Acknowledgements

Merci le Z.

Chapter 1

A Probabilistic Graph Coupling View of Dimension Reduction

1.1 Introduction

Dimensionality reduction (DR) is of central importance when dealing with high-dimensional data donoho2000high. 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) [**?**], multidimensional scaling [**?**] and more broadly kernel eigenmaps methods such as Isomap [**?**], Laplacian eigenmaps belkin2003laplacian and diffusion maps coifman2006diffusion. 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 ham2004kernel.

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 [**?**], its heavy-tailed symmetrized version t-SNE [**?**] or more recent approaches like LargeVis [**?**] and UMAP [**?**] 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 as the state-of-art techniques in many fields [**?**, **?**, **?**]. Their popularity is mainly due to their exceptional ability to preserve local structure, *i.e.* close points in the input space have close embeddings, as shown empirically [**?**]. They also demonstrate impressive performances in identifying clusters [**?**, **?**]. However this is done at the expense of global structure, that these methods struggle in preserving [**?**, **?**] *i.e.* the relative large-scale distances between embedded points do not necessarily correspond to the original ones.

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 [**?**, **?**, **?**] 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 (**??**). We demonstrate that such graphs essentially capture conditional independencies among rows through a pairwise Markov Random Field (MRF) model which construction can be found in **??**.
- 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 a degeneracy of the MRF when shift invariant kernels are used (**??**). Such degeneracy induces the loss of the relative positions of clusters corresponding to the connected components of the posterior latent graphs which distributions are identified (**??**). These findings are highlighted by a new initialization of the embeddings that is empirically tested (**??**).
- 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 (**??**). Such model does not suffer from the aforementioned degeneracy hence preserves the large-scale structure.

1.2 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.

1.2.1 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 1. The graph Laplacian operator is the map $L\colon \mathbb{R}_+^{n\times n} \to \mathcal{S}_+^n(\mathbb{R})$ 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 $\boldsymbol{L} = L(\overline{\mathbf{W}})$ 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 [**?**], the null space of $\boldsymbol{L}$ is spanned by the orthonormal vectors $\{\boldsymbol{U}_r\}_{r\in[R]}$ such that for $r \in [R]$, $\boldsymbol{U}_r = \left(n_r^{-1/2} \mathbb{1}_{i\in C_r}\right)_{i\in[\![n]\!]}$ with $n_r = \text{Card}(C_r)$. By the spectral theorem, $\boldsymbol{U}_{[R]}$ can be completed such that $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{U} = (\boldsymbol{U}_1, ..., \boldsymbol{U}_n)$ is orthogonal and $\boldsymbol{\Lambda} = \text{diag}((\lambda_i)_{i\in[\![n]\!]})$ with $0 = \lambda_1 = ... = \lambda_R < \lambda_{R+1} \leq ... \leq \lambda_n$. In the following, the data is split

into two parts: $\mathbf{X}_M$, the orthogonal projection of $\mathbf{X}$ on $\mathcal{S}_M = (\ker \boldsymbol{L}) \otimes \mathbb{R}^p$, and $\mathbf{X}_C$, the projection on $\mathcal{S}_C = (\ker \boldsymbol{L})^\perp \otimes \mathbb{R}^p$. 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.

1.2.2 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 $k : \mathbb{R}^p \to \mathbb{R}_+$. We consider the following pairwise MRF unnormalized density function:

$$f_k \colon (\mathbf{X}, \mathbf{W}) \mapsto \prod_{(i,j)\in[\![n]\!]^2} k(\mathbf{X}_i - \mathbf{X}_j)^{W_{ij}} \,. \tag{1.1}$$

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

$$\log f_k(\mathbf{X}, \mathbf{W}) = -\sum_{(i,j)\in[\![n]\!]^2} W_{ij} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2_{\boldsymbol{\Sigma}} = -\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}\right) \tag{1.2}$$

by property of the graph Laplacian (**??**). In this case, it is clear that due to the rank deficiency of $\boldsymbol{L}$, $f_k(\cdot, \mathbf{W})$ 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 [**?**]. Heavy-tailed kernels appear useful when the dimension of the embeddings is smaller than the intrinsic dimension of the data [**?**]. Our contribution provides flexibility by extending the previous result to a large class of kernels, as stated in the following theorem.

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

$$\mathbb{P}_k(d\mathbf{X}_C|\mathbf{W}) = \mathcal{C}_k(\mathbf{W})^{-1} f_k(\mathbf{X}_C, \mathbf{W}) \lambda_{\mathcal{S}_C}(d\mathbf{X}_C) \,. \tag{1.3}$$

Remark 3. 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 **??** 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 (**??**) boils down to a pairwise MRF model clifford1990markov 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 **??**) 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 $\overline{\mathbf{W}}$.

1.2.3 Uninformative Model for CC-wise Means

We showed that the MRF (**??**) 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.* [**?**] 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 $f^\varepsilon(\cdot,\mathbf{W})[\varepsilon\to 0]1$ almost everywhere. We now define the conditional distribution on $(\mathcal{S}_M,\mathcal{B}(\mathcal{S}_M))$ as follows:

$$\mathbb{P}^\varepsilon(d\mathbf{X}_M|\mathbf{W})=\mathcal{C}^\varepsilon(\mathbf{W})^{-1}f^\varepsilon(\mathbf{X}_M,\mathbf{W})\lambda_{\mathcal{S}_M}(d\mathbf{X}_M) \tag{1.4}$$

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 (**??**) and (**??**) 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. 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 loosing 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 $f^\varepsilon:\mathbf{X}\to\exp\left(-\frac{\varepsilon}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{X}^\top\boldsymbol{U}_{[:R]}\boldsymbol{\Theta}\boldsymbol{U}_{[R]}^\top\mathbf{X}\right)\right)$ such that: $\operatorname{vec}(\mathbf{X}_M)|\boldsymbol{\Theta}\sim\mathcal{N}\left(\mathbf{0},\left(\varepsilon\boldsymbol{U}_{[:R]}\boldsymbol{\Theta}\boldsymbol{U}_{[R]}^\top\right)^{-1}\otimes\boldsymbol{\Sigma}\right)$ where $\otimes$ denotes the Kronecker product.

1.3 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 non informative 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 [**?**], 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 constraint the topology of the graph. For a matrix $\boldsymbol{A}$, 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 [**?**].

Definition 5. Let $\boldsymbol{\pi}\in\mathbb{R}_+^{n\times n}$, $\varepsilon\in\mathbb{R}_+$, $\alpha\in\mathbb{R}$, k satisfies the assumptions of **??** and $\mathcal{P}\in\{B,D,E\}$. For $\mathbf{W}\in\mathcal{S}_W$ we introduce:

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

where $\Omega_B(\mathbf{W})=\prod_{ij}\mathbb{1}_{W_{ij}\le 1}$, $\Omega_D(\mathbf{W})=\prod_i\mathbb{1}_{W_{i+}=1}$ 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 $\mathbb{P}_{\mathcal{P}}(\mathbf{W};\boldsymbol{\pi})=\mathbb{P}_{\mathcal{P},k}^\varepsilon(\boldsymbol{W};\boldsymbol{\pi},0)$. 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 $\mathcal{P}=B$, $\forall(i,j)\in[\![n]\!]^2$, $W_{ij}\overset{\perp}{\sim}\mathcal{B}\left(\pi_{ij}/(1+\pi_{ij})\right)$.

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

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 **??**). In the following, $\odot$ stands for the Hadamard product and $\mathcal{D}$ for the convergence in distribution.

Proposition 6. *Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n \times n}$, k satisfies the assumptions of **??** with $\boldsymbol{K}_X = (k(\mathbf{X}_i - \mathbf{X}_j))_{(i,j)\in[\![n]\!]^2}$ and $\mathcal{P} \in \{B, D, E\}$. If $\mathbf{W}^\varepsilon \sim \mathbb{P}^\varepsilon_{\mathcal{P},k}(\cdot\,; \boldsymbol{\pi}, 1)$ then*

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

Remark 7. For all $\mathbf{W} \in \mathcal{S}_W$, $\mathcal{C}^\varepsilon(\mathbf{W})$ diverges as $\varepsilon \to 0$, hence the graph prior (**??**) is improper at the limit. This compensates for the uninformative diffuse conditional and allows to retrieve a well-defined tractable posterior limit.

1.3.1 Retrieving Well Known 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 **??** and we denote by $\boldsymbol{K}_X$ and $\boldsymbol{K}_Z$ the associated kernel matrices on $\mathbf{X}$ and 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 $\varepsilon \to 0$,

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

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 [**?**], where for $i \in [\![n]\!]$, $\boldsymbol{P}_i^D = \boldsymbol{K}_{X,i}/K_{X,i+}$ and $\boldsymbol{Q}_i^D = \boldsymbol{K}_{Z,i}/K_{Z,i+}$,

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

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

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

and the symmetrized objective of t-SNE [**?**] is recovered.

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

$$\mathcal{H}_{D,B} = -\sum_{i\neq j} P_{ij}^D \log Q_{ij}^B + \left(1 - P_{ij}^D\right) \log\left(1 - Q_{ij}^B\right) = -\sum_{i<j} \overline{P}_{ij}^D \log Q_{ij}^B + \left(2 - \overline{P}_{ij}^D\right) \log(1 - Q_{ij}^B)$$

which is the objective of LargeVis [**?**].

Table 1.1: 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

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, $\widetilde{W}_{X,ij} \sim \mathcal{B}\left(\widetilde{P}_{ij}^B\right)$ where $\widetilde{P}_{ij}^B = P_{ij}^B + P_{ji}^B - P_{ij}^B P_{ji}^B$ and $\boldsymbol{P}^B = \left(K_{X,ij}/(1+K_{X,ij})\right)_{(i,j)\in[\![n]\!]^2}$. Coupling $\widetilde{\mathbf{W}}_X$ and $\mathbf{W}_Z$ gives:

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

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

Remark 8. One can also consider $\mathcal{H}_{E,E}$ but as detailed in [**?**], 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.

1.3.2 Interpretations

As we have seen in **??**, 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 **??** and **??**. 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:

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times q}} - \sum_{(i,j)\in[\![n]\!]^2} \boldsymbol{P}_{ij}^{\mathcal{P}_X} \log k_z(\mathbf{Z}_i - \mathbf{Z}_j) + \log \mathbb{P}(\mathbf{Z}). \tag{1.5}$$

The first and second terms in **??** respectively summarize the attractive and repulsive forces of the objective. Recall from **??** that $\boldsymbol{P}^{\mathcal{P}_X}$ 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 $\boldsymbol{L}^\star = \mathbb{E}_{\mathbf{W}\sim\mathbb{P}_{\mathcal{P}_X}(\cdot;\boldsymbol{K}_X)}[L(\mathbf{W})]$, boiling down to the objective of Laplacian eigenmaps [**?**]. Therefore, for Gaussian MRFs, the attractive forces resume to an unconstrained Laplacian eigenmaps objective. Such link, already noted in [**?**], 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 (**??**). 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 [**?**] 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 **??**, 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 **??**. In the following section we propose to mitigate the global structure deficiency with non-degenerate MRF models.

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.

1.4.1 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 (**??**), 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 9. *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:*

$$\operatorname{vec}(\mathbf{X})|\mathbf{\Theta}_X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_X^{-1} \otimes \boldsymbol{I}_p), \tag{1.6}$$

$$\operatorname{vec}(\mathbf{Z})|\mathbf{\Theta}_Z \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q)\,. \tag{1.7}$$

Then the solution of 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 [**?**]. 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 **??**). Recall that the later always has a null-space which is spanned by the CC indicator vectors of $\mathbf{W}$ (**??**). 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 (**??**) and (**??**), as previously done when constructing the prior for neighbor embedding methods (definition **??**). 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 [**?**], 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 is 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 [**?**]. Therefore, as suspected in **??**, the degeneracy of the conditional distribution given the graph is key to determine the distance preservation properties of the embeddings. We propose in **??** to combine both graph coupling approaches to strike a balance between global and local structure preservation.

1.4.2 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 **??** and $\nu_X \geq n$:

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

where $\boldsymbol{U}_{[R]}$ 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 **??**. 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:

$$\begin{aligned} &\mathbb{E}_{(\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{X}} \left[\log \mathbb{P}((\mathbf{W}_Z, \boldsymbol{\Theta}_Z) = (\mathbf{W}_X, \boldsymbol{\Theta}_X) | \mathbf{Z})\right] \\ &= \mathbb{E}_{\mathbf{W}_X|\mathbf{X}} \left[\log \mathbb{P}(\mathbf{W}_Z = \mathbf{W}_X | \mathbf{Z})\right] + \qquad (1.8) \\ &\mathbb{E}_{(\mathbf{W}_X, \boldsymbol{\Theta}_X)|\mathbf{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X | \mathbf{W}_Z = \mathbf{W}_X, \mathbf{Z})\right] \qquad (1.9) \end{aligned}$$

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

Corollary 10. *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 $\boldsymbol{U}_{[:R]} \boldsymbol{U}_{[R]}^\top \mathbf{X}$ where $\boldsymbol{U}_{[:R]}$ are the CCs' membership vectors of $\overline{\mathbf{W}}_X$.

Remark 11. Note that while (**??**) approximates the objective of SNE-like methods when $\varepsilon \to 0$, the minimizer of (**??**) given by **??** is stable for all ε.

From this observation, we propose a simple heuristic to minimize (**??**) that consists in computing a PCA embedding of $\mathbb{E}_{\mathbb{P}_{\mathcal{P}_X}(\cdot; \boldsymbol{K}_X)} \left[\boldsymbol{U}_{[:R]} \boldsymbol{U}_{[R]}^\top\right] \mathbf{X}$. 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.

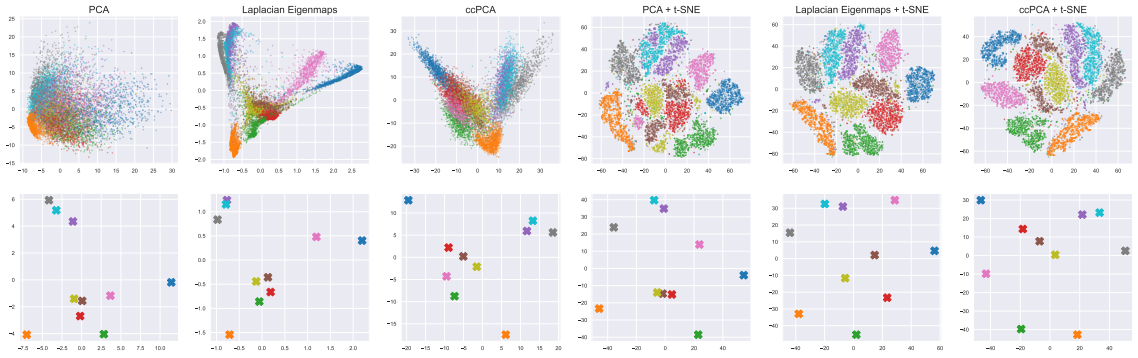


Figure 1.2: Top: MNIST embeddings produced by PCA, Laplacian eigenmaps, *ccPCA* and finally t-SNE launched after the previous three embeddings to improve the fine-grain structure. Bottom: mean coordinates for each digit using the embeddings of the first row. The color legend is the same as in **??**. t-SNE was trained during 1000 iterations using default parameters with the openTSNE implementation [**?**].

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 (**??**) which is done by running the DR method corresponding to the graph priors at hand (**??**). This second step essentially consists in refining the intra-CC structure.

1.4.3 Experiments with *ccPCA*

?? shows that a t-SNE embedding of a balanced MNIST dataset of 10000 samples [**?**] 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 $\mathbb{P}_{\mathcal{P}_X}(\cdot; \boldsymbol{K}_X)$ is very unlikely. Recall that the perplexity value [**?**] 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 **??**. To remedy this in practice, PCA or Laplacian eigenmaps are usually used as initialization [**?**].

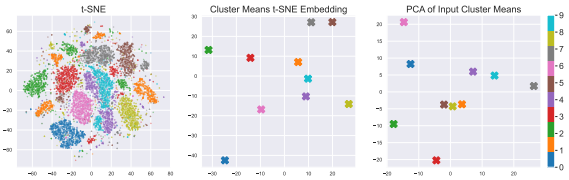


Figure 1.1: Left: MNIST t-SNE (perp : 30) embeddings initialized with i.i.d $\mathcal{N}(0, 1)$ coordinates. Middle: using these t-SNE embeddings, mean coordinates for each digit are represented. Right: we compute a matrix of mean input coordinates for each of the 10 digits and embed it using PCA. For t-SNE embeddings, the positions of clusters vary accross different runs and don't visually match the PCA embeddings of input mean vectors (right plot).

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 suffers 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.

1.5 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.

Chapter 2

SNEkhorn: Dimension Reduction with Symmetric Entropic Affinities

2.1 Introduction

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 [**?**], spectral clustering [**?**] and semi-supervised learning [**?**]. 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 [**?**]. 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.

Entropic Affinities and SNE/t-SNE. Entropic affinities (EAs) were first introduced in the seminal paper *Stochastic Neighbor Embedding* (SNE) [**?**]. 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 [**?**]. EAs are notoriously used to encode pairwise relations in a high-dimensional space for the DR algorithm t-SNE [**?**], among other DR methods including [**?**]. t-SNE is increasingly popular in many applied fields [**?**, **?**] mostly due to its ability to represent clusters in the data [**?**, **?**]. Nonetheless, one major flaw of EAs is that they are inherently directed and often require post-processing symmetrization.

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 [**?**] and it is well-established that it enhances spectral clustering performances [**?**, **?**, **?**]. Additionally, DS matrices present the benefit of being invariant to the various Laplacian normalizations [**?**]. 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 [**?**]. It also offers a more natural analog to the heat

kernel [**?**]. These properties have led to a growing interest in DS affinities, with their use expanding to various applications such as smoothing filters [**?**], subspace clustering [**?**] and transformers [**?**].

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) (**??**). We then propose in **??** 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 **??**, 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 [**?**]. We finally showcase the benefits of symmetric entropic affinities on a variety of applications in Section **??** including spectral clustering and DR experiments on datasets ranging from images to genomics data.

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 $\in \mathbb{R}^n_+$ is[1] $\mathrm{H}() = -\sum_i p_i(\log(p_i) - 1) = -\langle , \log - \mathbf{1} \rangle$. The Kullback-Leibler divergence between two matrices $\mathbf{P}, \mathbf{Q}$ with nonnegative entries such that $Q_{ij} = 0 \implies P_{ij} = 0$ is $\mathrm{KL}(\mathbf{P}|\mathbf{Q}) = \sum_{ij} P_{ij} \left(\log(\frac{P_{ij}}{Q_{ij}}) - 1 \right) = \langle \mathbf{P}, \log(\mathbf{P} \oslash \mathbf{Q}) - \mathbf{1}\mathbf{1}^\top \rangle$.

2.2 Entropic Affinities, Dimensionality Reduction and Optimal Transport

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 . This is generally achieved by optimizing $\mathbf{Z}$ such that the corresponding affinity matrix matches another affinity matrix defined from . 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* $\mathbf{C} \in \mathcal{D} := \{ \in \mathbb{R}_+^{n \times n} :=^\top$ 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 $\exp(-/\varepsilon)$, where $\varepsilon > 0$ is the bandwidth parameter.

Entropic Affinities (EAs). Another frequently used approach to generate affinities from $\mathbf{C} \in \mathcal{D}$ is to employ *entropic affinities* [**?**]. 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 [**?**]. 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 [**?**]. 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

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

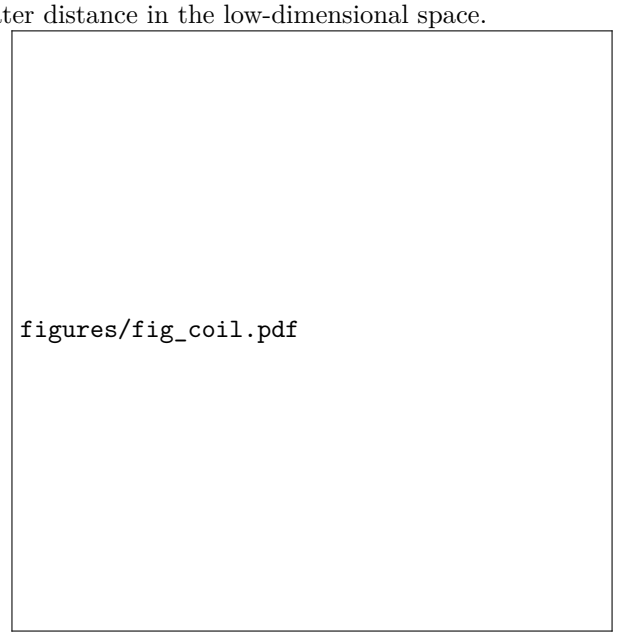
uniform distribution over ξ neighbors. Therefore, it provides the practitioner with an interpretable parameter specifying which scale of dependencies the affinity matrix should faithfully capture. In practice, a root-finding algorithm is used to find the bandwidth parameters $(\varepsilon_i^\star)_{i\in[\![n]\!]}$ that satisfy the constraints [**?**]. 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` [**?**]. We denote by $= \big(\|_{i:} - _{j:}\|_2^2\big)_{ij}$ and $= \big(\|_{i:} - _{j:}\|_2^2\big)_{ij}$ the cost matrices derived from the rows (*i.e.* the samples) of and respectively. `SNE` focuses on minimizing in the latent coordinates $\in \mathbb{R}^{n\times q}$ the objective $\mathrm{KL}(\mathbf{P}^{\mathrm{e}}|\mathbf{Q}_)$ where $\mathbf{P}^{\mathrm{e}}$ solves (**??**) with cost and $[\mathbf{Q}_]_{ij} = \exp(-[_]_{ij})/(\sum_\ell \exp(-[_]_{i\ell}))$. In the seminal paper van2008visualizing, 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 $[\widetilde{\mathbf{Q}}_]_{ij} = \exp(-[_]_{ij})/\sum_{\ell,t}\exp(-[_]_{\ell t})$ it consists in solving

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

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 $\mathcal{S}$: $\overline{\mathbf{P}^{\mathrm{e}}} = \mathrm{Proj}_{\mathcal{S}}^{\ell_2}(\mathbf{P}^{\mathrm{e}}) =_{\mathbf{P}\in\mathcal{S}} \|\mathbf{P} - \mathbf{P}^{\mathrm{e}}\|_2$ (see Appendix **??**). Instead of the Gaussian kernel, the popular extension t-SNE van2008visualizing considers a different distribution in the latent space $[\widetilde{\mathbf{Q}}_]_{ij} = (1 + []_{ij})^{-1}/\sum_{\ell,t}(1 + []_{\ell t})^{-1}$. In this formulation, $\widetilde{\mathbf{Q}}$ 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.

Considering symmetric similarities is appealing since the proximity between two points is inherently symmetric. Nonetheless, the Euclidean projection in (**??**) *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 **??**). This is arguably one of the main drawbacks of the approach. By contrast, the $\mathbf{P}^{\mathrm{se}}$ affinity that will be introduced in **??** can accurately set the entropy in each point to the desired value $\log\xi+1$. As shown in **??** this leads to more faithful embeddings with higher silhouette scores when combined with the SNEkhorn algorithm (**??**).



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

Symmetric Entropy-Constrained Optimal Transport. Entropy-regularized OT [**?**] 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 $\mathbf{P} \mapsto \langle \mathbf{P}, \rangle - \nu\sum_i \mathrm{H}(\mathbf{P}_{i:})$ 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 $\mathbf{P}^{\mathrm{ds}} = \mathrm{diag}(\mathbf{u})\,\mathrm{diag}(\mathbf{v})$ where $= \exp(-/\nu)$ is the Gibbs energy derived from and $\mathbf{u}, \mathbf{v}$ are positive vectors that can be found with the celebrated Sinkhorn-Knopp's algorithm [**?**, **?**]. Interestingly, when the cost is *symmetric* (*e.g.* $\in \mathcal{D}$) we can

take $\mathbf{u} = \mathbf{v}$ [**?**, Section 5.2] so that the unique optimal solution is itself symmetric and writes

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

In this case, by relying on convex duality as detailed in Appendix **??**, 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 **??**). This constrained formulation of symmetric entropic OT will provide new insights into entropic affinities, as detailed in the next sections.

2.3 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.

2.3.1 Entropic Affinities as Entropic Optimal Transport

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

$$\mathcal{H}_\xi = \{\mathbf{P} \in \mathbb{R}_+^{n \times n} \text{ s.t. } \mathbf{P}\mathbf{1} = \mathbf{1} \text{ and } \forall i,\, \mathrm{H}(\mathbf{P}_{i:}) \geq \log \xi + 1\} \,. \tag{2.1}$$

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 (**??**), $\mathcal{H}_\xi$ 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). propositionentropicaffinityaslinearprogram Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ without constant rows. Then $\mathbf{P}^{\text{e}}$ solves the entropic affinity problem (**??**) with cost $\mathbf{C}$ if and only if $\mathbf{P}^{\text{e}}$ 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, (**??**) can be seen as a specific *semi-relaxed* OT problem [**?**, **?**] (*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 (**??**) has *saturated entropy i.e.* $\forall i$, $\mathrm{H}(\mathbf{P}^\star_{i:}) = \log \xi + 1$. In other words, relaxing the equality constraint in (**??**) 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 12. The kernel bandwidth parameter $\boldsymbol{\varepsilon}$ from the original formulation of entropic affinities (**??**) is the Lagrange dual variable associated with the entropy constraint in (**??**). Hence computing $\boldsymbol{\varepsilon}^\star$ in (**??**) exactly corresponds to solving the dual problem of (**??**).

Remark 13. Let $_\sigma = \exp(-\mathbf{C}/\sigma)$. As shown in **??**, if $\boldsymbol{\varepsilon}^\star$ solves (**??**) and $\sigma \leq \min(\boldsymbol{\varepsilon}^\star)$, then $\mathbf{P}^{\text{e}} = \mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(_\sigma) =_{\mathbf{P} \in \mathcal{H}_\xi} \mathrm{KL}(\mathbf{P}|_\sigma)$. Therefore $\mathbf{P}^{\text{e}}$ can be seen as a KL Bregman projection [**?**] of a Gaussian kernel onto $\mathcal{H}_\xi$. Hence the input matrix in (**??**) is $\overline{\mathbf{P}^{\text{e}}} = \mathrm{Proj}^{\ell_2}_{\mathcal{S}}(\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi}(_\sigma))$ which corresponds to a surprising mixture of KL and orthogonal projections.

2.3.2 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 (**??**), in a similar fashion as (**??**). More precisely we consider the convex optimization problem

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

where we recall that $\mathcal{S}$ is the set of $n\times n$ symmetric matrices. Note that for any $\xi\leq n-1$, $\frac{1}{n}\mathbf{1}\mathbf{1}^\top\in\mathcal{H}_\xi\mathcal{S}$ hence the set $\mathcal{H}_\xi\mathcal{S}$ is a non-empty and convex set. We first detail some important properties of problem (**??**) (the proofs of the following results can be found in Appendix **??**). [Saturation of the entropies]propositionsaturation Let $\mathbf{C}\in\mathcal{S}$ with zero diagonal, then (**??**) with cost $\mathbf{C}$ has a *unique solution* that we denote by $\mathbf{P}^{\text{se}}$. If moreover $\mathbf{C}\in\mathcal{D}$, then for at least $n-1$ indices $i\in[\![n]\!]$ the solution satisfies $\mathrm{H}(\mathbf{P}^{\text{se}}_{i:})=\log\xi+1$. In other words, the unique solution $\mathbf{P}^{\text{se}}$ 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 (**??**) allows accurate control over the point-wise entropies while providing a symmetric doubly stochastic matrix, unlike $\overline{\mathbf{P}^{\text{e}}}$ defined in (**??**), as summarized in **??**. In the sequel, we denote by $\mathrm{H_r}(\mathbf{P})=(\mathrm{H}(\mathbf{P}_{i:}))_i$ the vector of row-wise entropies of $\mathbf{P}$. We rely on the following result to compute $\mathbf{P}^{\text{se}}$. [Solving for **??**]propositionsolvingsea 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 $q(\boldsymbol{\gamma},\boldsymbol{\lambda})=\min_{\mathbf{P}\in\mathbb{R}_+^{n\times n}\mathcal{S}}\mathcal{L}(\mathbf{P},\boldsymbol{\gamma},\boldsymbol{\lambda})$. Strong duality holds for (**??**). Moreover, let $\boldsymbol{\gamma}^\star,\boldsymbol{\lambda}^\star\in\operatorname{argmax}_{\boldsymbol{\gamma}\geq 0,\boldsymbol{\lambda}} q(\boldsymbol{\gamma},\boldsymbol{\lambda})$ 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 $\mathrm{H_r}(\mathbf{P}^{\text{se}})=(\log\xi+1)\mathbf{1}$ and $\mathbf{P}^{\text{se}}$ has the form

$$\mathbf{P}^{\text{se}}=\exp\left((\boldsymbol{\lambda}^\star\oplus\boldsymbol{\lambda}^\star-2\mathbf{C})\oslash(\boldsymbol{\gamma}^\star\oplus\boldsymbol{\gamma}^\star)\right). \tag{2.2}$$

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 $\boldsymbol{\gamma}^\star>0$, $\mathbf{P}^{\text{se}}=\mathbf{P}(\boldsymbol{\gamma}^\star,\boldsymbol{\lambda}^\star)$ 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 $\mathbf{P}^{\text{se}}$ we solve the problem (**??**). Although the form of $\mathbf{P}^{\text{se}}$ presented in Proposition **??** is only valid when $\boldsymbol{\gamma}^\star$ is positive and we have only proved it for $n-1$ indices, we emphasize that if (**??**) has a finite solution, then it is equal to $\mathbf{P}^{\text{se}}$. Indeed in this case the solution satisfies the KKT system associated with (**??**).

TABLE PROPERTIES OF $\mathbf{P}^{\text{e}}$, $\overline{\mathbf{P}^{\text{e}}}$, $\mathbf{P}^{\text{ds}}$ AND $\mathbf{P}^{\text{se}}$

AFFINITY MATRIX	$\mathbf{P}^{\text{e}}$	$\overline{\mathbf{P}^{\text{e}}}$	$\mathbf{P}^{\text{ds}}$	$\mathbf{P}^{\text{se}}$
REFERENCE	[**?**]	VAN2008VISUALIZING	[**?**]	(**??**)
$\mathbf{P}=\mathbf{P}^\top$	×			
$\mathbf{P}\mathbf{1}=\mathbf{P}^\top\mathbf{1}=\mathbf{1}$	×	×		
$\mathrm{H_r}(\mathbf{P})=(\log\xi+1)\mathbf{1}$		×	×	

Numerical optimization. The dual problem (**??**) is concave and can be solved with guarantees through a dual ascent approach with closed-form gradients (using *e.g.* `SGD`, `BFGS` [**?**] or `ADAM` [**?**]). 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.* [**?**, 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 14. In the same spirit as **??**, one can express $\mathbf{P}^{\text{se}}$ as a KL projection of $_\sigma=\exp(-\mathbf{C}/\sigma)$. Indeed, we show in **??** that if $0<\sigma\leq\min_i\gamma_i^\star$, then $\mathbf{P}^{\text{se}}=\mathrm{Proj}^{\mathrm{KL}}_{\mathcal{H}_\xi\mathcal{S}}(_\sigma)$. 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 **??** 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 $\mathbf{P}^{\text{ds}}$ (**??**) (OT problem with global entropy constraint (**??**)) , 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 [**?**]). 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).

2.4 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 [**?**], is used to measure the discrepancy between the two affinities. As detailed in sections **??** and **??**, $\mathbf{P}^{\text{se}}$ corrects for heterogeneity in the data density by imposing point-wise entropy constraints. As we do not need such correction for embedding coordinates since they must be optimized, we opt for the standard affinity (**??**) built as an OT transport plan with global entropy constraint (**??**). This OT plan can be efficiently computed using Sinkhorn's algorithm. More precisely, we propose the optimization problem

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

where $\mathbf{Q}^{\text{ds}}_{\mathbf{Z}} = \exp\left(\mathbf{f}_{\mathbf{Z}} \oplus \mathbf{f}_{\mathbf{Z}} - \mathbf{Z}\right)$ stands for the (**??**) affinity computed with cost $\mathbf{Z}$ 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 [**?**] 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 and relying on the double stochasticity of $\mathbf{P}^{\text{se}}$, the objective in (**??**) can be expressed as $\langle \mathbf{P}^{\text{se}}, \mathbf{Z} \rangle - 2\langle \mathbf{f}_{\mathbf{Z}}, \mathbf{1}\rangle$.

Heavy-tailed kernel in latent space. Since it is well known that heavy-tailed kernels can be beneficial in DR [**?**], 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 $[\mathbf{Z}]_{ij} = \big(\log(1 + \|_{i:} - _{j:}\|_2^2)\big)_{ij}$ instead of $(\|_{i:} - _{j:}\|_2^2)_{ij}$.

Inference. This new DR objective involves computing a doubly stochastic normalization for each update of . Interestingly, to compute the optimal dual variable $\mathbf{f}_{\mathbf{Z}}$ in $\mathbf{Q}^{\text{ds}}_{\mathbf{Z}}$, we leverage a well-conditioned Sinkhorn fixed point iteration knight2014symmetry, feydy2019interpolating, 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 **??**, we plot $\|\mathbf{Q}^{\text{ds}}_{\mathbf{Z}}\mathbf{1} - \mathbf{1}\|_\infty$ as a function of (**??**) iterations for a toy example presented in **??**. In most practical cases, we found that about 10 iterations were enough to reach

sufficiently small error. 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 $\mathbf{f_Z}$ of the last Sinkhorn ~~to initialize the current one~~.

Related work. Using doubly stochastic affinities for SNE has been proposed in [**?**], with two key differences from our work. First, they do not consider EAs and resort to $\mathbf{P}^{\text{ds}}$ (**??**). This affinity, unlike $\mathbf{P}^{\text{se}}$, is not adaptive to the data heterogeneous density (as illusrated in **??**). 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 . 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 **??**. In contrast, in SNEkhorn, the latent affinity *is also doubly stochastic* so that latent coordinates are not subject to spherical constraints anymore. The corresponding SNEkhorn embedding is shown in **??** (bottom right).

figures/snekhorn_not_DS.pdf

figureLeft: SNEkhorn embedding on the simulated data of **??** using $\widetilde{\mathbf{Q}}_{\mathbf{Z}}$ instead of $\mathbf{Q}^{\text{ds}}_{\mathbf{Z}}$ with $\xi = 30$. Right: number of iterations needed to achieve $\|\mathbf{Q}^{\text{ds}}_{\mathbf{Z}}\mathbf{1} - \mathbf{1}\|_\infty \leq \text{tol}$ with (**??**).

2.5 Numerical experiments

This section aims at illustrating the performances of the proposed affinity matrix $\mathbf{P}^{\text{se}}$ (**??**) and DR method SNEkhorn at faithfully representing dependencies and 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 [**?**] on real-world images and genomics datasets.

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

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

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 $_2$ exhibit different levels of noise due to various numbers of multinomial trials (10^3 and 10^4) to form an estimation of $_2$. 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 [**?**]). This phenomenon is known as *batch effect* [**?**]. In **??**, we show that,

figures/heteroscedastic_noise.pdf

figureTop: entries of $\overline{\mathbf{P}^{\text{e}}}$ (**??**) and $\mathbf{P}^{\text{se}}$ (**??**) matrices. Bottom: embeddings generated by symmetric-SNE and SNEkhorn using the above affinities. Perplexity $\xi = 30$.

unlike $\overline{\mathbf{P}^{\mathrm{e}}}$ (**??**), $\mathbf{P}^{\mathrm{se}}$ (**??**) manages to properly filter the noise (top row) to discriminate between samples generated by $_1$ and $_2$, 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 [**?**], OLIVETTI faces [**?**], UMNIST [**?**] and CIFAR 10 [**?**]. For CIFAR, we experiment with features obtained from the last hidden layer of a pre-trained ResNet [**?**] while for the other three datasets, we take as input the raw pixel data. Regarding genomics data, we consider the Curated Microarray Database (CuMiDa) [**?**] 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 [**?**]. 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 [**?**]. In the following experiments, when not specified the hyperparameters are set to the value leading to the 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 $\mathbf{P}^{\mathrm{st}}$ [**?**] 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.

Table 2.1: ARI ($\times 100$) 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**

Spectral Clustering. Building on the strong connections between spectral clustering mechanisms and t-SNE [**?**, **?**] we first consider spectral clustering tasks to evaluate the affinity matrix $\mathbf{P}^{\mathrm{se}}$ (**??**) and compare it against $\overline{\mathbf{P}^{\mathrm{e}}}$ (**??**). We also consider two versions of the Gaussian affinity with scalar bandwidth $= \exp(-/\nu)$: 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 normalized by row and $\mathbf{P}^{\mathrm{ds}}$ (**??**). We also consider the adaptive Self-Tuning $\mathbf{P}^{\mathrm{st}}$ affinity from [**?**] 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` [**?**] with default parameters which uses the unnormalized graph Laplacian. We measure the quality of clustering using the Adjusted Rand Index (ARI). Looking at both **??** and **??**, one can notice 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 $\mathbf{P}^{\mathrm{se}}$ outperforms all the other affinities in 8 out of 12 datasets. This shows that the adaptivity of EAs is crucial. **??** 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` [**?**]. All models were optimized using `ADAM` [**?**] 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

1.01

Table 2.2: 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	$\mathbf{96.8ß \pm 0.3}$
SNAREseq	6.8 ± 6.0	35.8 ± 5.2	$\mathbf{67.9ß \pm 1.2}$	93.1 ± 2.8	99.1 ± 0.1	$\mathbf{99.2 \pm 0.1}$

this same matrix to initialize all the methods that we wish to compare. To evaluate the embeddings' quality, we make use of the silhouette [**?**] and trustworthiness [**?**] scores from `scikit-learn` [**?**] 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 **??**, 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 **??**. Notably, one can notice that the use of t-SNEkhorn results in improved class separation compared to t-SNE.

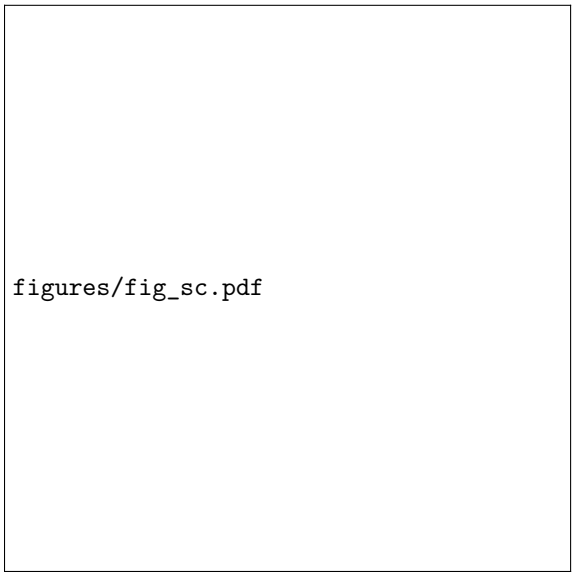
2.6 Conclusion

We have introduced a new principled and efficient method for constructing symmetric entropic affinities. Unlike the current formulation that enforces symmetry through an 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 [**?**]. 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 [**?**] as well as fast approximations for SNEkhorn forces by adapting *e.g.* Barnes-Hut [**?**] or interpolation-based methods [**?**] to the doubly stochastic setting. It could also be interesting to use SEAs in order to study the training dynamics of transformers [**?**].

Acknowledgments

The authors are grateful to Mathurin Massias, Jean Feydy and Aurélien Garivier for insightful discussions. This project was supported in part by the ANR projects AllegroAssai ANR-19-CHIA-0009, SingleStatOmics ANR-18-CE45-0023 and OTTOPIA ANR-20-CHIA-0030. This work was also supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.



figureSNAREseq embeddings produced by t-SNE and t-SNEkhorn with $\xi = 50$.

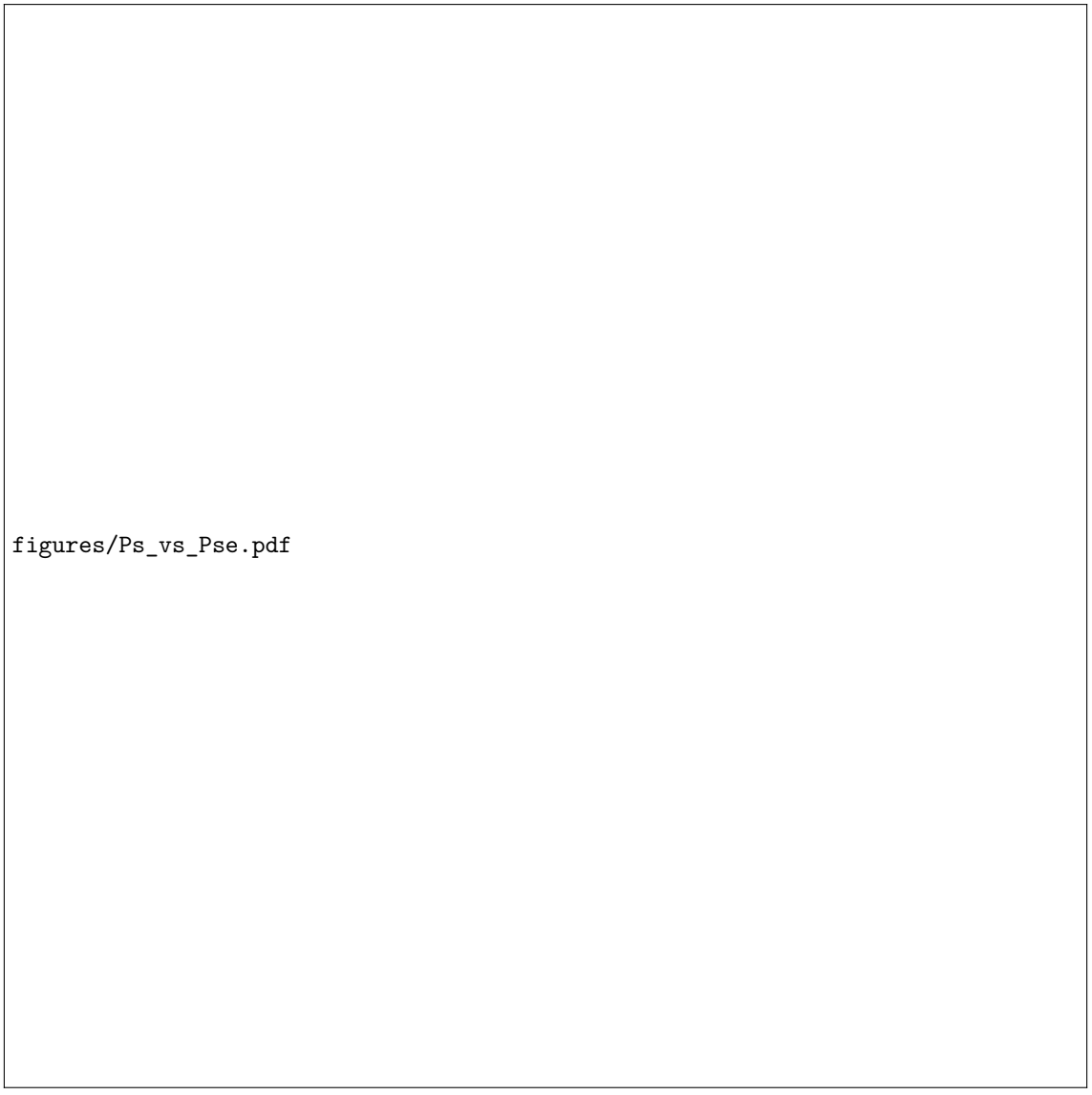


Figure 2.1: Samples from a mixture of three Gaussians with varying standard deviations. The edges' strength is proportional to the weights in the affinities $\mathbf{P}^{\text{ds}}$ (**??**) and $\mathbf{P}^{\text{se}}$ (**??**) computed with $\xi = 5$ (for $\mathbf{P}^{\text{ds}}$, ξ is the average perplexity such that $\sum_i \mathrm{H}(\mathbf{P}^{\text{ds}}_{i:}) = \sum_i \mathrm{H}(\mathbf{P}^{\text{se}}_{i:})$). Points' color represents the perplexity $e^{\mathrm{H}(\mathbf{P}_{i:})-1}$. Right plot: smallest eigenvalues of the Laplacian for the two affinities.

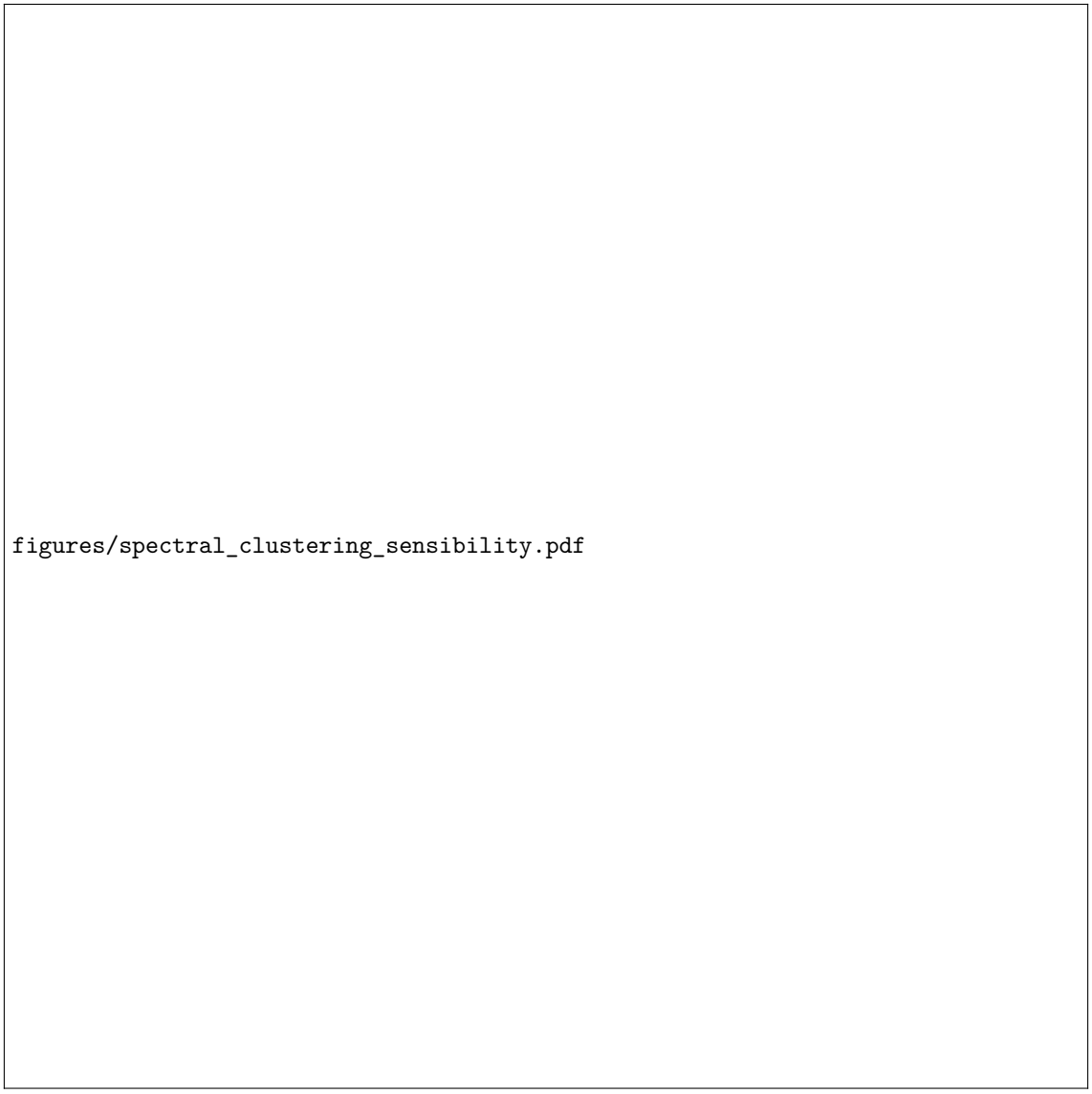


Figure 2.2: ARI spectral clustering score as a function of the perplexity parameter for image datasets.