cela [2013], the permutation σ that best aligns two similarity matrices. Lemma 5.1 therefore shows that the best permutation is the identity when we also optimize the embedding $\mathbf{Z}$. We can delve deeper into these comparisons and demonstrate that the general formulation of dimension reduction is also equivalent to minimizing the Gromov-Wasserstein objective, which serves as a relaxation of the Gromov-Monge problem Mémoli and Needham [2022].

DR as GW Minimization. We suppose that the distributions have the same number of points ($N = n$) and uniform weights ($\mathbf{h}_Z = \mathbf{h}_X = \frac{1}{N}\mathbf{1}_N$). We recall that a matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$ is conditionally positive definite (CPD), *resp.* conditionally negative definite (CND), if it is symmetric and $\forall \mathbf{x} \in \mathbb{R}^N, \mathbf{x}^\top \mathbf{1}_N = 0$ s.t. $\mathbf{x}^\top \mathbf{C}\mathbf{x} \geq 0$, *resp.* ≤ 0.

We thus have the following theorem that extends Lemma 5.1 to the GW problem (proof can be found in **??**).

Theorem 5.2. The minimum eq. (DR) is equal to $\min_{\mathbf{Z}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}), \mathbf{C}_Z(\mathbf{Z}), \frac{1}{N}\mathbf{1}_N, \frac{1}{N}\mathbf{1}_N)$ in the following settings:

(i) (spectral methods) $\mathbf{C}_X(\mathbf{X})$ is any matrix, $L = L_2$ and $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$.

(ii) (neighbor embedding methods) $\mathrm{Im}(\mathbf{C}_X) \subseteq \mathbb{R}^{N\times N}_{>0}$, $L = L_{\mathrm{KL}}$, the matrix $\mathbf{C}_X(\mathbf{X})$ is CPD and, for any $\mathbf{Z}$,

$$\mathbf{C}_Z(\mathbf{Z}) = \mathrm{diag}(\boldsymbol{\alpha}_\mathbf{Z})\mathbf{K}_\mathbf{Z}\,\mathrm{diag}(\boldsymbol{\beta}_\mathbf{Z})\,, \tag{5.2}$$

where $\boldsymbol{\alpha}_\mathbf{Z}, \boldsymbol{\beta}_\mathbf{Z} \in \mathbb{R}^N_{>0}$ and $\mathbf{K}_\mathbf{Z} \in \mathbb{R}^{N\times N}_{>0}$ is such that $\log(\mathbf{K}_\mathbf{Z})$ is CPD.

Remarkably, this result shows that *all spectral DR methods* can be seen as OT problems in disguise, as they all equivalently minimize a GW problem. The second

[1] Precisely $\min_{\sigma\in S_N} \sum_{ij} L([\mathbf{C}_X(\mathbf{X})]_{ij}, [\mathbf{C}_Z(\mathbf{Z})]_{\sigma(i)\sigma(j)})$ is the Gromov-Monge discrepancy between two discrete distributions, with the same number of atoms and uniform weights.

point of the theorem also provides some insights into this equivalence in the case of neighbor embedding methods. For instance, the Gaussian kernel $\mathbf{K_Z}$, used extensively in DR, satisfies the hypothesis as $\log(\mathbf{K_Z}) = (-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2)_{ij}$ is CPD (see *e.g.* Maron and Lipman 2018). The terms $\boldsymbol{\alpha}_\mathbf{Z}, \boldsymbol{\beta}_\mathbf{Z}$ also allow for considering all the usual normalizations of $\mathbf{K_Z}$: by a scalar so as to have $\sum_{ij}[\mathbf{C}_Z(\mathbf{Z})]_{ij} = 1$, but also any row-stochastic or doubly stochastic normalization (with the Sinkhorn-Knopp algorithm Sinkhorn and Knopp 1967).

Matrices satisfying $\log(\mathbf{K_Z})$ being CPD are well-studied in the literature and are known as infinitely divisible matrices Bhatia [2006]. It is noteworthy that the t-Student kernel does not fall into this category. Moreover, in the aforementioned neighbor embedding methods, the matrix $\mathbf{C}_X(\mathbf{X})$ is generally not CPD. The intriguing question of generalizing this result with weaker assumptions on $\mathbf{C}_Z$ and $\mathbf{C}_X$ remains open for future research. Interestingly, we have observed in the numerical experiments performed in **??** that the symmetric entropic affinity of Van Assel et al. [2023] was systematically CPD.

Remark 5.1. In **??** of the appendix we also provide other sufficient conditions for neighbor embedding methods with the cross-entropy loss $L(x, y) = x \log(x/y)$. They rely on specific structures for $\mathbf{C}_Z$ but do not impose any assumptions on $\mathbf{C}_X$. Additionally, in **??**, we provide a *necessary* condition based on a bilinear relaxation of the GW problem. Although its applicability is limited due to challenges in proving it in full generality, it requires minimal assumptions on $\mathbf{C}_X, \mathbf{C}_Z$ and L.

In essence, both Lemma 5.1 and Theorem 5.2 indicate that dimensionality reduction can be reframed from a distributional perspective, with the search for an empirical distribution that aligns with the data distribution in the sense of optimal transport, through the lens of GW. In other words, DR is informally solving $\min_{\mathbf{z}_1,\cdots,\mathbf{z}_N} \mathrm{GW}(\frac{1}{n}\sum_{i=1}^N \delta_{\mathbf{x}_i}, \frac{1}{n}\sum_{i=1}^N \delta_{\mathbf{z}_i})$.

5.3 Distributional Reduction

The previous interpretation is significant because it allows for two generalizations. Firstly, beyond solely determining the positions $\mathbf{z}_i$ of Diracs (as in classical DR) we can now optimize *the mass* of the distribution μ_Z. This is interpreted as finding the relative importance of each point in the embedding $\mathbf{Z}$. More importantly, due to the flexibility of GW, we can also seek a distribution in the embedding with a smaller number of points $n < N$. This will result in both reducing the dimension *and* clustering the points in the embedding space through the optimal coupling. Informally, our *Distributional Reduction* (DistR) framework aims at solving $\min_{\mu_Z \in \mathcal{P}_n(\mathbb{R}^d)} \mathrm{GW}(\frac{1}{n}\sum_{i=1}^N \delta_{\mathbf{x}_i}, \mu_Z)$.

5.3.1 Distributional Reduction Problem

Precisely, the optimization problem that we tackle in this paper can be formulated as follows

$$\min_{\substack{\mathbf{Z}\in\mathbb{R}^{n\times d}\\ \mathbf{h}_Z\in\Sigma_n}} \mathrm{GW}_L(\mathbf{C}_X(\mathbf{X}),\mathbf{C}_Z(\mathbf{Z}),\mathbf{h}_X,\mathbf{h}_Z) \tag{DistR}$$

This problem comes down to learning the closest graph $(\mathbf{C}_Z(\mathbf{Z}),\mathbf{h}_Z)$ parametrized by $\mathbf{Z}$ from $(\mathbf{C}_X(\mathbf{X}),\mathbf{h}_X)$ in the GW sense. When $n<N$, the embeddings $(\mathbf{z}_1,\cdots,\mathbf{z}_n)$ then act as *low-dimensional prototypical examples* of input samples, whose learned relative importance $\mathbf{h}_Z$ accommodates clusters of varying proportions in the input data $\mathbf{X}$ (see Figure 5.1). We refer to them as *prototypes*. The weight vector $\mathbf{h}_X$ is typically assumed to be uniform, that is $\mathbf{h}_X=\frac{1}{N}\mathbf{1}_N$, in the absence of prior knowledge. As discussed in Section 5.2, traditional DR amounts to setting $n=N,\mathbf{h}_Z=\frac{1}{N}\mathbf{1}_N$.

Clustering with DistR. One notable aspect of our model is its capability to simultaneously perform dimensionality reduction and clustering. Indeed, the optimal coupling $\mathbf{T}\in[0,1]^{N\times n}$ of problem eq. (DistR) is, by construction, a soft-assignment matrix from the input data to the embeddings. It allows each point $\mathbf{x}_i$ to be linked to one or more prototypes $\mathbf{z}_j$ (clusters). In Section 5.3.2 we explore conditions where these soft assignments transform into hard ones, such that each point is therefore linked to a unique prototype/cluster.

A semi-relaxed objective. For a given embedding $\mathbf{Z}$ and $L=L_2$, it is known that minimizing in $\mathbf{h}_Z$ the DistR objective is *equivalent* to a problem that is computationally simpler than the usual GW one, namely the semi-relaxed GW divergence srGW_L Vincent-Cuaz et al. [2022a]:

$$\min_{\mathbf{T}\in\mathcal{U}_n(\mathbf{h}_X)} E_L(\mathbf{C}_X(\mathbf{X}),\mathbf{C}_Z(\mathbf{Z}),\mathbf{T})\,, \tag{srGW}$$

where $\mathcal{U}_n(\mathbf{h}_X):=\left\{\mathbf{T}\in\mathbb{R}_+^{N\times n}:\mathbf{T}\mathbf{1}_n=\mathbf{h}_X\right\}$. To efficiently address eq. (DistR), we first observe that this equivalence holds for any inner divergence L with a straightforward adaptation of the proof in Vincent-Cuaz et al. [2022a]. Additionally, we prove that srGW_L remains a divergence as soon as L is itself a divergence. Consequently, srGW_L vanishes iff both measures are isomorphic in a weak sense Chowdhury and Mémoli [2019]. We emphasize that taking a proper divergence L is important (and basic assumptions on $\mathbf{X}$), as it avoids some trivial solutions as detailed in Appendix **??**.

Interestingly, srGW projections, *i.e.* optimizing only the weights $\mathbf{h}_Z$ over simple fixed supports $\mathbf{Z}$, have already remarkable representational capability. We illustrate this in **??**, by considering projections of a real-world dataset over 2D grids of increasing resolutions.

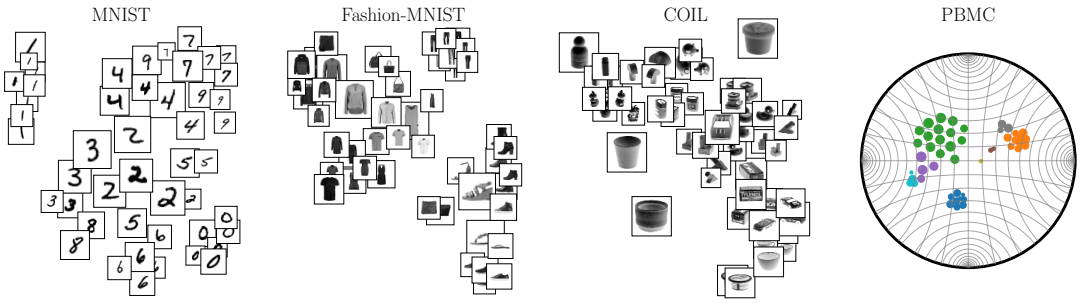

Figure 5.2: Examples of 2-dimensional embeddings produced by DistR using the SEA similarity for $\mathbf{C}_X$ and the Student's kernel for $\mathbf{C}_Z$. The latter computed in $\mathbb{R}^2$ for the first three datasets and the Poincaré ball for the last one. Displayed images are medoids for each cluster *i.e.* $\arg\max_i [\mathbf{C}_X(\mathbf{X})\mathbf{T}_{:,k}]_i$ for cluster k. The area of image k is proportional to $[\mathbf{h}_Z]_k$.

Setting $\mathbf{C}_X(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{C}_Z(\mathbf{X}) = \mathbf{Z}\mathbf{Z}^\top$, we can see that those projections recover faithful coarsened representations of the embeddings learned using PCA. DistR aims to exploit the full potential of this divergence by learning a few optimal prototypes that best represent the dataset.

5.3.2 Clustering Properties

In addition to the connections established between DistR and DR methods in Section 5.2, we elaborate now on the links between DistR and clustering methods.

In what follows, we call a coupling $\mathbf{T} \in [0,1]^{N\times n}$ with a single non-null element per row a *membership matrix*. When the coupling is a membership matrix each data point is associated with a single prototype thus achieving a hard clustering of the input samples.

We will see that a link can be drawn with kernel K-means using the analogy of *GW barycenters*. More precisely the *srGW barycenter* Vincent-Cuaz et al. [2022a] seeks for a closest target graph $(\overline{\mathbf{C}}, \overline{\mathbf{h}})$ from $(\mathbf{C}_X, \mathbf{h}_X)$ by solving

$$\min_{\overline{\mathbf{C}}\in\mathbb{R}^{n\times n}} \min_{\mathbf{T}\in\mathcal{U}_n(\mathbf{h}_X)} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})\,. \tag{srGWB}$$

We emphasize that the only (important) difference between eq. (srGWB) and eq. (DistR) is that there is no constraint imposed on $\overline{\mathbf{C}}$ in srGWB. In contrast, eq. (DistR) looks for minimizing over $\overline{\mathbf{C}} \in \{\mathbf{C}_Z(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N\times d}\}$. For instance, choosing $\mathbf{C}_Z(\mathbf{Z}) = \mathbf{Z}\mathbf{Z}^\top$ in eq. (DistR) is equivalent to enforcing $\text{rank}(\overline{\mathbf{C}}) \leq d$ in eq. (srGWB).

We establish below that srGWB is of particular interest for clustering. The motivation for this arises from the findings of Chen et al. [2023], which demonstrate that when $\mathbf{C}_X(\mathbf{X})$ is positive semi-definite and $\mathbf{T}$ is *constrained* to belong to the set of membership matrices (as opposed to couplings in $\mathcal{U}_n(\mathbf{h})$), eq. (srGWB) is equivalent to a kernel K-means whose samples are weighted by $\mathbf{h}_X$ Dhillon et al. [2004, 2007]. These additional constraints are in fact unnecessary since we show below that the original srGWB problem

admits membership matrices as the optimal coupling for a broader class of $\mathbf{C}_X(\mathbf{X})$ input matrices (see proofs in appendix **??**).

Theorem 5.3. Let $\mathbf{h}_X \in \Sigma_N$ and $L = L_2$. Suppose that for any $\mathbf{X} \in \mathbb{R}^{N\times p}$ the matrix $\mathbf{C}_X(\mathbf{X})$ is CPD or CND. Then the problem eq. (srGWB) admits a membership matrix as optimal coupling, *i.e.* , there is a minimizer of $\mathbf{T} \in \mathcal{U}_n(\mathbf{h}_X) \rightarrow \min_{\overline{\mathbf{C}}\in\mathbb{R}^{n\times n}} E_L(\mathbf{C}_X(\mathbf{X}), \overline{\mathbf{C}}, \mathbf{T})$ with only one non-zero value per row.

Theorem 5.3 and relations proven in Chen et al. [2023] provide that eq. (srGWB) is equivalent to the aforementioned kernel K-means when $\mathbf{C}_X(\mathbf{X})$ is positive semi-definite. Moreover, as the (hard) clustering property holds for more generic types of matrices, namely CPD and CND, srGWB stands out as a fully-fledged clustering method. Although these results do not apply directly to DistR, we argue that they further legitimize the use of GW projections for clustering. Interestingly, we also observe in practice that the couplings obtained by DistR are always membership matrices, regardless of $\mathbf{C}_Z$. Further research will be carried out to better understand this phenomenon.

5.3.3 Computation

DistR is a non-convex problem that we propose to tackle using a Block Coordinate Descent algorithm (BCD, Tseng 2001) guaranteed to converge to local optimum Grippo and Sciandrone [2000], Lyu and Li [2023]. The BCD alternates between the two following steps. First, we optimize in $\mathbf{Z}$ for a fixed transport plan using gradient descent with adaptive learning rates Kingma and Ba [2014]. Then we solve for a srGW problem given $\mathbf{Z}$. To this end, we benchmarked both the Conditional Gradient and Mirror Descent algorithms proposed in Vincent-Cuaz et al. [2022a], extended to support losses L_2 and L_{KL}.

Following Proposition 1 in [Peyré et al., 2016], a vanilla implementation leads to $\mathcal{O}(nN^2 + n^2N)$ operations to compute the loss or its gradient. In many DR methods, $\mathbf{C}_X(\mathbf{X})$ or $\mathbf{C}_Z(\mathbf{Z})$, or their transformations within the loss L, admit explicit low-rank factorizations. Including *e.g.* matrices involved in spectral methods and other similarity matrices derived from squared Euclidean distance matrices Scetbon et al. [2022]. In these settings, we exploit these factorizations to reduce the computational complexity of our solvers down to $\mathcal{O}(Nn(p+d) + (N+m)pd + n^2)$ when $L = L_2$, and $O(Nnd + n^2d)$ when $L = L_{\mathrm{KL}}$. We refer the reader interested in these algorithmic details to Appendix **??**.

5.3.4 Related Work

The closest to our work is the CO-Optimal-Transport (COOT) clustering approach proposed in Redko et al. [2020] that estimates simultaneously a clustering of samples

and features through the CO-Optimal Transport problem,

$$\min_{\substack{\mathbf{T}_1 \in \mathcal{U}(\mathbf{h}_1, \overline{\mathbf{h}}_1) \\ \mathbf{T}_2 \in \mathcal{U}(\mathbf{h}_2, \overline{\mathbf{h}}_2)}} \sum_{ijkl} (X_{ik} - Z_{jl})^2 [\mathbf{T}_1]_{ij} [\mathbf{T}_2]_{kl} \,, \tag{COOT}$$

where $\mathbf{h}_1 \in \Sigma_N$, $\overline{\mathbf{h}}_1 \in \Sigma_n$, $\mathbf{h}_2 \in \Sigma_p$ and $\overline{\mathbf{h}}_2 \in \Sigma_d$. We emphasize that COOT-clustering, which consists in optimizing the COOT objective above *w.r.t.* $\mathbf{Z}$, is a linear DR model as the reduction is done with the map $\mathbf{T}_2$. In contrast, DistR leverages the more expressive non-linear similarity functions of existing DR methods. Other joint DR-clustering approaches, such as Liu et al. [2022a], involve modeling latent variables by a mixture of distributions. In comparison, our framework is more versatile, as it can easily adapt to any $(L, \mathbf{C}_X, \mathbf{C}_Z)$ of existing DR methods (Section 2.5.1).

6

Proofs and Additional Results

Contents

6.1 Appendix of Chapter 4

6.1.1 Proof of Theorem 4.1

Theorem 4.1. Let $\nu \geq n$, $\boldsymbol{\Theta}_X \sim \mathcal{W}(\nu, \boldsymbol{I}_n)$ and $\boldsymbol{\Theta}_Z \sim \mathcal{W}(\nu + p - q, \boldsymbol{I}_n)$. Assume that $\boldsymbol{\Theta}_X$ and $\boldsymbol{\Theta}_Z$ structure the rows of respectively $\mathbf{X}$ and $\mathbf{Z}$ such that:

$$\mathrm{vec}(\mathbf{X})|\boldsymbol{\Theta}_X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_X^{-1} \otimes \boldsymbol{I}_p), \tag{4.1}$$

$$\mathrm{vec}(\mathbf{Z})|\boldsymbol{\Theta}_Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q)\,. \tag{4.2}$$

Then the solution to the precision coupling problem:

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times q}} -\mathbb{E}_{\boldsymbol{\Theta}_X|\mathbf{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X|\mathbf{Z})\right]$$

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

Proof. We consider the following hierarchical model, for $\nu_X, \nu_Z \geq n$:

$$\begin{aligned}
\boldsymbol{\Theta}_X &\sim \mathcal{W}(\nu_X, \boldsymbol{I}_n) \\
\mathrm{vec}(\mathbf{X})|\boldsymbol{\Theta}_X &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_X^{-1} \otimes \boldsymbol{I}_p) \\
\boldsymbol{\Theta}_Z &\sim \mathcal{W}(\nu_Z, \boldsymbol{I}_n) \\
\mathrm{vec}(\mathbf{Z})|\boldsymbol{\Theta}_Z &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q) \, .
\end{aligned}$$

With this at hand, the posteriors for $\boldsymbol{\Theta}_X$ and $\boldsymbol{\Theta}_Z$ can be derived in closed form:

$$\begin{aligned}
\boldsymbol{\Theta}_X|\mathbf{X} &\sim \mathcal{W}(\nu_X + p, \left(\boldsymbol{I}_n + \mathbf{X}\mathbf{X}^\top\right)^{-1}) \\
\boldsymbol{\Theta}_Z|\mathbf{Z} &\sim \mathcal{W}(\nu_Z + q, \left(\boldsymbol{I}_n + \mathbf{Z}\mathbf{Z}^\top\right)^{-1}) \, .
\end{aligned}$$

Keeping terms of $-\mathbb{E}_{\boldsymbol{\Theta}_X}[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X|\mathbf{Z})|\mathbf{X}]$ that depends on $\mathbf{Z}$, one has the optimization problem:

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times q}} \quad \frac{\nu_X + p}{2} \operatorname{tr}\left(\mathbf{Z}^\top (\boldsymbol{I}_n + \mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{Z}\right) - \frac{\nu_Z + q}{2} \log|\boldsymbol{I}_n + \mathbf{Z}\mathbf{Z}^\top|$$

Consider the eigendecomposition of the sample covariance matrices: $\mathbf{X}\mathbf{X}^\top = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$ and $\mathbf{Z}\mathbf{Z}^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{d})$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ such that $d_1 \geq ... \geq d_n$ and $\lambda_1 \geq ... \geq \lambda_n$. Denoting $\gamma = (\nu_X + q)/(\nu_Z + p)$, we consider the following problem:

$$\begin{aligned}
\min_{\boldsymbol{U} \in \mathcal{O}(n), \boldsymbol{\Lambda}} \quad & \operatorname{tr}\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \boldsymbol{V} (\boldsymbol{I}_n + \boldsymbol{D})^{-1} \boldsymbol{V}^\top\right) - \gamma \log|\boldsymbol{I}_n + \boldsymbol{\Lambda}| && (6.1) \\
\text{s.t.} \quad & \boldsymbol{\Lambda} \succcurlyeq \mathbf{0} && (6.2) \\
& \mathrm{rank}(\boldsymbol{\Lambda}) \leq q && (6.3)
\end{aligned}$$

Note that the above problem is non-convex because of the rank constraint (6.3).

First, we focus on finding the optimal eigenvectors. To that extent, let us denote, $\boldsymbol{R} = \boldsymbol{U}^\top \boldsymbol{V}$. Only the left term in (6.4) depends on $\boldsymbol{R}$. The optimization problem for eigenvectors writes:

$$\min_{\boldsymbol{R} \in \mathcal{O}(n)} \quad \operatorname{tr}\left(\boldsymbol{R}^\top \boldsymbol{\Lambda} \boldsymbol{R} (\boldsymbol{I}_n + \boldsymbol{D})^{-1}\right) \tag{6.4}$$

The objective (6.4) can be expressed as: $\sum_{(i,j) \in [\![n]\!]^2} \lambda_i (1 + d_j)^{-1} R_{ij}^2$. Now one can notice that since $\boldsymbol{R}$ is orthogonal, $\boldsymbol{R} \odot \boldsymbol{R}$ is doubly stochastic (*i.e.* sum of coefficients on each row and column is equal to one). Therefore thanks to the Birkhoff–von Neumann theorem, there exists $\theta_1, ..., \theta_L \geq 0$, $\sum_{\ell \in [\![L]\!]} \theta_\ell = 1$ and permutation matrices $\boldsymbol{P}_1, ..., \boldsymbol{P}_L$ such that:

$$\boldsymbol{R} \odot \boldsymbol{R} = \sum_{\ell \in [\![L]\!]} \theta_\ell \boldsymbol{P}_\ell$$

where for all $\ell \in [\![L]\!]$, there exists a permutation σ_ℓ of $[\![n]\!]$ such that $P_{\ell,ij} = \mathbb{1}_{\sigma_\ell(i)=j}$ for $(i,j) \in [\![n]\!]^2$.

With this at hand, objective (6.4) writes: $\sum_{\ell\in[\![L]\!]}\theta_\ell\sum_{i\in[\![n]\!]}\lambda_i(1+d_{\sigma_\ell(i)})^{-1}$. There exists a permutation $\sigma^\star$ such that the quantity $\sum_{i\in[\![n]\!]}\lambda_i(1+d_{\sigma_\ell(i)})^{-1}$ is minimal. Note that the identity permutation *i.e.* for $i\in[\![n]\!]$, $\sigma(i)=i$ is optimal in this case as the $(\lambda_i)_{i\in[\![n]\!]}$ and the $(d_i)_{i\in[\![n]\!]}$ are in decreasing order. Then choosing for $\ell\in[\![L]\!]$, $\theta_\ell=\mathbb{1}_{\sigma_\ell=\sigma^\star}$ minimizes the latter quantity. Therefore the solution of (6.4) $\boldsymbol{R}^\star$ is such that for $(i,j)\in[\![n]\!]^2$, $R^\star_{ij}=\pm\mathbb{1}_{\sigma^\star(i)=j}$. Thus an optimum in $\boldsymbol{U}$ of 6.4 is such that $\boldsymbol{U}^\star=\boldsymbol{V}\boldsymbol{R}^\star$.

Hence $\boldsymbol{U}=\boldsymbol{V}$, in particular, is optimal. We will choose this $\boldsymbol{U}$ in what follows as the sign of the axes do not influence the characterization of the final result in $\mathbf{Z}$ as a PCA embedding. Such a choice gives $\mathbf{Z}\mathbf{Z}^\top=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$.

Now it remains to find the optimal eigenvalues $(\lambda_i)_{i\in[\![n]\!]}$. The rank constraint (6.3) can be easily dealt with: since the eigenvalues are sorted in decreasing order, the constraint implies that for $i\geq q$, $\lambda_i=0$. Thus the eigenvalue problem can be formulated in $\mathbb{R}^q$:

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^q}\quad \boldsymbol{\lambda}^\top(\mathbf{1}+\boldsymbol{d})^{-1}-\gamma\mathbf{1}^\top\log(\mathbf{1}+\boldsymbol{\lambda}) \tag{6.5}$$

$$\text{s.t.}\quad \forall i\in[q],\quad \lambda_i\geq 0,\quad \lambda_1\geq\ldots\geq\lambda_q \tag{6.6}$$

where (6.6) accounts for (6.2). The above is convex. (6.5) is minimized for $\boldsymbol{\lambda}=\gamma(\mathbf{1}+\boldsymbol{d})-\mathbf{1}$. Taking the feasibility constraint (6.6) into account one has a solution $\boldsymbol{\lambda}^*$ such that:

$$\forall i\in[\![n]\!],\quad \lambda_i^*=\begin{cases}\max(0,\gamma(1+d_i)-1) & \text{if}\quad i\leq q\\ 0 & \text{otherwise}\,.\end{cases}$$

Note that this solution is not unique if there are repeated eigenvalues. Notice also that one has the freedom to choose the Wishart prior parameters such that $\gamma=1$. Doing so, the solution satisfies $\mathbf{Z}^\star{\mathbf{Z}^\star}^T=\boldsymbol{V}_{[:,q]}\boldsymbol{D}_{[q,q]}\boldsymbol{V}^\top_{[q,:]}$. Therefore there exists $\boldsymbol{R}$ an orthogonal matrix of size q such that $\mathbf{Z}^\star=\boldsymbol{V}_{[:,q]}\boldsymbol{D}^{\frac{1}{2}}_{[q,q]}\boldsymbol{R}$. The latter is the output of a PCA model of $\mathbf{X}$ with q components, which is defined up to a rotation. □

6.1.2 Proof of Theorem 4.2

Theorem 4.2. If k is $\lambda_{\mathbb{R}^p}$-integrable and bounded above $\lambda_{\mathbb{R}^p}$-almost everywhere then $f_k(\cdot,\mathbf{W})$ is $\lambda_{\mathcal{S}_C}$-integrable.

Proof. $\mathbf{W}\in\mathcal{S}_W$ is the weight matrix of a graph with R connected components $\{C_1,\ldots,C_R\}$ partitioning $[\![n]\!]$. Since k is upper bounded by a constant, there exists $M_+>1$ that upper bounds k. Let $\boldsymbol{\mathcal{T}}$ be the adjacency matrix of a spanning forest of

$\mathbf{W}$, since each edge of $\mathbf{W}$ is bounded by n, one has:

$$\begin{aligned}\int f_k(\mathbf{X},\mathbf{W})\lambda_{\mathcal{S}_C}(d\mathbf{X}) &= \int \prod_{(i,j)\in[\![n]\!]^2} k(\mathbf{x}_i-\mathbf{x}_j)^{W_{ij}}\lambda_{\mathcal{S}_C}(d\mathbf{X}) \\ &\leq M_+^{n^3}\int \prod_{(i,j)\in[\![n]\!]^2} k(\mathbf{x}_i-\mathbf{x}_j)^{\mathcal{T}_{ij}}\lambda_{\mathcal{S}_C}(d\mathbf{X}) \\ &\leq M_+^{n^3}\prod_{r\in[\![R]\!]}\int \prod_{(i,j)\in C_r^2} k(\mathbf{x}_i-\mathbf{x}_j)^{\mathcal{T}_{ij}}\lambda_{\mathcal{S}_C}(d\mathbf{X})\,. \end{aligned} \tag{6.7}$$

Let $r\in[\![R]\!]$. The spanning tree corresponding to the r^{th} connected component called $\boldsymbol{\mathcal{T}}^r$ has exactly n_r-1 edges. There exists a leaf node $\ell\in[\![n]\!]$ of $\boldsymbol{\mathcal{T}}^r$ and let $\tilde{\ell}$ be the node linked to it. Consider a bijective map $\sigma\colon C_r\backslash\{\ell\}\to[\![n_r-1]\!]$ such that $\sigma(\tilde{\ell})=1$ and for $(i,j)\in(C_r\backslash\{\ell\})^2$, $\sigma(i)\leq\sigma(j)$ implies that node i has a shorter path on $\overline{\boldsymbol{\mathcal{T}}^r}$[1] to ℓ than node j. There exists a bijective map $e\colon[\![2:n_r-1]\!]\to[\![n_r-2]\!]$ such that for $i\in[\![2:n_R-1]\!]$, $\overline{\boldsymbol{\mathcal{T}}^r}_{\sigma^{-1}(i),\sigma^{-1}(e(i))}>0$ and node $\sigma^{-1}(e(i))$ has a shorter path on $\overline{\boldsymbol{\mathcal{T}}^r}$ to node ℓ than node $\sigma^{-1}(i)$.

Recall that since $\mathbf{X}\in\mathcal{S}_C$ one has: $\sum_{i\in C_r}\mathbf{x}_i=0$ hence $\mathbf{x}_\ell=-\sum_{i\neq\ell}\mathbf{x}_i$. Let us now consider the linear map ϕ^r such that:

$$\forall i\in[n_r-1],\quad \phi^r(\mathbf{x}_i)=\begin{cases}\mathbf{x}_{\sigma^{-1}(i)}+\sum_{j\in[\![n_r-1]\!]}\mathbf{x}_{\sigma^{-1}(j)} & \text{if i = 1}\\ \mathbf{x}_{\sigma^{-1}(i)}-\mathbf{x}_{\sigma^{-1}(e(i))} & \text{otherwise}\,.\end{cases}$$

We now show that the change of variable ϕ^r is a $\mathcal{C}^1$ diffeomorphism by proving that its Jacobian has full rank. Ordering the columns with the map σ, the latter takes the form:

$$\mathbf{J}_{\phi^r}=\begin{pmatrix}2 & 1 & 1 & \dots & 1\\ & 1 & 0 & \dots & 0\\ & & \ddots & \ddots & \vdots\\ & \mathbf{A} & & \ddots & 0\\ & & & & 1\end{pmatrix}$$

where $\mathbf{A}$ is a strictly lower triangular matrix such that for all $i\in[\![2:n_r-1]\!]$, $A_{ie(i)}=-1$ and for all $t\neq e(i)$, $A_{it}=0$. The above can be factorized as:

$$\mathbf{J}_{\phi^r}=\begin{pmatrix}\alpha_{n_r-1} & \alpha_{n_r-2} & \dots & \alpha_2 & \alpha_1\\ 0 & 1 & 0 & \dots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \ddots & 0\\ 0 & \dots & \dots & 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}1 & 0 & \dots & \dots & 0\\ & 1 & \ddots & & \vdots\\ & & \ddots & \ddots & \vdots\\ & \mathbf{A} & & \ddots & 0\\ & & & & 1\end{pmatrix}$$

[1]Symmetrized version *i.e.* $\overline{\boldsymbol{\mathcal{T}}^r}=\boldsymbol{\mathcal{T}}^r+(\boldsymbol{\mathcal{T}}^r)^\top$.

where $\alpha_1 = -1$ and for $\ell > 1$, $\alpha_\ell = \sum_{j<l} \alpha_j \mathbb{1}_{e(n_r-j)=n_r-\ell} - 1$. With this in place, for $i \in [\![n_r - 1]\!]$, $\alpha_i \neq 0$ in particular $\alpha_{n_r-1} \neq 0$ therefore $|\mathbf{J}_{\phi^r}| \neq 0$ and ϕ^r is a $\mathcal{C}^1$ diffeomorphism. This change of variable yields:

$$\begin{aligned}\int \prod_{(i,j)\in C_r^2} k(\mathbf{x}_i - \mathbf{x}_j)^{\mathcal{T}_{ij}} \lambda_{\mathcal{S}_C}(d\mathbf{X}) &= \int \bigotimes_{i\in[\![n_r-1]\!]} k(\mathbf{y}_i)|\mathbf{J}_{\phi^r}(\mathbf{Y})|^{-1}\lambda_{\mathbb{R}^p}(d\mathbf{Y}) \\ &= |\mathbf{J}_{\phi^r}|^{-1} \prod_{i\in[\![n_r-1]\!]} \int k(\mathbf{y}_i)\lambda_{\mathbb{R}^p}(d\mathbf{y}_i)\end{aligned}$$

using the Fubini Tonelli theorem. The result follows from $\lambda_{\mathbb{R}^p}$-integrability of k and upper bound 6.7. □

6.1.3 Proof of Proposition 6.1

Proposition 6.1. Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$, k satisfies the assumptions of theorem 4.2 with $\mathbf{K}_X = (k(\mathbf{X}_i - \mathbf{X}_j))_{(i,j)\in[\![n]\!]^2}$ and $\mathcal{P} \in \{B, D, E\}$. If $\mathbf{W}^\varepsilon \sim \mathbb{P}^\varepsilon_{\mathcal{P},k}(\cdot\,;\boldsymbol{\pi}, 1)$ then

$$\mathbf{W}^\varepsilon|\mathbf{X} \xrightarrow[\varepsilon\to 0]{\mathcal{D}} \mathbb{P}_{\mathcal{P}}(\cdot\,;\boldsymbol{\pi}\odot\mathbf{K}_X)\,.$$

Proof. Let $\mathcal{P} \in \{B, D, E\}$, k be a valid kernel (assumptions of theorem 4.2) with $\mathbf{K}_X = (k(\mathbf{x}_i - \mathbf{x}_j))_{(i,j)\in[\![n]\!]^2}$ and $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$. Let $\mathbf{W} \sim \mathbb{P}^\varepsilon_{\mathcal{P},k}(\cdot;\boldsymbol{\pi}, 1)$. Inversion of conditional with Bayes rule gives:

$$\forall \boldsymbol{W} \in \mathcal{S}_W, \quad \mathbb{P}(\mathbf{W}|\boldsymbol{X}) \propto \mathcal{C}_k^\varepsilon(\mathbf{W})^{-1} f^\varepsilon(\mathbf{X}, \mathbf{W}) f_k(\mathbf{X}, \mathbf{W}) \mathbb{P}^\varepsilon_{\mathcal{P},k}(\mathbf{W}; \boldsymbol{\pi}, 1) \tag{6.8}$$

where the prior reads:

$$\mathbb{P}^\varepsilon_{\mathcal{P},k}(\boldsymbol{W}; \boldsymbol{\pi}, 1) \propto \mathcal{C}_k^\varepsilon(\mathbf{W})\Omega_{\mathcal{P}}(\mathbf{W}) \prod_{(i,j)\in[\![n]\!]^2} \pi_{ij}^{W_{ij}}\,. \tag{6.9}$$

Hence the joint normalizing constant simplifies such that:

$$\forall \boldsymbol{W} \in \mathcal{S}_W, \quad \mathbb{P}(\mathbf{W}|\boldsymbol{X}) \propto f^\varepsilon(\mathbf{X}, \mathbf{W})\Omega_{\mathcal{P}}(\mathbf{W}) \prod_{(i,j)\in[\![n]\!]^2} (\pi_{ij} k(\mathbf{x}_i - \mathbf{x}_j))^{W_{ij}} \tag{6.10}$$

$$\xrightarrow[\varepsilon\to 0]{} \Omega_{\mathcal{P}}(\mathbf{W}) \prod_{(i,j)\in[\![n]\!]^2} (\pi_{ij} k(\mathbf{x}_i - \mathbf{x}_j))^{W_{ij}} \tag{6.11}$$

which ends the proof. As a complement, we now explicit the simple forms taken by the posterior limit graph in each case.

***B*-Prior.** Recall that in this case the prior reads:

$$\mathbb{P}^\varepsilon_B(\boldsymbol{W}; \boldsymbol{\pi}, 1) \propto \mathcal{C}_k^\varepsilon(\boldsymbol{W}) \prod_{(i,j)\in[\![n]\!]^2} \pi_{ij}^{W_{ij}} \mathbb{1}_{W_{ij}\leq 1}\,.$$

Therefore the posterior limit graph has the distribution:

$$\begin{aligned}\mathbb{P}_B(\mathbf{W};\boldsymbol{\pi}\odot\boldsymbol{K}_X) &= \frac{\prod_{(i,j)\in[\![n]\!]^2}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{ij}\leq 1}}{\sum_{\mathbf{W}\in\mathcal{S}_W}\prod_{(i,j)\in[\![n]\!]^2}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{ij}\leq 1}}\\ &= \prod_{(i,j)\in[\![n]\!]^2}\left(\frac{\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}{1+\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}\right)^{W_{ij}}\left(\frac{1}{1+\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}\right)^{1-W_{ij}}\mathbb{1}_{W_{ij}\leq 1}\,.\end{aligned}$$

This distribution amounts to: $\forall(i,j)\in[\![n]\!]^2,\quad \mathbf{W}_{ij}\overset{\perp\!\!\!\perp}{\sim}\mathcal{B}\left(\frac{\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}{1+\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}\right)$.

***D*-Prior.** The prior writes:

$$\mathbb{P}^{\varepsilon}_D(\boldsymbol{W};\boldsymbol{\pi},1)\propto\mathcal{C}^{\varepsilon}_k(\boldsymbol{W})\prod_{(i,j)\in[\![n]\!]^2}\pi_{ij}^{W_{ij}}\mathbb{1}_{W_{i+}=1}\,.$$

The distribution of the posterior limit then becomes:

$$\begin{aligned}\mathbb{P}_D(\mathbf{W};\boldsymbol{\pi}\odot\boldsymbol{K}_X) &= \frac{\prod_{(i,j)\in[\![n]\!]^2}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{i+}=1}}{\sum_{\mathbf{W}\in\mathcal{S}_W}\prod_{(i,j)\in[\![n]\!]^2}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{i+}=1}}\\ &= \frac{\prod_{(i,j)\in[\![n]\!]^2}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{i+}=1}}{\prod_{i\in[\![n]\!]}\sum_{\ell\in[\![n]\!]}\pi_{i\ell}k(\mathbf{x}_i-\mathbf{x}_\ell)}\\ &= \prod_{(i,j)\in[\![n]\!]^2}\left(\frac{\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}{\sum_{\ell\in[\![n]\!]}\pi_{i\ell}k(\mathbf{x}_i-\mathbf{x}_\ell)}\right)^{W_{ij}}\mathbb{1}_{W_{i+}=1}\,.\end{aligned}$$

This distribution amounts to: $\forall i\in[\![n]\!],\quad \mathbf{W}_i\overset{\perp\!\!\!\perp}{\sim}\mathcal{M}\left(1,\left(\frac{\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}{\sum_{\ell\in[\![n]\!]}\pi_{i\ell}k(\mathbf{x}_i-\mathbf{x}_\ell)}\right)_{j\in[\![n]\!]}\right)$.

***E*-Prior.** In this case the prior reads:

$$\mathbb{P}^{\varepsilon}_E(\boldsymbol{W};\boldsymbol{\pi},1)\propto\mathcal{C}^{\varepsilon}_k(\boldsymbol{W})\prod_{(i,j)\in[\![n]\!]^2}\frac{\pi_{ij}^{W_{ij}}}{W_{ij}!}\mathbb{1}_{W_{++}=n}\,.$$

Finally, deriving the distribution of the posterior graph limit:

$$\begin{aligned}\mathbb{P}_E(\mathbf{W};\boldsymbol{\pi}\odot\boldsymbol{K}_X) &= \frac{\prod_{(i,j)\in[\![n]\!]^2}(W_{ij}!)^{-1}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{++}=n}}{\sum_{\mathbf{W}\in\mathcal{S}_W}\prod_{(i,j)\in[\![n]\!]^2}(W_{ij}!)^{-1}\left(\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)\right)^{W_{ij}}\mathbb{1}_{W_{++}=n}}\\ &= n!\prod_{(i,j)\in[\![n]\!]^2}(W_{ij})^{-1}\left(\frac{\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}{\sum_{(\ell,t)\in[\![n]\!]^2}\pi_{\ell t}k(\mathbf{x}_\ell-\mathbf{x}_t)}\right)^{W_{ij}}\mathbb{1}_{W_{++}=n}\,.\end{aligned}$$

This distribution amounts to: $\mathbf{W}\sim\mathcal{M}\left(n,\left(\frac{\pi_{ij}k(\mathbf{x}_i-\mathbf{x}_j)}{\sum_{(\ell,t)\in[\![n]\!]^2}\pi_{\ell t}k(\mathbf{x}_\ell-\mathbf{x}_t)}\right)_{(i,j)\in[\![n]\!]^2}\right)$. □

6.1.4 Towards Capturing Large-Scale Dependencies

In this section, we investigate the ability of graph coupling to faithfully represent the global structure in low dimensions. To gain intuition on the case where the distribution induced by the graph is not degenerate, we consider a proper Gaussian graph coupling model and show its equivalence with PCA. We then provide a new initialization procedure to alleviate the large-scale deficiency of graph coupling when degenerate MRFs are used.

Hierarchical Graph Coupling

The goal of this section is to show that global structure in SNE-like embeddings can be improved by structuring the CCs' positions. We consider the following hierarchical model for $\mathbf{X}$, where $\mathcal{P}_X \in \{B, D, E\}$, k_x satisfies the assumptions of theorem 4.2 and $\nu_X \geq n$:

$$\begin{aligned} \mathbf{W}_X &\sim \mathbb{P}^{\varepsilon}_{\mathcal{P}_X, k_x}(\cdot\,; \mathbf{1}, 1), \quad \boldsymbol{\Theta}_X|\mathbf{W}_X \sim \mathcal{W}(\nu_X, \boldsymbol{I}_R) \\ \mathbf{X}_C|\mathbf{W}_X &\sim \mathbb{P}_{k_x}(\cdot\,|\mathbf{W}_X), \quad \mathrm{vec}(\mathbf{X}_M)|\boldsymbol{\Theta}_X \sim \mathcal{N}\left(\mathbf{0}, \left(\varepsilon \mathbf{U}_{[:R]} \boldsymbol{\Theta}_X \mathbf{U}_{[R]}^\top\right)^{-1} \otimes \boldsymbol{I}_p\right) \end{aligned}$$

where $\mathbf{U}_R$ are the eigenvectors associated to the Laplacian null-space of $\overline{\mathbf{W}}_X$. Given a graph $\mathbf{W}_X$, the idea is to structure the CCs' relative positions with a full-rank Gaussian model. The same model is considered for $\mathbf{W}_Z$, $\boldsymbol{\Theta}_Z$ and $\mathbf{Z}$, choosing $\nu_Z = \nu_X + p - q$ for the Wishart prior to satisfy the assumption of theorem 4.1. With this in place, we aim at providing a complete coupling objective, matching the pairs $(\mathbf{W}_X, \boldsymbol{\Theta}_X)$ and $(\mathbf{W}_Z, \boldsymbol{\Theta}_Z)$. The joint negative cross-entropy can be decomposed as follows:

$$\begin{aligned} &\mathbb{E}_{(\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{X}} \left[\log \mathbb{P}((\mathbf{W}_Z, \boldsymbol{\Theta}_Z) = (\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{Z})\right] \\ &= \mathbb{E}_{\mathbf{W}_X|\mathbf{X}} \left[\log \mathbb{P}(\mathbf{W}_Z = \mathbf{W}_X|\mathbf{Z})\right] + \qquad (6.12) \\ &\mathbb{E}_{(\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X|\mathbf{W}_Z = \mathbf{W}_X, \mathbf{Z})\right] \qquad (6.13) \end{aligned}$$

where (6.12) is the usual coupling criterion of $\mathbf{W}_X$ and $\mathbf{W}_Z$ capturing intra-CC variability while (6.13) is a penalty resulting from the Gaussian structure on $\mathcal{S}_M$. Constructed as such, the above objective allows a trade-off between local and global structure preservation. Following current trends in DR Kobak and Linderman [2021], we propose to take care of the global structure first *i.e.* focusing on (6.13) before (6.12). The difficulty of dealing with (6.13) lies in the hierarchical construction of the graph and the Gaussian precision (see **??**). We state the following result.

Corollary 6.1. Let $\mathbf{W}_X \in \mathcal{S}_W$, $\boldsymbol{L} = L(\overline{\mathbf{W}}_X)$ and $\mathcal{S}_M^q = (\ker \boldsymbol{L}) \otimes \mathbb{R}^q$, then for all $\varepsilon > 0$, given the above hierarchical model, the solution of the problem:

$$\min_{\mathbf{Z} \in \mathcal{S}_M^q} -\mathbb{E}_{\boldsymbol{\Theta}_X|\mathbf{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X|\mathbf{W}_Z = \mathbf{W}_X, \mathbf{Z})\right]$$

is a PCA embedding of $\mathbf{U}_{[:R]}\mathbf{U}_{[R]}^\top\mathbf{X}$ where $\mathbf{U}_{[:R]}$ are the CCs' membership vectors of $\overline{\mathbf{W}}_X$.

Remark 6.1. Note that while (6.12) approximates the objective of SNE-like methods when $\varepsilon \to 0$, the minimizer of (6.13) given by corollary 6.1 is stable for all ε.

From this observation, we propose a simple heuristic to minimize (6.13) that consists in computing a PCA embedding of $\mathbb{E}_{\mathbb{P}_{\mathcal{P}_X}(\cdot;\mathbf{K}_X)}\left[\mathbf{U}_{[:R]}\mathbf{U}_{[R]}^\top\right]\mathbf{X}$. The distribution of the connected components of the posterior of $\mathbf{W}_X$ being intractable, we resort to a Monte-Carlo estimation of the above expectation. The latter procedure called *ccPCA* aims at recovering the inter-CC structure that is filtered by SNE-like methods. *ccPCA* may then be used as initialization for optimizing (6.12) which is done by running the DR method corresponding to the graph priors at hand (section 4.4.1). This second step essentially consists in refining the intra-CC structure.

Experiments with *ccPCA*

?? shows that a t-SNE embedding of a balanced MNIST dataset of 10000 samples Deng [2012] with isotropic Gaussian initialization performs poorly in conserving the relative positions of clusters. As each digit cluster contains approximately 1000 points, with a perplexity of 30, sampling an edge across digit clusters in the graph posterior $\mathbb{P}_{\mathcal{P}_X}(\cdot;\mathbf{K}_X)$ is very unlikely. Recall that the perplexity value van der Maaten and Hinton [2008] corresponds to the approximate number of effective neighbors of each point. Hence images of different digits are with very high probability in different CCs of the graph posterior and their CC-wise means are not coupled as discussed in section 4.4.2. To remedy this in practice, PCA or Laplacian eigenmaps are usually used as initialization Kobak and Linderman [2021].

These strategies are tested (**??**) together with *ccPCA*. This shows that *ccPCA* manages to retrieve the digits that mostly support the large-scale variability as measured by the peripheral positioning of digits 0 (blue), 2 (green), 6 (pink) and 7 (grey) given by the right side of **??**. Other perplexity values for *ccPCA* are explored in appendix **??** while the experimental setup is detailed in appendix **??**. In appendix **??**, we perform quantitative evaluations of *ccPCA* for both t-SNE and UMAP on various datasets using K-ary neighborhood criteria. We find that using *ccPCA* as initialization is in general more reliable than PCA and Laplacian eigenmaps for preserving global structure using both t-SNE and UMAP.

Compared to PCA, *ccPCA* manages to aggregate points into clusters, thus filtering the intra-cluster variability and focusing solely on the inter-cluster structure. Compared to Laplacian eigenmaps which perform well at identifying clusters but suffer from the same deficiency as t-SNE for positioning them, *ccPCA* retains more of the coarse-grain structure. These observations support our unifying probabilistic framework and the

theoretical results about the MRF degeneracy which are the leading contributions of this article. The *ccPCA* initialization appears as a first stepping stone towards more grounded DR methods based on the probabilistic model presented in this article.

6.2 Appendix of Chapter 3

References

Akshay Agrawal, Alnur Ali, Stephen Boyd, et al. Minimum-distortion embedding. *Foundations and Trends® in Machine Learning*, 14(3):211–378, 2021.

David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

Friedrich Anders, Cristina Chiappini, Basilio Xavier Santiago, Gal Matijevič, Anna B Queiroz, Matthias Steinmetz, and Guillaume Guiglion. Dissecting stellar chemical abundance space with t-sne. *Astronomy & Astrophysics*, 619:A125, 2018.

Farzana Anowar, Samira Sadaoui, and Bassant Selim. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, 40:100378, 2021.

Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-sne algorithm for data visualization. In *Conference On Learning Theory*, pages 1455–1462. PMLR, 2018.

Mukund Balasubramanian, Eric L Schwartz, Joshua B Tenenbaum, Vin de Silva, and John C Langford. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

Mario Beauchemin. On affinity matrix normalization for graph cuts and spectral clustering. *Pattern Recognition Letters*, 68:90–96, 2015.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48 (3):334–334, 1997.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA, 1997.

Rajendra Bhatia. Infinitely divisible matrices. *The American Mathematical Monthly*, 113(3): 221–235, 2006.

Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. A unifying perspective on neighbor embeddings along the attraction-repulsion spectrum. *arXiv preprint arXiv:2007.08902*, 2020.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

T. Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research (JMLR)*, 23(301):1–54, 2022. URL http://jmlr.org/papers/v23/21-0524.html.

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature communications*, 12(1):124, 2021.

Miguel A Carreira-Perpinán. The elastic embedding algorithm for dimensionality reduction. In *ICML*, volume 10, pages 167–174. Citeseer, 2010.

Eranda Cela. *The quadratic assignment problem: theory and algorithms*, volume 1. Springer Science & Business Media, 2013.

Ines Chami, Albert Gu, Dat Nguyen, and Christopher Ré. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1419–1429. PMLR, 2021.

David M Chan, Roshan Rao, Forrest Huang, and John F Canny. t-sne-cuda: Gpu-accelerated t-sne and its applications to modern data. In *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 330–338. IEEE, 2018.

Yifan Chen, Rentian Yao, Yun Yang, and Jie Chen. A gromov–wasserstein geometric view of spectrum-preserving graph coarsening. *arXiv preprint arXiv:2306.08854*, 2023.

Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

Samir Chowdhury and Tom Needham. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021.

F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997a.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997b.

Peter Clifford. Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32, 1990.

Andy Coenen and Adam Pearce. Understanding umap. *Google PAIR*, 2019.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Neural Information Processing Systems (NeurIPS)*, 26, 2013.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020a. doi: 10.1101/2020.04.28.066787. URL https://www.biorxiv.org/content/early/2020/11/11/2020.04.28.066787.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, pages 2020–04, 2020b.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11): 1944–1957, 2007.

Tianjiao Ding, Derek Lim, Rene Vidal, and Benjamin D Haeffele. Understanding doubly stochastic clustering. In *International Conference on Machine Learning (ICML)*, 2022.

David L Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

Danielle Ezuz, Justin Solomon, Vladimir G Kim, and Mirela Ben-Chen. Gwcnn: A metric alignment layer for deep shape analysis. In *Computer Graphics Forum*, volume 36, pages 49–57. Wiley Online Library, 2017.

Xiran Fan, Chun-Hao Yang, and Baba C. Vemuri. Nested hyperbolic spaces for dimensionality reduction and hyperbolic nn design. In *Conference on Computer Vision and Pattern Recognition*, pages 356–365, June 2022.

Bruno César Feltes, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Márcio Dorn. Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 26(4):376–386, 2019. doi: 10.1089/cmb.2018.0238. URL https://doi.org/10.1089/cmb.2018.0238. PMID: 30789283.

Samaria Ferdinando and Harter Andy. Parameterisation of a stochastic model for human face identification, 1994.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2681–2690. PMLR, 2019.

Rémi Flamary, Cédric Févotte, Nicolas Courty, and Valentin Emiya. Optimal spectral transportation with application to music transcription. *Neural Information Processing Systems (NeurIPS)*, 29, 2016.

Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. 1996.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Unified framework for spectral dimensionality reduction, maximum variance unfolding, and kernel learning by semidefinite programming: Tutorial and survey. *arXiv preprint arXiv:2106.15379*, 2021.

Daniel B Graham and Nigel M Allinson. Characterising virtual eigensignatures for general purpose face recognition. *Face recognition: from theory to applications*, pages 446–456, 1998.

Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

Y. Guo, H. Guo, and S. X. Yu. CO-SNE: Dimensionality reduction and visualization for hyperbolic data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, Los Alamitos, CA, USA, jun 2022.

Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47, 2004.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Unsupervised Learning*. Springer New York, New York, NY, 2009.

Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer, 2002.

Geoffrey E Hinton and Sam T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 857–864. MIT Press, 2003. URL http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.pdf.

Martin Idel. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Philip A Knight, Daniel Ruiz, and Bora Uçar. A symmetry preserving algorithm for matrix scaling. *SIAM journal on Matrix Analysis and Applications*, 35(3):931–955, 2014.

Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14, 2019.

Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature biotechnology*, 39(2):156–157, 2021.

Dmitry Kobak, George Linderman, Stefan Steinerberger, Yuval Kluger, and Philipp Berens. Heavy-tailed kernels reveal a finer cluster structure in t-sne visualisations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 124–139. Springer, 2019.

Dmitry Kobak, George Linderman, Stefan Steinerberger, Yuval Kluger, and Philipp Berens. Heavy-tailed kernels reveal a finer cluster structure in t-sne visualisations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 124–139. Springer, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Joseph B Kruskal. *Multidimensional scaling*. Number 11. Sage, 1978.

Boris Landa, Ronald R Coifman, and Yuval Kluger. Doubly stochastic normalization of the gaussian kernel is robust to heteroskedastic noise. *SIAM journal on mathematics of data science*, 3(1):388–413, 2021.

John A Lee, Diego H Peluffo-Ordóñez, and Michel Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

Tianxi Li, Cheng Qian, Elizaveta Levina, and Ji Zhu. High-dimensional gaussian graphical models on network-linked data. *J. Mach. Learn. Res.*, 21:74–1, 2020.

Wentian Li, Jane E Cerise, Yaning Yang, and Henry Han. Application of t-sne to human genetic data. *Journal of bioinformatics and computational biology*, 15(04):1750017, 2017.

Derek Lim, René Vidal, and Benjamin D Haeffele. Doubly stochastic subspace clustering. *arXiv preprint arXiv:2011.14859*, 2020.

Ya-Wei Eileen Lin, Ronald R. Coifman, Gal Mishne, and Ronen Talmon. Hyperbolic diffusion embedding and distance for hierarchical representation learning. In *International Conference on Machine Learning*, 2023.

George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.

George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245, 2019.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, Xiang Zhou, Xingjie Shi, and Jin Liu. Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic acids research*, 50(12):e72–e72, 2022a.

Weijie Liu, Jiahao Xie, Chao Zhang, Makoto Yamada, Nenggan Zheng, and Hui Qian. Robust graph dictionary learning. In *The Eleventh International Conference on Learning Representations*, 2022b.

Yao Lu, Jukka Corander, and Zhirong Yang. Doubly stochastic neighbor embedding on spheres. *Pattern Recognition Letters*, 128:100–106, 2019.

Hanbaek Lyu and Yuchen Li. Block majorization-minimization with diminishing radius for constrained nonconvex optimization. 08 2023.

Haggai Maron and Yaron Lipman. (probably) concave graph matching. *Advances in Neural Information Processing Systems*, 31, 2018.

Nicholas F Marshall and Ronald R Coifman. Manifold learning with bi-stochastic kernels. *IMA Journal of Applied Mathematics*, 84(3):455–482, 2019.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Binu Melit Devassy, Sony George, and Peter Nussbaum. Unsupervised clustering of hyperspectral paper data using t-sne. *Journal of Imaging*, 6(5):29, 2020.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Facundo Mémoli and Tom Needham. Gromov-monge quasi-metrics and distance distributions. *arXiv*, 2018, 2018.

Facundo Mémoli and Tom Needham. Comparison results for gromov-wasserstein and gromov-monge distances. *arXiv preprint arXiv:2212.14123*, 2022.

Peyman Milanfar. Symmetrizing smoothing filters. *SIAM Journal on Imaging Sciences*, 6(1): 263–284, 2013. doi: 10.1137/120875843.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Nicola Pezzotti, Julian Thijssen, Alexander Mordvintsev, Thomas Höllt, Baldur Van Lew, Boudewijn PF Lelieveldt, Elmar Eisemann, and Anna Vilanova. Gpgpu linear complexity t-sne optimization. *IEEE transactions on visualization and computer graphics*, 26(1):1172–1181, 2019.

Huy Phan. Pytorch models trained on cifar-10 dataset. https://github.com/huyvnphan/PyTorch_CIFAR10, 2021.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE international conference on image processing (ICIP)*, pages 4852–4856. IEEE, 2014.

Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. Co-optimal transport. *Advances in Neural Information Processing Systems*, 33(17559-17570):2, 2020.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63): 94, 2015.

Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

Satu Elisa Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pages 287–297, 2016.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Alexis Thual, Quang Huy TRAN, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in Neural Information Processing Systems*, 35:21792–21804, 2022.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109:475–494, 2001.

Hugues Van Assel, Thibault Espinasse, Julien Chiquet, and Franck Picard. A probabilistic graph coupling view of dimension reduction. *Neural Information Processing Systems (NeurIPS)*, 2022.

Hugues Van Assel, Titouan Vayer, Rémi Flamary, and Nicolas Courty. SNEkhorn: Dimension reduction with symmetric entropic affinities. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Laurens Van Der Maaten. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*, 2013.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008.

Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71), 2009.

L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(nov):2579–2605, 2008. ISSN 1532-4435. Pagination: 27.

David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks—ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings 11*, pages 485–491. Springer, 2001.

Elias Ventre, Ulysse Herbach, Thibault Espinasse, Gérard Benoit, and Olivier Gandrillon. One model fits all: combining inference and simulation of gene regulatory networks. *PLoS Computational Biology*, 19(3):e1010962, 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International conference on machine learning*, pages 10564–10574. PMLR, 2021.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*, 2022a.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Template based graph neural network with optimal transport distances. *Advances in Neural Information Processing Systems*, 35:11800–11814, 2022b.

Max Vladymyrov and Miguel Carreira-Perpinan. Entropic affinities: Properties and efficient numerical computation. In *International conference on machine learning*, pages 477–485. PMLR, 2013.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *J Mach. Learn. Res*, 22:1–73, 2021.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1 (10):e2, 2016.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

Hongtengl Xu. Gromov-wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6478–6485, 2020.

Ron Zass and Amnon Shashua. A unifying approach to hard and probabilistic clustering. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 294–301. IEEE, 2005.

Ron Zass and Amnon Shashua. Doubly stochastic normalization for spectral clustering. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Neural Information Processing Systems (NeurIPS)*. MIT Press, 2006.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in neural information processing systems*, 17, 2004.

Zhichen Zeng, Ruike Zhu, Yinglong Xia, Hanqing Zeng, and Hanghang Tong. Generative graph dictionary learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40749–40769. PMLR, 23–29 Jul 2023.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Jason Ramapuram, Dan Busbridge, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. σreparam: Stable transformer training with spectral reparametrization. 2023.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Neural Information Processing Systems (NeurIPS)*, 16, 2003.