Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–12

# Towards One Model for Classical Dimensionality Reduction: A Probabilistic Perspective on UMAP and t-SNE

**Aditya Ravuri** * and **Neil D. Lawrence**
*University of Cambridge*

## Abstract

This paper shows that dimensionality reduction methods such as UMAP and t-SNE, can be approximately recast as MAP inference methods corresponding to a model introduced in Ravuri et al. (2023), that describes the graph Laplacian (an estimate of the data precision matrix) using a Wishart distribution, with a mean given by a non-linear covariance function evaluated on the latents. This interpretation offers deeper theoretical and semantic insights into such algorithms, and forging a connection to Gaussian process latent variable models by showing that well-known kernels can be used to describe covariances implied by graph Laplacians. We also introduce tools with which similar dimensionality reduction methods can be studied.

## 1. Introduction

In domains with complex, high-dimensional data, such as single-cell biology, dimensionality reduction (DR) algorithms are essential tools for uncovering the underlying structure of data. These algorithms, which include very widely-used techniques like t-SNE (van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2020), are especially valuable for visualizing data manifolds, enabling downstream processing, and the discovery of insightful clusters and trajectories. A deeper understanding of these algorithms and their theoretical underpinnings is crucial for advancing their applicability (particularly when prior information is available) and improving their interpretability. Our work builds on and aims to unify the Probabilistic Dimensionality Reduction (ProbDR) framework Ravuri et al. (2023), which interprets classical DR methods through a probabilistic lens to enable the communication of assumptions, integration of prior knowledge, and model extension for new applications.

Ravuri et al. (2023) introduced ProbDR as a framework with two main interpretations:

1. UMAP and t-SNE having a variational interpretations, describing inference over a nearest neighbour adjacency matrix, and,
2. other classical eigendecomposition-based DR algorithms as inference algorithms of a Wishart model that describes a covariance/precision matrix with a linear kernel evaluated on the latents.

**Our contribution:** In this work, we simplify the framework, moving away from the variational interpretations and propose that **all algorithms** with ProbDR interpretations (and hence most classical DR methods) in the large-$n$ limit can be written as MAP inference algorithms given the model [1],

$$\mathbf{S}|\mathbf{X} \sim \mathcal{W}^{(-1)}(\mathbf{X}\mathbf{X}^T + \alpha\mathbf{H}K_t(\mathbf{X})\mathbf{H} + \gamma\mathbf{I}, \nu), \tag{1}$$

* aditya.ravuri@cl.cam.ac.uk

1. Note that we assume an improper prior on the latents, $p(\mathbf{X}) \propto 1$, throughout this work.

where $\mathbf{S} \in S_n^+$ is an estimate of a covariance matrix generated using the high-dimensional data $\mathbf{Y} \in \mathbb{R}^{n,d}$, $\mathbf{X} \in \mathbb{R}^{n,q}$ corresponds to the set of low ($q$-)dimensional latent variables, and $K$ is a covariance function used to construct a positive-definite matrix constructed using latent variables. This enables a direct comparison between algorithms such as Laplacian Eigenmaps, t-SNE and UMAP.[2]

The rest of the paper will take the following structure. In section 2 we outline relevant background work that we build on in developing the unified framework which we derive in section 3. In section 4, we discuss the significance of this framework in the ability it affords in comparing UMAP and t-SNE based on differences in kernel type. We also present empirical results and connections to Gaussian process latent variable models Lawrence (2005). Section 5 concludes the paper.

## 2. Background

We will now discuss related background work which we build on in developing the unified framework in the next section. Our main objective is to simplify the ProbDR framework by moving away from its variational interpretations[3] and to instead consider algorithms such as t-SNE and UMAP as MAP algorithms. If a Wishart model was found, inference wherein approximated t-SNE and UMAP, assumptions within that model could be compared to the model that leads to traditional eigendecomposition-based methods such as Laplacian Eigenmaps (Belkin and Niyogi, 2001).

For this work, we need to recap only the eigendecomposition view of Ravuri et al. (2023), and we do so with **Laplacian Eigenmaps** as an example: the algorithm involves the calculation of a nearest neighbour graph using the high-dimensional datapoints $\mathbf{Y}$, which can be represented using a graph Laplacian $\hat{\mathbf{L}} = \mathbf{D} - \mathbf{A}$ (where $\mathbf{A}$ is the corresponding adjacency matrix, and $\mathbf{D}$ is the diagonal degree matrix, $\mathbf{D}_{ii} = \sum_k \mathbf{A}_{ik}$) and then obtaining the embedding $\mathbf{X}$ as the eigenvectors of $\hat{\mathbf{L}}$ corresponding to the lowest eigenvalues.

ProbDR showed that this corresponds to inference for $\mathbf{X}$ by maximising $\log \mathcal{W}(\nu\hat{\mathbf{L}}|(\mathbf{X}\mathbf{X}^T + \beta\mathbf{I})^{-1}, \nu)$, i.e. maximising the likelihood of,

$$\nu * \hat{\mathbf{L}}|\mathbf{X} \sim \mathcal{W}((\mathbf{X}\mathbf{X}^T + \beta\mathbf{I})^{-1}, \nu), \tag{2}$$

where $\hat{\mathbf{L}}$ is interpreted to be an estimate of a precision matrix (see Ravuri et al. (2023) for more detail).[4] The model is intuitive as the implied covariance ($\hat{\mathbf{L}}^+$) is modelled by a linear covariance function acting on the latents $\mathbf{X}$, which is familiar in models such as (dual probabilistic) PCA and GPLVMs (Lawrence, 2005).

UMAP and t-SNE, in Ravuri et al. (2023), were interpreted in a variational way, i.e. as KL-minimising algorithms acting on the binary adjacency matrix $\mathbf{A}$. Under certain

2. Moreover, certain constructions used in this paper, such as double centered distance matrices and exponentiated kernel matrices may be useful for building covariances in practice.
3. Van Assel et al. (2022) also offer a similar *variational* view on UMAP and t-SNE.
4. Although the proof method there uses results used by Williams and Agakov (2002) (for probabilistic minor components analysis), a similar result can be seen easily. Consider the matrix $\hat{\mathbf{S}} = (\hat{\mathbf{L}} + \epsilon\mathbf{I})^{-1}$. Maximising the likelihood of $d * \hat{\mathbf{S}}|\mathbf{X} \sim \mathcal{W}(\mathbf{X}\mathbf{X}^T + \beta\mathbf{I}, d)$ recovers the same solution as the one above. This can be seen as the major eigenvectors of $\hat{\mathbf{S}}$ are the minor eigenvectors of $\hat{\mathbf{L}}$. This model is simply dual probabilistic PCA (Lawrence, 2005), but with a non-standard covariance estimator, instead of the traditional $\mathbf{Y}\mathbf{Y}^T/d$, written in terms of the covariance, as $\mathbf{Y} \sim \mathcal{MN}(\mathbf{0}, \mathbf{C}, \mathbf{I}) \Rightarrow \mathbf{Y}\mathbf{Y}^T \sim \mathcal{W}(\mathbf{C}, d)$.

circumstances[5], the interpretation becomes equivalent to MAP estimation (due to Ravuri et al. (2023), Appendix B.7, Lemma 13, also presented in Damrich et al. (2022)).

Critically, however, Damrich et al. (2022) note that the optimisation process is equally as important. As part of an extensive study on the nature of the t-SNE and UMAP loss functions, Damrich et al. (2022) then show how the stochastic optimisation of t-SNE and UMAP can be interpreted to be contrastive estimation with the energy function (negative loss),

$$\mathcal{E}(\mathbf{X}) \propto \sum_{ij} \mathbf{A}_{ij} \log\left(\frac{1}{d_{ij}(\mathbf{X})^2+1}\right) + \frac{4n_{\text{neg}}n_{\text{neigh}}}{3n}\sum_{ij}(1-\mathbf{A}_{ij})\log\left(1-\frac{1}{d_{ij}(\mathbf{X})^2+1}\right), \quad (3)$$

where $\mathbf{A}_{ij}$ represents whether data points $\mathbf{Y}_i$ and $\mathbf{Y}_j$ are neighbours, and $d_{ij}^2(\mathbf{X}) = \|\mathbf{X}_{i:} - \mathbf{X}_{j:}\|^2$. We will refer to this as the **CNE objective**. Note that we've made some substitutions in the original formation that appears in Damrich et al. (2022), for example, we set their parameter $\tilde{q}_{ij} = 1/d_{ij}^2$, which corresponds to the UMAP setting (and a full derivation is given in appendix B). The hyperparameter $n_{\text{neg}}$ sets the number of contrastive negatives (set to be five) that affects the strength of repulsion, and $n_{\text{neigh}}$ corresponds to the number of neighbours set for a point (fifteen in this work). In this work, we will work with this loss function and aim to interpret it as a likelihood, but over the latents $\mathbf{X}$.[6]

## 3. Towards a distribution on the graph Laplacian

In this section, we derive the Wishart distribution inference in which leads to UMAP and t-SNE-like algorithms. Eq. 3 is not a likelihood due to the multiplicative constant weighting the contribution of points that are not adjacent. The second term, $\mathcal{T}_b$ can be expressed as follows,

$$\mathcal{T}_b = \sum_{ij}(1-\mathbf{A}_{ij})\log\left(\left[1-\frac{1}{d_{ij}(\mathbf{X})^2+1}\right]^{\tilde{\epsilon}}\right).$$

The implied probability of adjacency $\tilde{\mathbf{p}}_{ij}$, assuming that this is the second term of a Bernoulli likelihood is,

$$\begin{aligned}\tilde{\mathbf{p}}_{ij} &= 1-\left[1-\frac{1}{d_{ij}(\mathbf{X})^2+1}\right]^{\tilde{\epsilon}} = 1-\exp\left[\tilde{\epsilon}\log\left(1-\frac{1}{d_{ij}(\mathbf{X})^2+1}\right)\right] \\ &= 1-\exp\left[-\tilde{\epsilon}\log\left(1+\frac{1}{d_{ij}(\mathbf{X})^2}\right)\right] \approx 1-1+\tilde{\epsilon}\log\left(1+\frac{1}{d_{ij}(\mathbf{X})^2}\right) \qquad \text{large n.}\end{aligned}$$

5. if the variational probabilities are zero or one, signifying if two points are nearest neighbours; this simplification is reasonable due to the findings of Damrich and Hamprecht (2021), where it was found that the relatively complex calculation of the variational probabilities in t-SNE and UMAP can be replaced with simply the adjacency matrices without loss of performance. Our initial experiments closely aligned with these findings.
6. We were particularly inspired to look toward contrastive methods by Nakamura et al. (2023), who showed that contrastive learning methods could be seen as variational algorithms (hence suggesting a link between t-SNE/UMAP and contrastive learning) and by Gutmann and Hyvärinen (2010), which shows that contrastive losses are estimators of negative log-likelihoods. Damrich et al. (2022) offer us a constrastive objective for t-SNE and UMAP, greatly simplify the optimisation process and offer a clear objective to work with, i.e. eq. (3).

Now, we observe that $\log(1+1/x) >= 1/(1+x)$, so,

$$\mathcal{T}_b = \sum_{ij}(1-\mathbf{A}_{ij})\log\left(1-\tilde{\mathbf{p}}_{ij}\right) \leq \sum_{ij}(1-\mathbf{A}_{ij})\log\left(1-\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right).$$

Note that the multiplication of $\tilde{\epsilon}$ within the log in the first term adds just a constant to the first term of eq. (3)). Therefore,

$$\mathcal{E}(\mathbf{X}) \leq \sum_{ij}\mathbf{A}_{ij}\log\left(\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right) + \sum_{ij}(1-\mathbf{A}_{ij})\log\left(1-\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right) + c.$$

We conclude that the CNE objective lower bounds the Bernoulli likelihood implied by the model,

$$\mathbf{A}_{ij}|\mathbf{X} \sim \text{Bernoulli}\left(\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right). \tag{4}$$

Although this is a valid probabilistic interpretation, we will go slightly further and outline an argument that makes this interpretation comparable to the Wishart interpretations in ProbDR.[7] As Wisharts have supports over positive definite matrices, we will try to reconsider this model with the graph Laplacian $\mathbf{L}$ as the observed statistic. The likelihood for $\mathbf{X}$ implied by eq. (4) is,

$$\begin{aligned}
\log p(\mathbf{X}|\mathbf{A}) &= \sum_{ij}\mathbf{A}_{ij}\log\left(\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right) + \sum_{ij}(1-\mathbf{A}_{ij})\log\left(1-\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right) \\
&\approx \sum_{ij}\mathbf{A}_{ij}\log\left(\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right) + \sum_{ij}\log\left(1-\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}\right) && \text{large n} \\
&\approx \text{tr}(\mathbf{AP}') - \sum_{ij}\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2} && \text{small } \tilde{\epsilon}
\end{aligned}$$

$$\Rightarrow \log p(\mathbf{X}|\mathbf{L}) = -\text{tr}(\mathbf{LHP}'\mathbf{H}) - \sum_{ij}\tilde{\epsilon}\frac{1}{1+d_{ij}(\mathbf{X})^2}, \qquad \text{centd. } \mathbf{L} \text{ and } \text{tr}(\mathbf{DP}') = 0$$

where, $\mathbf{P}^u_{ij} = 1/(1+d^2_{ij})$, $\mathbf{P}_{ij} = \tilde{\epsilon}\mathbf{P}^u_{ij}$, $\mathbf{P}'_{ij} = \log \mathbf{P}_{ij}$[8] and $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/n$ is a centering matrix. Note that $\text{tr}(\mathbf{HMH}) = n - \sum_{ij}\mathbf{M}_{ij}/n$. Therefore[9],

$$\log p(\mathbf{X}|\mathbf{L}) = -\text{tr}(\mathbf{LHP}'\mathbf{H}) + n\text{tr}(\mathbf{HPH}) + k \approx -\text{tr}(\mathbf{LHP}'\mathbf{H}) + n\log|\mathbf{I}+\mathbf{HPH}| + k. \quad \text{small } \tilde{\epsilon}$$

7. This is reasonable as exponential families share similar likelihood forms, and a Wishart interpretation, despite being over discrete matrices, may correspond to a similar statistical learning scenario as performing classification using linear regression (i.e. using the $L^2$ norm serparator).
8. $\mathbf{P}'$ is conditionally positive definite (CPD) as $\mathbf{P}_{ij}$ defines the student-t kernel, which enforces PDness and an elementwise log of a PD matrix with all positive elements will at least be CPD (Bhatia (2007)). Double centering such matrices makes them PSD.
9. Note that our derivation also produces a similar objective to the DK-LLE objective of Draganov and Dohn (2023) (Lemma 4), which was found by gradient-based arguments, adding credibility to our arguments.

We approximate $\log \mathbf{P}_{ij}$ within the vicinity of $\tilde{\epsilon}$[10] by matching the gradient and function value using an ansatz,

$$\begin{aligned}
\log \mathbf{P}_{ij} &\approx \frac{\mathbf{P}_{ij}}{2\tilde{\epsilon}} - \frac{\tilde{\epsilon}}{2\mathbf{P}_{ij}} + \log \tilde{\epsilon}. \\
\Rightarrow \log p(\mathbf{X}|\mathbf{L}) &\approx -\text{tr}(\mathbf{L}\mathbf{H}(0.5\mathbf{P}^u - 0.5[1/\mathbf{P}^u]_{ij})\mathbf{H}) + n\log|\mathbf{I} + \mathbf{H}\mathbf{P}\mathbf{H}| + c \\
&= -\text{tr}(\mathbf{L}(0.5\tilde{\epsilon}^{-1}\mathbf{I} + 0.5\mathbf{H}\mathbf{P}^u\mathbf{H} + \mathbf{X}\mathbf{X}^T)) + n\log|0.5\tilde{\epsilon}^{-1}\mathbf{I} + 0.5\mathbf{H}\mathbf{P}^u\mathbf{H}| + k \\
&\leq \log \mathcal{W}(\mathbf{L}|(0.5\tilde{\epsilon}^{-1}\mathbf{I} + 0.5\mathbf{H}\mathbf{P}^u\mathbf{H} + \mathbf{X}\mathbf{X}^T)^{-1}, n).
\end{aligned}$$

Therefore, the model below approximates the model in eq. (4).

$$\mathbf{L}|\mathbf{X} \sim \mathcal{W}\left((0.5\tilde{\epsilon}^{-1}\mathbf{I} + 0.5\mathbf{H}\mathbf{P}^u\mathbf{H} + \mathbf{X}\mathbf{X}^T)^{-1}, n\right) \tag{5}$$

## 4. Results and Discussion

In this paper, we've presented an interpretation for UMAP and t-SNE-like algorithms and proposed a comparable distributional assumption to ProbDR's. Our model for UMAP and t-SNE-like algorithms uses a covariance based on a double-centred non-linear kernel[11] however, as opposed to the linear kernel used as part of ProbDR's Wishart-based algorithms. Our work provides shows a direct connection to the model behind Laplacian Eigenmaps; the model of eq. (5) is a non-linear extension (albeit with an interesting choice of kernel scales) of ProbDR's Laplacian Eigenmaps with a non-linear kernel in eq. (2).

fig. 1 shows a comparison between embeddings found with our interpretations compared with those presented in Damrich et al. (2022) within different datasets (found within the openTSNE repositories (Poličar et al., 2024)). We resample each dataset such that it has exactly 10 groups, with up to 25k points in total. The optimisation was done using a small-range GPU capable of inverting matrices of 25k rows/columns, using pytorch's Adam optimiser, with an initial learning rate set to 1.0, with each experiment being run for 50 epochs. We see that the embeddings recovered using our methods are qualitatively quite similar to that of CNE and not PCA/GPLVM.

The Wishart model of eq. (5) is quite elegant, as it implies that the data covariance is modelled by a covariance function, connecting Gaussian process latent variable models Lawrence (2005) (that use kernels not unlike the Student-t/Cauchy kernel to model the covariance) to ProbDR models that describe the graph Laplacian (which describes a precision matrix). Moreover, it is interesting to note that the implied covariance of our model (the inverse of the Wishart parameter) is non-stationary and can be justified by the fact that the adjacency probabilities go to zero as a function of distance. The semantic, probabilistic, and modelling implications behind other assumptions (such as the high noise-scale parameter within the Wishart parameter) made within our simplified ProbDR remain an area that can

10. This neighbourhood was chosen as $\tilde{\epsilon}$ is the maximal value of $\mathbf{P}_{ij}$.
11. Other empirical arguments can also be made to show that, assuming a model such as: $\mathbf{X} \to \mathbf{Y} \to \mathbf{A}$, where $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}$, the adjacency probabilities of eq. (4) cannot be found using simply a linear kernel, and **can** be found using a non-linear kernel such as a Student-t kernel, as we've used here. For this, one can write out the approximate distribution on distances (assuming a Gaussian process prior on $\mathbf{Y}$, $\mathbb{E}(d^2_{ij}(\mathbf{Y})) = d * (k_{ii} + k_{jj} - 2k_{ij})$ and $\text{Cov}(d^2_{ij}, d^2_{mn}) = 2d * (k_{im} + k_{jn} - k_{in} - k_{jm})^2$) and study the probability with which such distances attain extreme values, as a function of latent distance.

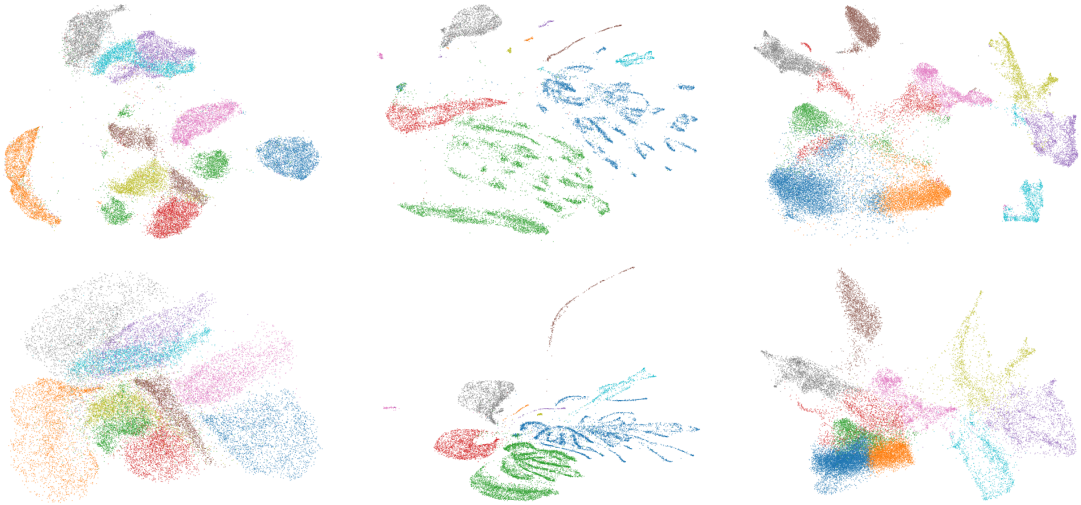

Figure 1: Comparison between embeddings obtained using the CNE objective (**top**) and our inference (**bottom**). From left to right: MNIST digits, transcriptomic data from Macosko et al. (2015), and larger-scale transcriptomic data from Zheng et al. (2017). appendix C shows that PCA and GPLVMs with a similar kernel do not produce similar embeddings, presumably due to the Laplacian encoding different statistics of the data w.r.t. the empirical covariance.

be studied. In addition to these insights, we show non-traditional ways in which covariances can be constructed as part of Gaussian process models and/or dimensionality reduction methods, which revolve around double-centering CPD matrices (which can be based on distance matrices[12] or element-wise exponentiated kernel matrices[13]).

## 5. Conclusion

In conclusion, this study presents a novel theoretical framework that reinterprets UMAP and t-SNE-like algorithms as maximum a posteriori (MAP) inference algorithms within a model for a graph Laplacian described by a Wishart distribution. This result bridges the gap between popular dimensionality reduction methods with Gaussian process latent variable models, enhancing our understanding of their mechanisms. The insights gained show the importance of specific modelling assumptions in optimizing these algorithms' effectiveness and interpretability, setting a foundation for further research on model specifications and their practical implications. Next steps include studying whether eigenfunctions of RBF-like

12. For example, take $-\mathbf{D} = \log \frac{1}{1+\|\mathbf{X}_i - \mathbf{X}_j\|^2}$, where $\mathbf{D}$ represents a distance matrix. The inner term is a kernel, hence is PSD. The log function (like the square root, in this context) preserves CPSD-ness (Bhatia, 2007). Therefore, $-\mathbf{HDH}$ is PSD. It's also interesting to observe, due to the results of Khare (2019); Schoenberg (1935), that there exists an isometric Euclidean embedding for such a distance matrix (and vice versa).
13. with fractional powers.

kernels (which roughly behave as $\phi_k(x) \sim \gamma \cos(2x\kappa - k\pi/2)$) can aid the characterization of solutions of ProbDR models.

## Acknowledgements

AR would like to thank Diana Robinson and Francisco Vargas for helpful discussions, and the Accelerate Programme for Scientific Discovery for a studentship funding this work.

## References

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/f106b7f99d2cb30c3db1c3cc0fde9ccb-Paper.pdf.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007. ISBN 9780691129181. URL http://www.jstor.org/stable/j.ctt7rxv2.

Sebastian Damrich and Fred A Hamprecht. On umap's true loss function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5798–5809. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

Sebastian Damrich, Niklas Böhm, Fred A Hamprecht, and Dmitry Kobak. From $t$-sne to umap with contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Andrew Draganov and Simon Dohn. Unexplainable explanations: Towards interpreting tsne and umap embeddings. *arXiv preprint*, 06 2023. doi: 10.48550/arXiv.2306.11898. URL http://xkdd2023.isti.cnr.it/papers/286.pdf.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

Apoorva Khare. Schoenberg: from metric geometry to matrix positivity, Apr 2019. URL https://math.iisc.ac.in/seminar-slides/2019/2019-04-12-ApoorvaKhare.pdf.

Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, dec 2005. ISSN 1532-4435.

Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A.

McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2015.05.002. URL https://www.sciencedirect.com/science/article/pii/S0092867415005498.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.

Hiroki Nakamura, Masashi Okada, and Tadahiro Taniguchi. Representation uncertainty in self-supervised learning as variational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16484–16493, 2023.

Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. opentsne: A modular python library for t-sne dimensionality reduction and embedding. *Journal of Statistical Software*, 109(3): 1–30, 2024. doi: 10.18637/jss.v109.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v109i03.

Aditya Ravuri, Francisco Vargas, Vidhi Lalchand, and Neil D Lawrence. Dimensionality reduction as probabilistic inference. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023. URL https://arxiv.org/pdf/2304.07658.pdf.

I. J. Schoenberg. Remarks to maurice frechet's article "sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de hilbert. *Annals of Mathematics*, 36(3):724–732, 1935. ISSN 0003486X. URL http://www.jstor.org/stable/1968654.

Hugues Van Assel, Thibault Espinasse, Julien Chiquet, and Franck Picard. A probabilistic graph coupling view of dimension reduction. *Advances in Neural Information Processing Systems*, 35:10696–10708, 2022.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Christopher K. I. Williams and Felix V. Agakov. Products of Gaussians and Probabilistic Minor Component Analysis. *Neural Computation*, 14(5):1169–1182, 05 2002. ISSN 0899-7667. doi: 10.1162/089976602753633439. URL https://doi.org/10.1162/089976602753633439.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, Jan 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. URL https://doi.org/10.1038/ncomms14049.

## Appendix A. Approximate distribution of a distance matrix

**Theorem 1 (Distribution of normal distances)** *Assume that* $\mathbf{Y}$ *is distributed as,*

$$\begin{bmatrix}\mathbf{Y}_i\\ \mathbf{Y}_j\end{bmatrix} \sim \mathcal{MN}\left(\mu, \begin{bmatrix}k_{ii} & k_{ij}\\ k_{ji} & k_{jj}\end{bmatrix}, \mathbf{I}_d\right).$$

*Then, the following hold. Firstly, denoting* $d_{ij}^2 = \|\mathbf{Y}_i - \mathbf{Y}_j\|^2$*, the marginal distribution is given by,*

$$d_{ij}^2 \sim \Gamma\left(k = \frac{d}{2}, \theta = 2(k_{ii} + k_{jj} - 2k_{ij})\right).$$

*As a consequence,* $\mathbb{E}(d_{ij}^2) = d * \tilde{k}_{ij}$ *and* $\mathbb{V}(d_{ij}^2) = 2d * \tilde{k}_{ij}^2$*, where* $\tilde{k}_{ij} = k_{ii} + k_{jj} - 2k_{ij}$*. Additionally,*

$$\mathbb{C}(d_{ij}^2, d_{mn}^2) = 2d * (k_{im} + k_{jn} - k_{in} - k_{jm})^2.$$

*This is a useful fact as the upper triangle of the distance matrix is approximately normal due to the central limit theorem with increasing* $d$*.*

The first part of the theorem is given in ProbDR (Ravuri et al. 2023), reproduced below.

$$\begin{aligned}
\forall k : d'_{ij} \equiv y_i^k - y_j^k \sim \mathcal{N}(0, k_{ii} + k_{jj} - 2k_{ij}) &\overset{d}{=} \sqrt{k_{ii} + k_{jj} - 2k_{ij}}Z\\
\Rightarrow d_{ij}^2 \equiv \|\mathbf{Y}_i - \mathbf{Y}_j\|^2 = \sum_k^d (y_i^k - y_j^k)^2 &\overset{d}{=} (k_{ii} + k_{jj} - 2k_{ij})\sum_k^d Z_k^2\\
&\overset{d}{=} (k_{ii} + k_{jj} - 2k_{ij})\chi_d^2\\
&\overset{d}{=} \Gamma(k = d/2, \theta = 2(k_{ii} + k_{jj} - 2k_{ij})).
\end{aligned}$$

The covariance between $d_{ij}^2$ and $d_{mn}^2$ can be computed as follows. Let,

$$d'_{ij} = \mathbf{Y}_{id} - \mathbf{Y}_{jd} \text{ and } d'_{mn} = \mathbf{Y}_{md} - \mathbf{Y}_{nd}.$$

We can then derive some important moments as follows,

$$\mathbb{E}(d'_{ij}) = 0, \mathbb{V}(d'_{ij}) = \mathbb{E}(d'^2_{ij}) = k_{ii} + k_{jj} - 2k_{ij},$$

$$\begin{aligned}
\mathbb{C}(d'_{ij}, d'_{mn}) &= \mathbb{C}(Y_{id} - Y_{jd}, Y_{md} - Y_{nd})\\
&= \mathbb{C}(Y_{id}, Y_{md}) - \mathbb{C}(Y_{id}, Y_{nd}) - \mathbb{C}(Y_{jd}, Y_{md}) + \mathbb{C}(Y_{jd}, Y_{nd})\\
&= k_{im} + k_{jn} - k_{in} - k_{jm},
\end{aligned}$$

Then,

$$
\begin{aligned}
\mathbb{C}(d_{ij}^2, d_{mn}^2) &= \mathbb{C}\left(\sum_{d_1}(\mathbf{Y}_{id_1} - \mathbf{Y}_{jd_1})^2, \sum_{d_2}(\mathbf{Y}_{md_2} - \mathbf{Y}_{nd_2})^2\right) \\
&= \sum_{d_1}\sum_{d_2}\mathbb{C}((\mathbf{Y}_{id_1} - \mathbf{Y}_{jd_1})^2, (\mathbf{Y}_{md_2} - \mathbf{Y}_{nd_2})^2) && \text{linearity} \\
&= \sum_{d}\mathbb{C}(d_{ij}^{'2}, d_{mn}^{'2}) && \text{independence} \\
&= d * \mathbb{C}(d_{ij}^{'2}, d_{mn}^{'2}) \\
&= d * \left[\mathbb{E}[d_{ij}^{'2}d_{mn}^{'2}] - \mathbb{E}[d_{ij}^{'2}]\mathbb{E}[d_{mn}^{'2}]\right] \\
&= d * \left[\mathbb{E}(d_{ij}^{'2})\mathbb{E}(d_{mn}^{'2}) + 2\mathbb{E}^2(d'_{ij}d'_{mn}) - \mathbb{E}[d_{ij}^{'2}]\mathbb{E}[d_{mn}^{'2}]\right] && \text{Isserlis' theorem} \\
&= 2d * \mathbb{E}^2(d'_{ij}d'_{mn}) \\
&= 2d * (k_{im} + k_{jn} - k_{in} - k_{jm})^2.
\end{aligned}
$$

## Appendix B. Derivation of eq. (3)

Eqn. 8 of Damrich et al. (2022) reads,

$$\mathcal{L}^{\mathrm{NEG}}(\theta) = -\mathbb{E}_{x\sim p}\log\left(\frac{q_\theta(x)}{q_\theta(x)+1}\right) - m\mathbb{E}_{x\sim\xi}\log\left(1-\frac{q_\theta(x)}{q_\theta(x)+1}\right).$$

We use the notation $n_{\mathrm{neg}} = m$, and we consider the objective (resulting in eq. (3)) in terms of the negative loss $\mathcal{E} = -\mathcal{L}$. Next, Lemma 3 of Damrich et al. (2022) specifies the choice of $q_\theta$ corresponding to UMAP to be $q_\theta(x) = 1/d^2_{ij}(\mathbf{X})$. The objective simplifies to,

$$\begin{aligned}\mathcal{E}(\mathbf{X}) &= \frac{1}{\sum_{i>j}\mathbf{A}_{ij}}\sum_{i>j}\mathbf{A}_{ij}\log\left(\frac{1}{1+d^2_{ij}(\mathbf{X})}\right) + \frac{n_{\mathrm{neg}}}{\sum_{i>j}(1-\mathbf{A}_{ij})}\sum_{i>j}(1-\mathbf{A}_{ij})\log\left(1-\frac{1}{1+d^2_{ij}(\mathbf{X})}\right)\\ &\propto \sum_{i>j}\mathbf{A}_{ij}\log\left(\frac{1}{1+d^2_{ij}(\mathbf{X})}\right) + \frac{n_{\mathrm{neg}}\sum_{i>j}\mathbf{A}_{ij}}{\sum_{i>j}(1-\mathbf{A}_{ij})}\sum_{i>j}(1-\mathbf{A}_{ij})\log\left(1-\frac{1}{1+d^2_{ij}(\mathbf{X})}\right)\end{aligned}$$

The multiplicative constant is approximated as,

$$\tilde{n} \equiv \frac{n_{\mathrm{neg}}\sum_{i>j}\mathbf{A}_{ij}}{\sum_{i>j}1-\mathbf{A}_{ij}} \approx \frac{n * n_{\mathrm{neg}} * n_{\mathrm{neigh}}/1.5}{(n^2-n)/2} \approx \frac{4n_{\mathrm{neg}}n_{\mathrm{neigh}}}{3n}.$$

Therefore, the objective becomes,

$$\begin{aligned}\mathcal{E}(\mathbf{X}) &\approx \sum_{i>j}\mathbf{A}_{ij}\log\left(\frac{1}{1+d^2_{ij}(\mathbf{X})}\right) + \frac{4n_{\mathrm{neg}}n_{\mathrm{neigh}}}{3n}\sum_{i>j}(1-\mathbf{A}_{ij})\log\left(1-\frac{1}{1+d^2_{ij}(\mathbf{X})}\right)\\ &\propto \sum_{ij}\mathbf{A}_{ij}\log\left(\frac{1}{1+d^2_{ij}(\mathbf{X})}\right) + \frac{4n_{\mathrm{neg}}n_{\mathrm{neigh}}}{3n}\sum_{ij}(1-\mathbf{A}_{ij})\log\left(1-\frac{1}{1+d^2_{ij}(\mathbf{X})}\right),\end{aligned}$$

which is eq. (3).

## Appendix C. Comparison with GPLVMs

Figure 2: MNIST digits embedded using PCA (**left**), GPLVM using a `linear + constant + t + noise` kernels, with the inits scaled towards zero (**center**), and the same GPLVM with unscaled PCA inits (**right**). In each case, the GPLVM hyperparameters were first "pre-trained" using the PCA-initialized embeddings for 10 epochs, and the embeddings were trained for a further 40. In every case, note the visual dissimilarity w.r.t. our versions of UMAP/t-SNE.