A Probabilistic Graph Coupling View of Dimension Reduction

Hugues Van Assel
UMPA, LBMC
ENS Lyon
hugues.van_assel@ens-lyon.fr

Thibault Espinasse
Institut Camille Jordan Lyon 1
Inria Dracula
espinasse@math.univ-lyon1.fr

Julien Chiquet
AgroParisTech
INRAE
julien.chiquet@inrae.fr

Franck Picard
LBMC, CNRS
ENS Lyon
franck.picard@ens-lyon.fr

Abstract

Most popular dimension reduction (DR) methods like t-SNE and UMAP are based on minimizing a cost between input and latent pairwise similarities. Though widely used, these approaches lack clear probabilistic foundations to enable a full understanding of their properties and limitations. To that extent, we introduce a unifying statistical framework based on the coupling of hidden graphs using cross entropy. These graphs induce a Markov random field dependency structure among the observations in both input and latent spaces. We show that existing pairwise similarity DR methods can be retrieved from our framework with particular choices of priors for the graphs. Moreover this reveals that these methods relying on shift-invariant kernels suffer from a statistical degeneracy that explains poor performances in conserving coarse-grain dependencies. New links are drawn with PCA which appears as a non-degenerate graph coupling model.

1 Introduction

Dimensionality reduction (DR) is of central importance when dealing with high-dimensional data [15]. 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) [31], multidimensional scaling [23] and more broadly kernel eigenmaps methods such as Isomap [4], Laplacian eigenmaps [5] and diffusion maps [12]. 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 [16].

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 [18], its heavy-tailed symmetrized version t-SNE [38] or more recent approaches like LargeVis [36] and UMAP [30] 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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[27, 19, 1]. 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 [40]. They also demonstrate impressive performances in identifying clusters [3, 29]. However this is done at the expense of global structure, that these methods struggle in preserving [41, 11] *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 [8, 33, 28] 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 $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^\top \in \mathbb{R}^{n\times p}$, an input dataset that consists of n vectors of dimension p. Our task is to embed $\boldsymbol{X}$ in a lower dimensional space of dimension $q < p$ (typically $q = 2$ for visualization), and we denote by $\boldsymbol{Z} = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n)^\top \in \mathbb{R}^{n\times q}$ the unknown embeddings. The rationale of our framework is to suppose that the observations $\boldsymbol{X}$ and $\boldsymbol{Z}$ are structured by two latent graphs with $\boldsymbol{W}_X$ and $\boldsymbol{W}_Z$ 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 $\boldsymbol{Z}$ of $\boldsymbol{X}$ such that $\boldsymbol{W}_X$ and $\boldsymbol{W}_Z$ 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 $\boldsymbol{W}_X$ and $\boldsymbol{W}_Z$. Note that as they share the same dimensionality the latter graphs can be easily compared unlike $\boldsymbol{X}$ and $\boldsymbol{Z}$. 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 3). 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 section 2.
- 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 (theorem 1). 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 (proposition 1). These findings are highlighted by a new initialization of the embeddings that is empirically tested (section 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 2). Such model does not suffer from the aforementioned degeneracy hence preserves the large-scale structure.

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 $\boldsymbol{W} \in \mathcal{S}_W$ where:

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

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.

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*

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

With an abuse of notation, let $\boldsymbol{L} = L(\overline{\boldsymbol{W}})$ where $\overline{\boldsymbol{W}} = \boldsymbol{W} + \boldsymbol{W}^\top$. Let $(C_1, ..., C_R)$ be a partition of $[n]$ (*i.e.* the set $\{1, 2, ..., n\}$) corresponding to the connected components (CCs) of $\overline{\boldsymbol{W}}$. As well known in spectral graph theory [9], 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 $n_r = \mathrm{Card}(C_r)$. 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 $\boldsymbol{\Lambda} = \mathrm{diag}((\lambda_i)_{i\in[n]})$ with $0 = \lambda_1 = ... = \lambda_R < \lambda_{R+1} \leq ... \leq \lambda_n$. In the following, the data is split into two parts: $\boldsymbol{X}_M$, the orthogonal projection of $\boldsymbol{X}$ on $\mathcal{S}_M = (\ker \boldsymbol{L}) \otimes \mathbb{R}^p$, and $\boldsymbol{X}_C$, the projection on $\mathcal{S}_C = (\ker \boldsymbol{L})^\perp \otimes \mathbb{R}^p$. For $i\in[n]$, $\boldsymbol{X}_{M,i} = \sum_{r\in[R]} n_r^{-1}\mathbb{1}_{i\in C_r}\sum_{\ell\in C_r}\boldsymbol{X}_\ell$ hence $\boldsymbol{X}_M$ stands for the empirical means of $\boldsymbol{X}$ on CCs, thus modelling the CC positions, while $\boldsymbol{X}_C = \boldsymbol{X} - \boldsymbol{X}_M$ is CC-wise centered, thus modeling the relative positions of the nodes within CCs. We now introduce the probability distribution of these variables.

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:

$$f_k\colon (\boldsymbol{X}, \boldsymbol{W}) \mapsto \prod_{(i,j)\in[n]^2} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{W_{ij}}\,. \tag{1}$$

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 $\boldsymbol{X}$ vary smoothly on the graph $\boldsymbol{W}$. Thus we can expect $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ to be close if there is an edge between node i and node j in $\boldsymbol{W}$. A key remark is that f_k is kept invariant by translating $\boldsymbol{X}_M$. Namely for all $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $f_k(\boldsymbol{X}, \boldsymbol{W}) = f_k(\boldsymbol{X}_C, \boldsymbol{W})$. This invariance results in $f_k(\cdot, \boldsymbol{W})$ 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 $k : \boldsymbol{x} \mapsto e^{-\frac{1}{2}\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2}$ where $\boldsymbol{\Sigma}$ stands for the covariance among columns. One has:

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

by property of the graph Laplacian (definition 1). In this case, it is clear that due to the rank deficiency of $\boldsymbol{L}$, $f_k(\cdot, \boldsymbol{W})$ 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 [38]. Heavy-tailed kernels appear useful when the dimension of the embeddings is smaller than the intrinsic dimension of the data [20]. Our contribution provides flexibility by extending the previous result to a large class of kernels, as stated in the following theorem.

Theorem 1 *If k is $\lambda_{\mathbb{R}^p}$-integrable and bounded above $\lambda_{\mathbb{R}^p}$-almost everywhere then $f_k(\cdot, \boldsymbol{W})$ is $\lambda_{\mathcal{S}_C}$-integrable.*

We refer to appendix A.1 for the proof. We can now define a distribution on $(\mathcal{S}_C, \mathcal{B}(\mathcal{S}_C))$, where $\mathcal{C}_k(\boldsymbol{W}) = \int f_k(\cdot, \boldsymbol{W})d\lambda_{\mathcal{S}_C}$:

$$\mathbb{P}_k(d\boldsymbol{X}_C|\boldsymbol{W}) = \mathcal{C}_k(\boldsymbol{W})^{-1} f_k(\boldsymbol{X}_C, \boldsymbol{W})\lambda_{\mathcal{S}_C}(d\boldsymbol{X}_C)\,. \tag{3}$$

Remark 1 *Kernels may have node-specific bandwidths $\boldsymbol{\tau}$, set during a pre-processing step, giving $f_k(\boldsymbol{X}, \boldsymbol{W}) = \prod_{(i,j)} k((\boldsymbol{X}_i - \boldsymbol{X}_j)/\tau_i)^{W_{ij}}$. Note that such bandwidth does not affect the degeneracy of the distribution and theorem 1 still holds.*

Between-Rows Dependency Structure. By symmetry of k, reindexing gives: $f_k(\boldsymbol{X}, \boldsymbol{W}) = \prod_{j\in[n]} \prod_{i\in[j]} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\overline{W}_{ij}}$. Hence distribution (3) boils down to a pairwise MRF model [10] with respect to the undirected graph $\overline{\boldsymbol{W}}$, $\mathcal{C}_k$ playing the role of the partition function. Note that since f_k (Equation 1) trivially factorize according to the cliques of $\overline{\boldsymbol{W}}$, the Hammersley-Clifford theorem ensures that the rows of $\boldsymbol{X}_C$ satisfy the local and global Markov properties with respect to $\overline{\boldsymbol{W}}$.

2.3 Uninformative Model for CC-wise Means

We showed that the MRF (1) is only integrable on $\mathcal{S}_C$, the definition of which depends on the connectivity structure of $\boldsymbol{W}$. 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.* [35] for a similar construction in the Gaussian case. We introduce the Borel function $f^\varepsilon(\cdot, \boldsymbol{W})\colon \mathbb{R}^{n\times p} \to \mathbb{R}_+$ for $\varepsilon > 0$ such that for all $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $f^\varepsilon(\boldsymbol{X}, \boldsymbol{W}) = f^\varepsilon(\boldsymbol{X}_M, \boldsymbol{W})$. To allow f^ε to become arbitrarily non-informative, we assume that for all $\boldsymbol{W} \in \mathcal{S}_W$, $f^\varepsilon(\cdot, \boldsymbol{W})$ is $\lambda_{\mathcal{S}_M}$-integrable for all $\varepsilon \in \mathbb{R}_+^*$ and $f^\varepsilon(\cdot, \boldsymbol{W}) \xrightarrow[\varepsilon\to 0]{} 1$ almost everywhere. We now define the conditional distribution on $(\mathcal{S}_M, \mathcal{B}(\mathcal{S}_M))$ as follows:

$$\mathbb{P}^\varepsilon(d\boldsymbol{X}_M|\boldsymbol{W}) = \mathcal{C}^\varepsilon(\boldsymbol{W})^{-1} f^\varepsilon(\boldsymbol{X}_M, \boldsymbol{W}) \lambda_{\mathcal{S}_M}(d\boldsymbol{X}_M) \tag{4}$$

where $\mathcal{C}^\varepsilon(\boldsymbol{W}) = \int f^\varepsilon(\cdot, \boldsymbol{W}) d\lambda_{\mathcal{S}_M}$. With this at hand, the joint conditional is defined as the product measure of (3) and (4) over the row axis, the integrability of which is ensured by the Fubini-Tonelli theorem. In the following we will use the compact notation $\mathcal{C}_k^\varepsilon(\boldsymbol{W}) = \mathcal{C}_k(\boldsymbol{W})\mathcal{C}^\varepsilon(\boldsymbol{W})$ for the joint normalizing constant.

Remark 2 *At the limit $\varepsilon \to 0$ the above construction amounts to setting an infinite variance on the distribution of the empirical means of $\boldsymbol{X}$ 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 $f^\varepsilon : \boldsymbol{X} \to \exp\left(-\frac{\varepsilon}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{U}_{[:R]}\boldsymbol{\Theta}\boldsymbol{U}_{[R]}^\top \boldsymbol{X}\right)\right)$ such that: $vec(\boldsymbol{X}_M)|\boldsymbol{\Theta} \sim \mathcal{N}\left(\boldsymbol{0}, \left(\varepsilon\boldsymbol{U}_{[:R]}\boldsymbol{\Theta}\boldsymbol{U}_{[R]}^\top\right)^{-1} \otimes \boldsymbol{\Sigma}\right)$ where $\otimes$ denotes the Kronecker product.

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 $\boldsymbol{W}$. Our methodology is similar to that of constructing conjugate priors for distributions in the exponential family [39], 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 [26].

Definition 2 *Let $\boldsymbol{\pi} \in \mathbb{R}_+^{n\times n}$, $\varepsilon \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$, k satisfies the assumptions of theorem 1 and $\mathcal{P} \in \{B, D, E\}$. For $\boldsymbol{W} \in \mathcal{S}_W$ we introduce:*

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

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

When $\alpha = 0$, the above no longer depends on ε and k. We will use the compact notation $\mathbb{P}_{\mathcal{P}}(\boldsymbol{W};\boldsymbol{\pi}) = \mathbb{P}^{\varepsilon}_{\mathcal{P},k}(\boldsymbol{W};\boldsymbol{\pi},0)$. Note that by $\boldsymbol{W} \sim \mathbb{P}_{\mathcal{P}}(\cdot\,;\boldsymbol{\pi})$ 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 $\mathcal{P} = D$, $\forall i \in [n]$, $\boldsymbol{W}_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{M}\left(1, \boldsymbol{\pi}_i/\pi_{i+}\right)$.
- If $\mathcal{P} = E$, $\boldsymbol{W} \sim \mathcal{M}\left(n, \boldsymbol{\pi}/\pi_{++}\right)$.

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 appendix A.2). In the following, $\odot$ stands for the Hadamard product and $\mathcal{D}$ for the convergence in distribution.

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

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

Remark 3 *For all $\boldsymbol{W} \in \mathcal{S}_W$, $\mathcal{C}^{\varepsilon}(\boldsymbol{W})$ diverges as $\varepsilon \to 0$, hence the graph prior (definition 2) is improper at the limit. This compensates for the uninformative diffuse conditional and allows to retrieve a well-defined tractable posterior limit.*

3.1 Retrieving Well Known DR 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 1 and we denote by $\boldsymbol{K}_X$ and $\boldsymbol{K}_Z$ the associated kernel matrices on $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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 $\varepsilon \to 0$,

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

defining a coupling criterion to be optimized with respect to embedding coordinates $\boldsymbol{Z}$. We now go through each couple $(\mathcal{P}_X, \mathcal{P}_Z)$ such that $\mathrm{supp}\left(\mathbb{P}_{\mathcal{P}_X}\right) \subset \mathrm{supp}\left(\mathbb{P}_{\mathcal{P}_Z}\right)$ 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 [17], where for $i \in [n]$, $\boldsymbol{P}^D_i = \boldsymbol{K}_{X,i}/K_{X,i+}$ and $\boldsymbol{Q}^D_i = \boldsymbol{K}_{Z,i}/K_{Z,i+}$,

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

Symmetric-SNE. Choosing $\mathcal{P}_X = D$ and $\mathcal{P}_Z = E$, we define for $(i,j) \in [n]^2$, $\boldsymbol{Q}^E_{ij} = K_{Z,ij}/K_{Z,++}$ and $\overline{P}^D_{ij} = P^D_{ij} + P^D_{ji}$. The symmetry of $\boldsymbol{Q}^E$ 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 [38] 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

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

which is the objective of LargeVis [36].

Table 1: Prior distributions for $\boldsymbol{W}_X$ and $\boldsymbol{W}_Z$ 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 $\widetilde{\boldsymbol{W}}_X = \mathbb{1}_{\boldsymbol{W}_X + \boldsymbol{W}_X^\top \geq 1}$. 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 $\boldsymbol{P}^B = (K_{X,ij}/(1+K_{X,ij}))_{(i,j)\in[n]^2}$. Coupling $\widetilde{\boldsymbol{W}}_X$ and $\boldsymbol{W}_Z$ 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 [30], the construction of $\widetilde{\boldsymbol{W}}_X$ being borrowed from section 3.1 of the paper.

Remark 4 *One can also consider $\mathcal{H}_{E,E}$ but as detailed in [38], 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.*

3.2 Interpretations

As we have seen in section 3.1, 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 4 and section 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:

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times q}} -\sum_{(i,j)\in[n]^2} \boldsymbol{P}_{ij}^{\mathcal{P}_X} \log k_z(\boldsymbol{Z}_i - \boldsymbol{Z}_j) + \log \mathbb{P}(\boldsymbol{Z}). \tag{5}$$

The first and second terms in eq. (5) respectively summarize the attractive and repulsive forces of the objective. Recall from proposition 1 that $\boldsymbol{P}^{\mathcal{P}_X}$ is the posterior expectation of $\boldsymbol{W}_X$. Hence in SNE-like methods, the attractive forces resume to a pairwise MRF log likelihood with respect to a graph posterior expectation given $\boldsymbol{X}$. For instance if k_z is the Gaussian kernel, this attractive term reads $\operatorname{tr}\left(\boldsymbol{Z}^\top \boldsymbol{L}^\star \boldsymbol{Z}\right)$ where $\boldsymbol{L}^\star = \mathbb{E}_{\boldsymbol{W}\sim\mathbb{P}_{\mathcal{P}_X}(\cdot;\boldsymbol{K}_X)}[L(\boldsymbol{W})]$, boiling down to the objective of Laplacian eigenmaps [5]. Therefore, for Gaussian MRFs, the attractive forces resume to an unconstrained Laplacian eigenmaps objective. Such link, already noted in [7], is easily unveiled in our framework. Moreover, one can notice that only this attractive term depends on $\boldsymbol{X}$ as the repulsion is given by the marginal term in (5). The latter reads $\mathbb{P}(\boldsymbol{Z}) = \sum_{\boldsymbol{W}\in\mathcal{S}_W} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W})$ with $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) \propto f_k(\boldsymbol{Z},\boldsymbol{W})\Omega_{\mathcal{P}_Z}(\boldsymbol{W})$. Such penalty notably prevents a trivial solution, as $\mathbf{0}$, like any constant vector, is a mode of $f_k(\cdot,\boldsymbol{W})$ for all $\boldsymbol{W}$. Also note that the prior for $\boldsymbol{W}_X$ only conditions attraction while the prior for $\boldsymbol{W}_Z$ 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 [6] for a quantitative study of attraction and repulsion in these methods.

Global Structure Preservation. To gain intuition, consider that $\boldsymbol{W}_X$ is observed. As we showed in section 2.2, 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 $\varepsilon \to 0$, $\boldsymbol{X}_M$ and $\boldsymbol{Z}_M$ 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 $\boldsymbol{W}_X$ is considered, especially when clusters are well distinguishable in $\boldsymbol{X}$. 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 1. In the following section we propose to mitigate the global structure deficiency with non-degenerate MRF models.

4 Towards Capturing Large-Scale Dependencies

In this section, we investigate the ability of graph coupling to faithfully represent global structure in low dimension. 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.

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 (3), a natural solution would be to consider graphical models that are well defined and integrable on the entire definition spaces of $\boldsymbol{X}$ and $\boldsymbol{Z}$. 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 $\boldsymbol{\Theta}_Z$, $\mathbb{P}(\boldsymbol{\Theta}_Z = \cdot)$ denotes its density.

Theorem 2 *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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ such that:*

$$vec(\boldsymbol{X})|\boldsymbol{\Theta}_X \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta}_X^{-1} \otimes \boldsymbol{I}_p), \tag{6}$$

$$vec(\boldsymbol{Z})|\boldsymbol{\Theta}_Z \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta}_Z^{-1} \otimes \boldsymbol{I}_q)\,. \tag{7}$$

Then the solution of the precision coupling problem:

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

is a PCA embedding of $\boldsymbol{X}$ 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 [35]. When choosing the Gaussian kernel for neighbor embedding methods, we saw that the graph Laplacian $\boldsymbol{L}_X$ of $\boldsymbol{W}_X$ was playing the role of the among-row precision matrix, as we had $\boldsymbol{X}|\boldsymbol{W}_X \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{L}_X^{-1} \otimes \boldsymbol{I}_p)$ (equation 2). Recall that the later always has a null-space which is spanned by the CC indicator vectors of $\boldsymbol{W}$ (section 2.1). Here, the key difference is that we impose a full-rank constraint on the precision $\boldsymbol{\Theta}$. Concerning the priors, we choose the ones that are conjugate to the conditionals (6) and (7), as previously done when constructing the prior for neighbor embedding methods (definition 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 [37], in the above the linear relation between $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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 [2]. Therefore, as suspected in section 3.2, the degeneracy of the conditional distribution given the graph is key to determine the distance preservation properties of the embeddings. We propose in section 4.2 to combine both graph coupling approaches to strike a balance between global and local structure preservation.

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 $\boldsymbol{X}$, where $\mathcal{P}_X \in \{B, D, E\}$, k_x satisfies the assumptions of theorem 1 and $\nu_X \geq n$:

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

where $\boldsymbol{U}_{[R]}$ are the eigenvectors associated to the Laplacian null-space of $\overline{\boldsymbol{W}}_X$. Given a graph $\boldsymbol{W}_X$, the idea is to structure the CCs' relative positions with a full-rank Gaussian model. The same model is considered for $\boldsymbol{W}_Z$, $\boldsymbol{\Theta}_Z$ and $\boldsymbol{Z}$, choosing $\nu_Z = \nu_X + p - q$ for the Wishart prior to satisfy the assumption of theorem 2. With this in place, we aim at providing a complete coupling objective, matching the pairs $(\boldsymbol{W}_X, \boldsymbol{\Theta}_X)$ and $(\boldsymbol{W}_Z, \boldsymbol{\Theta}_Z)$. The joint negative cross-entropy can be decomposed as follows:

$$\mathbb{E}_{(\boldsymbol{W}_X, \boldsymbol{\Theta}_X)|\boldsymbol{X}} \left[\log \mathbb{P}((\boldsymbol{W}_Z, \boldsymbol{\Theta}_Z) = (\boldsymbol{W}_X, \boldsymbol{\Theta}_X) | \boldsymbol{Z})\right]$$
$$= \mathbb{E}_{\boldsymbol{W}_X|\boldsymbol{X}} \left[\log \mathbb{P}(\boldsymbol{W}_Z = \boldsymbol{W}_X | \boldsymbol{Z})\right] + \tag{8}$$
$$\mathbb{E}_{(\boldsymbol{W}_X, \boldsymbol{\Theta}_X)|\boldsymbol{X}} \left[\log \mathbb{P}(\boldsymbol{\Theta}_Z = \boldsymbol{\Theta}_X | \boldsymbol{W}_Z = \boldsymbol{W}_X, \boldsymbol{Z})\right] \tag{9}$$

where (8) is the usual coupling criterion of $\boldsymbol{W}_X$ and $\boldsymbol{W}_Z$ capturing intra-CC variability while (9) 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 [21], we propose to take care of the global structure first *i.e.* focusing on (9) before (8). The difficulty of dealing with (9) lies in the hierarchical construction of the graph and the Gaussian precision (see fig. 3). We state the following result.

Corollary 1 *Let $\boldsymbol{W}_X \in \mathcal{S}_W$, $\boldsymbol{L} = L(\overline{\boldsymbol{W}}_X)$ 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:*

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

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

Remark 5 *Note that while (8) approximates the objective of SNE-like methods when $\varepsilon \to 0$, the minimizer of (9) given by corollary 1 is stable for all ε.*

From this observation, we propose a simple heuristic to minimize (9) that consists in computing a PCA embedding of $\mathbb{E}_{\mathbb{P}_{\mathcal{P}_X}(\cdot; \boldsymbol{K}_X)} \left[\boldsymbol{U}_{[:R]}\boldsymbol{U}_{[R]}^\top\right] \boldsymbol{X}$. The distribution of the connected components of the posterior of $\boldsymbol{W}_X$ 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 (8) which is done by running the DR method corresponding to the graph priors at hand (section 3.1). This second step essentially consists in refining the intra-CC structure.

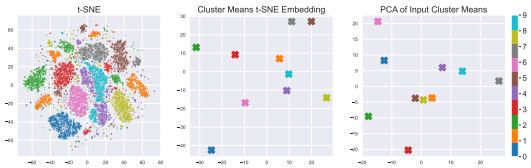

Figure 1: Left: MNIST t-SNE (perp : 30) embeddings initialized with i.i.d $\mathcal{N}(0, 1)$ coordinates. Middle: using these t-SNE embeddings, mean coordinates for each digit are represented. Right: we compute a matrix of mean input coordinates for each of the 10 digits and embed it using PCA. For 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).

4.3 Experiments with *ccPCA*

Figure 1 shows that a t-SNE embedding of a balanced MNIST dataset of 10000 samples [14] 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.

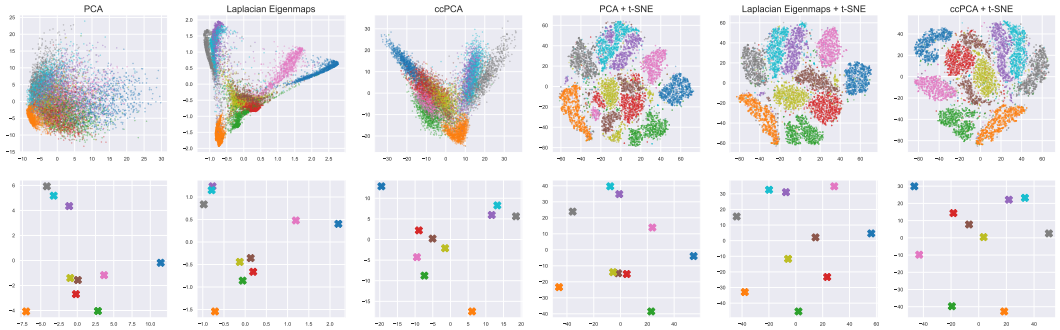

Figure 2: Top: MNIST embeddings produced by PCA, Laplacian eigenmaps, *ccPCA* and finally t-SNE launched after the previous three embeddings to improve the fine-grain structure. Bottom: mean coordinates for each digit using the embeddings of the first row. The color legend is the same as in fig. 1. t-SNE was trained during 1000 iterations using default parameters with the openTSNE implementation [34].

Recall that the perplexity value [38] 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 3.2. To remedy this in practice, PCA or Laplacian eigenmaps are usually used as initialization [21].

These strategies are tested (fig. 2) 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 fig. 1. Other perplexity values for *ccPCA* are explored in appendix B.2 while the experimental setup is detailed in appendix B.1. In appendix B.3, 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 variablity and focusing solely on the inter-cluster structure. Compared to Laplacian eigenmaps which performs 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.

5 Conclusion and Perspectives

In this work we shed a new light on 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.

Acknowledgments. The authors would like to thank the anonymous reviewers whose comments and questions helped improve the clarity of this manuscript, as well as Aurélien Garivier and Floshi Poshi for helpful discussions. This work was supported by the Agence Nationale de la Recherche ANR-18-CE45-0023 SingleStatOmics.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

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

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

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

[13] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

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

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009.

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

[24] Nadia S Kurd, Zhaoren He, Tiani L Louis, J Justin Milner, Kyla D Omilusik, Wenhao Jin, Matthew S Tsai, Christella E Widjaja, Jad N Kanbar, Jocelyn G Olvera, et al. Early precursors and molecular determinants of tissue-resident memory cd8+ t lymphocytes revealed by single-cell rna sequencing. *Science immunology*, 5(47):eaaz6894, 2020.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[33] 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.

[34] Pavlin G Poličar, Martin Stražar, and Blaž Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *BioRxiv*, page 731877, 2019.

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

[36] 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.

[37] 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.

[38] 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. Pagination: 27.

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

[40] 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.

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

[42] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In appendix B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

A.1 Proof of theorem 1

$\boldsymbol{W} \in \mathcal{S}_W$ is the weight matrix of a graph with R connected components $\{C_1, ..., C_R\}$ 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 $\boldsymbol{W}$, since each edge of $\boldsymbol{W}$ is bounded by n, one has:

$$\begin{aligned}\int f_k(\boldsymbol{X}, \boldsymbol{W}) \lambda_{\mathcal{S}_C}(d\boldsymbol{X}) &= \int \prod_{(i,j)\in[n]^2} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{W_{ij}} \lambda_{\mathcal{S}_C}(d\boldsymbol{X}) \\ &\leq M_+^{n^3} \int \prod_{(i,j)\in[n]^2} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathcal{T}_{ij}} \lambda_{\mathcal{S}_C}(d\boldsymbol{X}) \\ &\leq M_+^{n^3} \prod_{r\in[R]} \int \prod_{(i,j)\in C_r^2} k(\boldsymbol{X}_i - \boldsymbol{X}_j)^{\mathcal{T}_{ij}} \lambda_{\mathcal{S}_C}(d\boldsymbol{X}) \,. \end{aligned} \tag{10}$$

Let $r \in [R]$. The spanning tree corresponding to the rth 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 $\boldsymbol{X} \in \mathcal{S}_C$ one has: $\sum_{i\in C_r} \boldsymbol{X}_i = 0$ hence $\boldsymbol{X}_\ell = -\sum_{i\neq\ell} \boldsymbol{X}_i$. Let us now consider the linear map ϕ^r such that:

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

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:

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

where $\boldsymbol{A}$ 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:

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

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 $|\boldsymbol{J}_{\phi^r}| \neq 0$ and ϕ^r is a $\mathcal{C}^1$ diffeomorphism. This change of variable yields:

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

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

using the Fubini Tonelli theorem. The result follows from $\lambda_{\mathbb{R}^p}$-integrability of k and upper bound 10.

A.2 Proof of proposition 1

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

$$\forall \boldsymbol{W} \in \mathcal{S}_W, \quad \mathbb{P}(\boldsymbol{W}|\boldsymbol{X}) \propto \mathcal{C}^{\varepsilon}_k(\boldsymbol{W})^{-1} f^{\varepsilon}(\boldsymbol{X}, \boldsymbol{W}) f_k(\boldsymbol{X}, \boldsymbol{W}) \mathbb{P}^{\varepsilon}_{\mathcal{P},k}(\boldsymbol{W}; \boldsymbol{\pi}, 1) \tag{11}$$

where the prior reads:

$$\mathbb{P}^{\varepsilon}_{\mathcal{P},k}(\boldsymbol{W}; \boldsymbol{\pi}, 1) \propto \mathcal{C}^{\varepsilon}_k(\boldsymbol{W}) \Omega_{\mathcal{P}}(\boldsymbol{W}) \prod_{(i,j)\in[n]^2} \pi_{ij}^{W_{ij}} \,. \tag{12}$$

Hence the joint normalizing constant simplifies such that:

$$\forall \boldsymbol{W} \in \mathcal{S}_W, \quad \mathbb{P}(\boldsymbol{W}|\boldsymbol{X}) \propto f^{\varepsilon}(\boldsymbol{X}, \boldsymbol{W}) \Omega_{\mathcal{P}}(\boldsymbol{W}) \prod_{(i,j)\in[n]^2} (\pi_{ij} k(\boldsymbol{X}_i - \boldsymbol{X}_j))^{W_{ij}} \tag{13}$$

$$\xrightarrow[\varepsilon\to 0]{} \Omega_{\mathcal{P}}(\boldsymbol{W}) \prod_{(i,j)\in[n]^2} (\pi_{ij} k(\boldsymbol{X}_i - \boldsymbol{X}_j))^{W_{ij}} \tag{14}$$

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:

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

Therefore the posterior limit graph has the distribution:

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

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

***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:

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

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

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

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

Finally, deriving the distribution of the posterior graph limit:

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

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

A.3 Proof of theorem 2

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

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

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

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

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

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

Our strategy is to first find the optimal sample covariance matrix $\boldsymbol{Z}\boldsymbol{Z}^\top$ and then focus on the solution in $\boldsymbol{Z}$. To that extend, consider the eigendecomposition of the sample covariance matrices: $\boldsymbol{X}\boldsymbol{X}^\top=\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$ and $\boldsymbol{Z}\boldsymbol{Z}^\top=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{D}=\operatorname{diag}(\boldsymbol{d})$ and $\boldsymbol{\Lambda}=\operatorname{diag}(\boldsymbol{\lambda})$ such that $d_1\geq\ldots\geq d_n$ and $\lambda_1\geq\ldots\geq\lambda_n$. Denoting $\gamma=(\nu_X+q)/(\nu_Z+p)$, we consider the following problem:

$$\min_{\boldsymbol{U}\in\mathcal{O}(n),\boldsymbol{\Lambda}} \quad \operatorname{tr}\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{V}(\boldsymbol{I}_n+\boldsymbol{D})^{-1}\boldsymbol{V}^\top\right)-\gamma\log|\boldsymbol{I}_n+\boldsymbol{\Lambda}| \tag{15}$$

$$\text{s.t.} \quad \boldsymbol{\Lambda}\succcurlyeq\boldsymbol{0} \tag{16}$$

$$\operatorname{rank}(\boldsymbol{\Lambda})\leq q \tag{17}$$

The above problem is non-convex because of the rank constraint (17). Nonetheless it can be simplified as we now show.

We focus on finding the optimal eigenvectors first. To that extent, let us denote, $\boldsymbol{R}=\boldsymbol{U}^\top\boldsymbol{V}$. Only the left term in (18) depends on $\boldsymbol{R}$. The optimization problem for eigenvectors writes:

$$\min_{\boldsymbol{R}\in\mathcal{O}(n)} \quad \operatorname{tr}\left(\boldsymbol{R}^\top\boldsymbol{\Lambda}\boldsymbol{R}(\boldsymbol{I}_n+\boldsymbol{D})^{-1}\right) \tag{18}$$

The objective (18) 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 (18) 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 $\ell \in [L]$, $\theta_\ell =_{\sigma_\ell = \sigma^\star}$ minimizes the latter quantity. Therefore the solution of (18) $\boldsymbol{R}^\star$ 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 18 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 $\boldsymbol{Z}$ as a PCA embedding. Such a choice gives $\boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$.

Now it remains to find the optimal eigenvalues $(\lambda_i)_{i \in [n]}$. The rank constraint (17) 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 $\mathbb{R}^q$:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^q} \quad \boldsymbol{\lambda}^\top (\mathbf{1} + \boldsymbol{d})^{-1} - \gamma \mathbf{1}^\top \log(\mathbf{1} + \boldsymbol{\lambda}) \tag{19}$$

$$\text{s.t.} \quad \forall i \in [q], \quad \lambda_i \geq 0, \quad \lambda_1 \geq ... \geq \lambda_q \tag{20}$$

where (20) accounts for (16). The above is convex. (19) is minimized for $\boldsymbol{\lambda} = \gamma(\mathbf{1} + \boldsymbol{d}) - \mathbf{1}$. Taking the feasibility constraint (20) 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 $\boldsymbol{Z}^\star \boldsymbol{Z}^{\star\,T} = \boldsymbol{V}_{[:,q]} \boldsymbol{D}_{[q,q]} \boldsymbol{V}_{[q,:]}^\top$. Therefore there exists $\boldsymbol{R}$ an orthogonal matrix of size q such that $\boldsymbol{Z}^\star = \boldsymbol{V}_{[:,q]} \boldsymbol{D}_{[q,q]}^{\frac{1}{2}} \boldsymbol{R}$. The latter is the output of a PCA model of $\boldsymbol{X}$ with q components, which is defined up to a rotation.

A.4 Proof of Corollary 1

With the presented hierarchical model (fig. 3), the coupling problem is the following:

$$\min_{\boldsymbol{Z} \in \mathcal{S}_M^q} \quad \operatorname{tr}\left(\boldsymbol{U}_{[:R]} \boldsymbol{Z}^\top (\boldsymbol{I}_R + \varepsilon \boldsymbol{U}_{[R]}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{U}_{[:R]})^{-1} \boldsymbol{U}_{[R]}^\top \boldsymbol{Z}\right) - \log |\boldsymbol{I}_R + \varepsilon \boldsymbol{U}_{[R]}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{U}_{[:R]}| \tag{21}$$

where $\boldsymbol{U}_{[:R]}$ are the eigenvectors associated to the Laplacian null-space of $\overline{\boldsymbol{W}}_X$.

Let us denote $\bar{\boldsymbol{Z}} = \boldsymbol{U}_{[R]}^\top \boldsymbol{Z} \in \mathbb{R}^{R \times q}$ and $\bar{\boldsymbol{X}} = \boldsymbol{U}_{[R]}^\top \boldsymbol{X} \in \mathbb{R}^{R \times p}$. Note that $\boldsymbol{Z} \to \boldsymbol{U}_{[R]}^\top \boldsymbol{Z}$ is a bijective linear map from $\mathcal{S}_M^q$ to $\mathbb{R}^{R \times q}$ with inverse $\bar{\boldsymbol{Z}} \to \boldsymbol{U}_{[R]} \bar{\boldsymbol{Z}}$ (and equivalently for $\mathbb{R}^{R \times p}$). Hence (21) is equivalent to:

$$\min_{\bar{\boldsymbol{Z}} \in \mathbb{R}^{R \times q}} \quad \operatorname{tr}\left(\bar{\boldsymbol{Z}}^\top (\boldsymbol{I}_R + \varepsilon \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top)^{-1} \bar{\boldsymbol{Z}}\right) - \log |\boldsymbol{I}_R + \varepsilon \bar{\boldsymbol{Z}}\bar{\boldsymbol{Z}}^\top| \tag{22}$$

According to theorem 2, the solution of problem (22) is such that there exists $\boldsymbol{R}$ orthogonal, $\bar{\boldsymbol{Z}}^\star = \boldsymbol{V}_{[:,q]} \boldsymbol{S}_{[q,q]} \boldsymbol{R}$ where $\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top = \boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^\top$ is the eigendecomposition in an orthogonal basis of the among-row covariance matrix of $\bar{\boldsymbol{X}}$. Note that the solution does not depend on ε.

Therefore (21) is solved for $\boldsymbol{Z}^\star = \boldsymbol{U}_{[:R]} \boldsymbol{V}_{[:,q]} \boldsymbol{S}_{[q,q]} \boldsymbol{R}$. One can notice that since the singular value decomposition (*i.e.* SVD) of $\boldsymbol{U}_{[R]}^\top \boldsymbol{X}$ takes the form $\boldsymbol{V}\boldsymbol{S}\boldsymbol{B}$ where $\boldsymbol{B}$ is an semi-orthogonal matrix of size p, then $\boldsymbol{U}_{[:R]} \boldsymbol{U}_{[R]}^\top \boldsymbol{X} = \boldsymbol{U}_{[:R]} \boldsymbol{V}\boldsymbol{S}\boldsymbol{B}$. Noticing that $\boldsymbol{V}' = \boldsymbol{U}_{[:R]} \boldsymbol{V}$ is orthogonal, one has that $\boldsymbol{V}'\boldsymbol{S}\boldsymbol{B}$ is a compact SVD of $\boldsymbol{U}_{[:R]} \boldsymbol{U}_{[R]}^\top \boldsymbol{X}$. Therefore, since $\boldsymbol{Z}^\star = \boldsymbol{V}'\boldsymbol{S}$, $\boldsymbol{Z}^\star$ is a PCA embedding of $\boldsymbol{U}_{[:R]} \boldsymbol{U}_{[R]}^\top \boldsymbol{X}$.

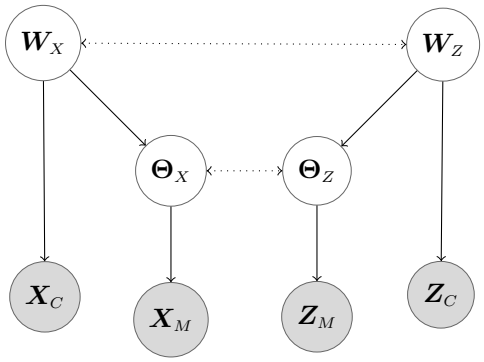

Figure 3: Graphical representation of the hierarchical model considered in section 4.2. Plain directed arrows represent conditional dependencies while dotted arrows represent the coupling links. Corollary 1 provides a solution for the coupling between $\boldsymbol{\Theta}_X$ and $\boldsymbol{\Theta}_Z$.

B Experiments Supplementary Material

B.1 Experimental Setup and Details About *ccPCA*

Implementation of existing methods. For t-SNE, we rely on the openTSNE implementation [34] for both computing the kernel $\boldsymbol{K}_X$ with appropriate bandwidths and running the tSNE algorithm. We keep the training default parameters and 1000 iterations of gradient descent. For all experiments, the default perplexity of 30 was used to set the kernel bandwidths. For UMAP, we use the default Python implementation of [30] with default parameters. For PCA and Laplacian eigenmaps, the scikit-learn implementation is used [32] with default parameters as well.

ccPCA. The pseudo code of the algorithm is given in algorithm 1. CCs' memberships (*i.e.* eigenvectors $\boldsymbol{U}_{[R]}$) are computed using igraph [13]. Regarded the time complexity of ccPCA, one can sample the posterior graph with constant time if $\mathcal{P}_X = E$, linear time if $\mathcal{P}_X = D$ and quadratic time if $\mathcal{P}_X = B$. Moreover, computing $\boldsymbol{U}_{[R]}$ can be done with linear complexity *w.r.t.* the number of nodes. Hence the time complexity is $O(N \times n)$ for E and D priors and $O(N \times n^2)$ for the B prior, where N is the number of Monte Carlo samples. In practice we found that $N \approx 100$ Monte Carlo samples produce a consistent *ccPCA* embedding for $n \approx 10000$. Note that the time complexity of PCA is $O(\min(p^3, n^3))$ where p is the dimensionality (*i.e.* number of columns) of $\boldsymbol{X}$. Hence in most common applications involving images or biological sequencing data (where p is very large), the additional time complexity brought by *ccPCA* compared to PCA is negligible.

Algorithm 1 *ccPCA*

Input: $\boldsymbol{K}_X$, $\mathcal{P}_X$, N
for $\ell = 1$ **to** N **do**
  Sample $\boldsymbol{W}^\ell \sim \mathbb{P}_{\mathcal{P}_X}(\cdot; \boldsymbol{K}_X)$
  Compute CCs' memberships $\boldsymbol{U}^\ell_{[R]}$ of $\boldsymbol{W}^\ell$
end for
Output: PCA of $\left(N^{-1} \sum_{\ell \in [N]} \boldsymbol{U}^\ell_{[R]} \boldsymbol{U}^{\ell\,T}_{[R]}\right) \boldsymbol{X}$

All experiments are performed on a machine with four Intel Core i5 processors and 16 GB memory.

B.2 *ccPCA* with Varying Perplexity Values

Recall that the *ccPCA* algorithm retrieves the same latent graph as neighbor embedding methods. As shown in section 3.1, these graphs' distributions depend on the type of prior considered, and take simple forms as follows, when $\boldsymbol{\pi}_X = \mathbf{1}$:

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

and $\boldsymbol{K}_X$ is the kernel matrix evaluated on the data such that:

$$\forall (i,j) \in [n]^2, \quad \boldsymbol{K}_{X,ij} = k((\boldsymbol{X}_i - \boldsymbol{X}_j)/\tau_i)$$

where $\boldsymbol{\tau} \in \mathbb{R}^n$ is set using an heuristic depending on the method considered [38, 30, 36]. In fig. 4, we focus on the effect of the kernel bandwidths on *ccPCA*, choosing the example of t-SNE.

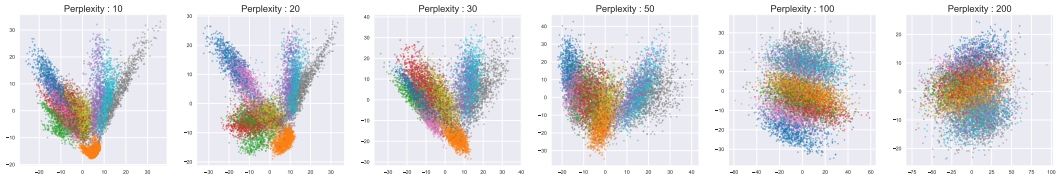

Figure 4: *ccPCA* launched for different values of the perplexity parameter. The latter determines the kernel bandwidths and can be interpreted as the number of effective neighbors of each point [38]. As the perplexity grows, the probability of connecting different clusters of digit by sampling through the graph posterior $\mathbb{P}_{\mathcal{P}_X}(\cdot; \boldsymbol{K}_X)$ increases. Therefore clusters are less and less identifiable as the perplexity increases.

From fig. 4, one can notably notice that using a high perplexity leads to a more connected graph and therefore a PCA-like embedding with less degeneracy and no clustering effect. Recall that *ccPCA* computes the same clusters as t-SNE through the CCs of the latent MRF and manage to position t-SNE clusters by focusing on their relative positions (that are filtered by t-SNE). In the case of a connected graph (high perplexity), *ccPCA* will show little advantage over classical PCA since there will not be any cluster to position. Note that this discussion can be extended to other neighbor embedding methods equivalently. Therefore, our probabilistic framework allows us to indentify which part of information is filtered by the posterior graphs with given kernel bandwidth.

B.3 Quantitative Evaluation of *ccPCA*

For quantitative assessment of *ccPCA*, we focused on t-SNE [38] and UMAP [30] which are the most popular neighbor embedding methods. Note that for these algorithms the initialization is crucial for the global structure of the embeddings as shown in [21]. In addition to MNIST [14], we considered the datasets cifar-10, cifar-100 [22], fashion-MNIST [42] as well as the CD8+ T lymphocytes single cell RNA-seq dataset from [24].

We used the quantitative criterion of [25] to assess the quality of the embeddings. As mentionned in this paper, the use of this criterion appears as the general consensus in dimension reduction, a field in which building meaningful criteria is tedious. The criterion measures the rescaled average agreement between the K-ary neighbourhoods in the input and output spaces. It is constructed as follows.

We first define the following quantities for $(i,j) \in [n]^2$, $\rho_{ij} =\mid \{k : ||\boldsymbol{X}_i - \boldsymbol{X}_k||_2^2 < ||\boldsymbol{X}_i - \boldsymbol{X}_j||_2^2\} \mid, r_{ij} =\mid \{k : ||\boldsymbol{Z}_i - \boldsymbol{Z}_k||_2^2 < ||\boldsymbol{Z}_i - \boldsymbol{Z}_j||_2^2\} \mid, \nu_i^K = \{j : 1 \leq \rho_{ij} \leq K\}$ and $\gamma_i^K = \{j : 1 \leq r_{ij} \leq K\}$. The average K-ary neighbourhood preservation is rescaled to indicate the improvement over a random embedding such that:

$$R_n(K) = \frac{(n-1)Q_n(K) - K}{n-1-K} \tag{23}$$

where $Q_n(K) = \frac{1}{Kn}\sum_{i=1}^n \mid \nu_i^K \cap \gamma_i^K \mid$, n is the number of data points and K is the hyperparameter that adjusts the scale at which we are looking.

To focus on large-scale structure, K was chosen as either n/4 or n/2. As summarized by [21], current practice consists in using PCA or Laplacian eigenmaps as initialization for these

algorithms, thus we compare to these strategies. Results are displayed in table 2 and table 3, each entry being an average over 5 random seeds, with standard deviation displayed below each entry. Note that when not specified, tSNE and UMAP are initialized with an isotropic Gaussian variable.

These results show that using *ccPCA* is a reliable alternative to PCA and Laplacian eigenmaps for reproducing large-scale neighborhoods.

Table 2: $100 \times R_n(K)$ (23) for embeddings produced using t-SNE with various initializations.

	tSNE		PCA + tSNE		LE + tSNE		ccPCA + tSNE	
K of K-ary	n/4	n/2	n/4	n/2	n/4	n/2	n/4	n/2
MNIST	18.7 ±2.2	7.4 ±5.1	28.4 ±0.3	21.9 ±0.2	26.7 ±0.7	18.5 ±0.4	**31.3** ±0.4	**28.5** ±1.2
cifar-10	20.3 ±3.2	16.4 ±4.8	**36.9** ±0.6	41.9 ±1.1	25.8 ±0.6	24.1 ±1.5	36.4 ±0.4	**43.4** ±1.6
cifar-100	21.6 ±3.6	18.2 ±5.5	38.1 ±0.4	**47.5** ±0.4	23.3 ±1.5	26.5 ±1.8	**39.6** ±0.7	43.6 ±1.1
fashion-MNIST	27.2 ±4.3	12.3 ±7.8	36.9 ±0.1	28.5 ±0.2	32.0 ±0.8	25.1 ±2.2	**41.6** ±0.9	**35.7** ±1.5
Single Cell data	25.7 ±4.8	22.4 ±10.6	37.7 ±2.7	29.0 ±4.7	28.1 ±1.5	31.5 ±1.4	**40.1** ±1.7	**34.6** ±2.6

Table 3: $100 \times R_n(K)$ (23) for embeddings produced using UMAP with various initializations.

	UMAP		PCA + UMAP		LE + UMAP		ccPCA + UMAP	
K of K-ary	n/4	n/2	n/4	n/2	n/4	n/2	n/4	n/2
MNIST	29.5 ±1.4	22.7 ±2.2	**36.6** ±0.2	31.1 ±0.5	34.6 ±0.2	24.9 ±0.7	33.4 ±0.3	**32.3** ±0.5
cifar-10	39.2 ±2.6	47.6 ±1.1	44.3 ±0.2	**53.4** ±0.1	44.2 ±0.1	52.6 ±0.2	**44.6** ±0.2	53.2 ±0.2
cifar-100	41.6 ±1.8	42.2 ±0.9	45.4 ±0.2	45.2 ±0.1	44.2 ±0.3	43.4 ±0.1	**49.9** ±0.4	**52.9** ±0.6
fashion-MNIST	48.7 ±2.6	33.6 ±9.5	56.2 ±0.5	54.3 ±0.6	58.1 ±0.5	53.4 ±0.6	**58.9** ±0.5	**55.8** ±0.3
Single Cell data	39.5 ±1.4	34.3 ±6.1	52.3 ±0.8	47.2 ±6.9	**55.9** ±0.3	45.7 ±0.9	53.6 ±0.3	**53.9** ±1.3