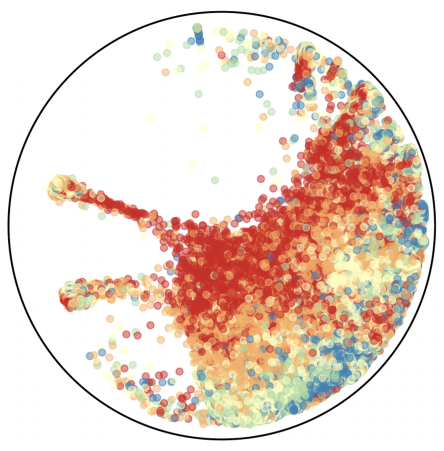
Dimensionality Reduction: a Probabilistic Perspective

Aditya Ravuri Neil D. Lawrence
UNIVERSITY OF CAMBRIDGE

Background

Dimensionality Reduction (DR)

Premise: high dimensional data sitting on low dimensional manifold.



gene expression projected into a Poinaré disk showing cell differentiation

Often a need for:

- uncertainty quantification
- incorporation of prior knowledge (e.g. confounders)
- identifiability and interpretability

Existing Interpretations

PCA, **FA**, **ICA**, **CMDS** algorithms correspond to MLE in

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \epsilon \tag{1}$$
$$\Leftrightarrow \mathbf{Y}|\mathbf{X} \sim \mathcal{MN}(\mathbf{0}, \mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I}, \mathbf{I}) \tag{2}$$
$$\Leftrightarrow \mathbf{Y}|\mathbf{W} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}), \tag{3}$$

with different priors (and hence implications) on $\mathbf{X}$,

- isotropic Gaussian $\leftrightarrow$ maximal variance
- non-isotropic $\leftrightarrow$ identifiability
- non-Gaussian $\leftrightarrow$ causation

Are there probabilistic interpretations to other DR methods? ... Yes!

Case Study A: Eigencomponent based DR

1. Non-constant likelihood terms of eq. (1) are equivalent in,
$$\mathbf{S}|(\cdot) \sim \mathcal{W}((\cdot)(\cdot)^T + \sigma^2\mathbf{I}, d), \tag{4}$$
where $\mathbf{S} \equiv \mathbf{Y}\mathbf{Y}^T$ or $\mathbf{Y}^T\mathbf{Y}$ is the sample covariance, **an estimate** and $(\cdot) = \mathbf{X}$ or $\mathbf{W}$.
2. Eigencomponents result in MLEs for latents in such models.

EC	interpretation	Cov($\mathbf{S}$|$\mathbf{X}$)	Cov($\mathbf{S}$|$\mathbf{W}$)
Major	data linear in latent	$\mathbf{X}\mathbf{X}^\mathbf{T} + \sigma^2\mathbf{I}$	$\mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}$
Minor	incidence linear in latent	$(\mathbf{X}\mathbf{X}^\mathbf{T} + \beta\mathbf{I})^{-1}$	$(\mathbf{W}\mathbf{W}^\mathbf{T} + \beta\mathbf{I})^{-1}$

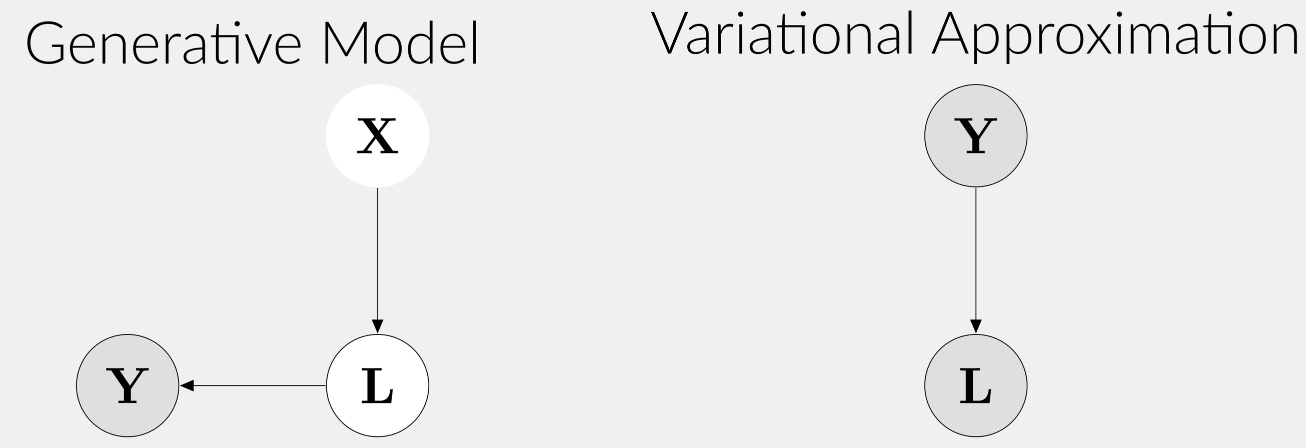
Our interpretation: DR algo.s outputting embeddings as eigencomponents of PSD matrices are two step inference processes:

1. estimate a covariance $\hat{\mathbf{S}}$ or graph Laplacian(/precision) $\hat{\mathbf{L}}$,
2. obtain emedding through MLE in the class of models above (minor eig.s if precision, major if covariance).

List of our interpretations:

- Laplacian Eigenmaps, Spectral Embeddings:
$$\hat{\mathbf{L}}(\mathbf{Y})_{ij} = -k(\mathbf{Y}_i, \mathbf{Y}_j)$$
- LLE:
$$\hat{\mathbf{L}}(\mathbf{Y}) = \text{MLE in a GMRF/Matérn-1 graph GP}$$
- Kernel PCA:
$$\hat{\mathbf{S}}(\mathbf{Y}) = k(\mathbf{Y}_i, \mathbf{Y}_j)$$
- Diffusion Maps ($\alpha = 1$): (this is a Matérn-∞ graph GP covariance)
$$\hat{\mathbf{S}}(\mathbf{Y}) = \exp[-t\mathbf{\Delta}(\mathbf{Y})]$$

Case Study B: VI based DR



a class of models whose ELBOs give rise to (t-)SNE and UMAP objectives

(t-)SNE & UMAP are variational inference (VI) algorithms in graph Laplacian estimation models, potentially with a generative model.

ELBOs here take the form `constant - cost_func_of_algo`:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{L}|\mathbf{Y})}[\log p(\mathbf{Y}|\mathbf{L})] - \text{KL}(q(\mathbf{L}|\mathbf{Y})||p(\mathbf{L}|\mathbf{X}))$$

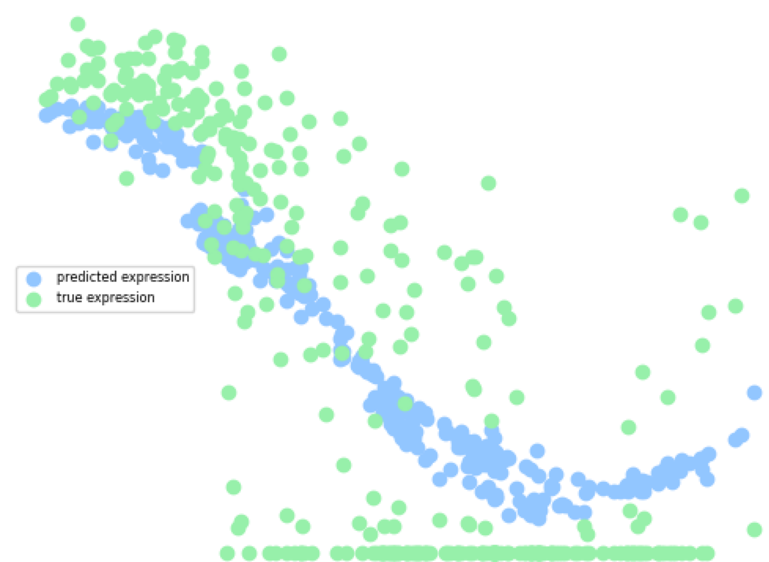
The variational & model probabilities $q(\mathbf{A}|\mathbf{Y}), p(\mathbf{A}|\mathbf{X})$ factorize as,

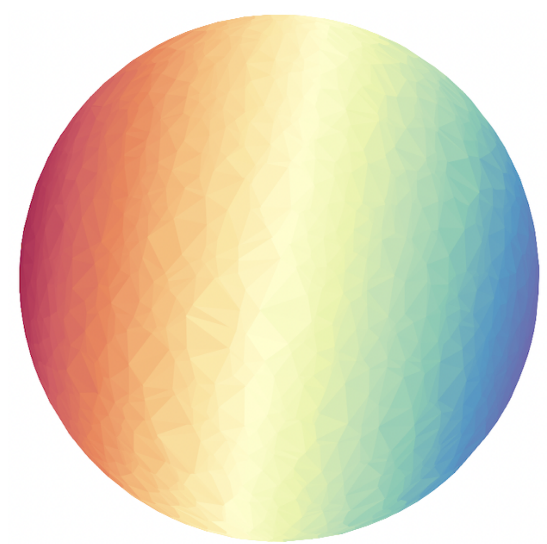
$$\begin{cases} \prod_{i\neq j}^n \text{Bernoulli}(\mathbf{A}_{ij}|\{v \text{ or } w\}_{ij}^U) & \text{if UMAP} \\ \prod_i^n \text{Categorical}(\mathbf{A}_{i:}|\{v \text{ or } w\}_{ij}^S) & \text{if (t-)SNE} \end{cases}.$$

Methods in literature approximating components above using neural networks are simple extensions.

Graph GP generative models on $\mathbf{L} \to \mathbf{Y}$ can outperform GPLVMs, VAEs.

Applications & Extensions

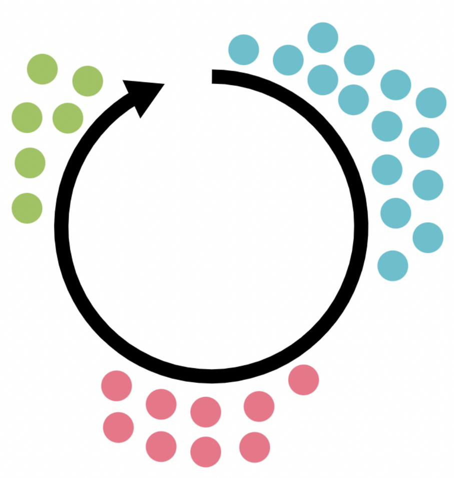




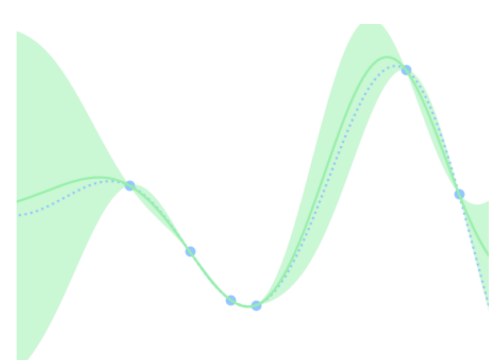
left: predicted gene expression given latents using a graph GP generative model, difficult to obtain without a generative model **right:** illustration of convergence of eigenvectors of $\mathbf{L}$ to eigenfunctions of $\mathbf{\Delta}$ - thus hinting at a convergence of $\mathbf{X} \to \mathbf{L} \to \mathbf{Y}$ to a Matérn GP on a manifold, thus GPLVMs.

The Why & Connections to other projects

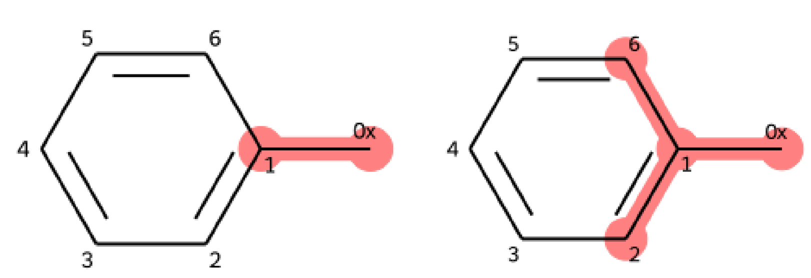
Probabilistic interpretations:



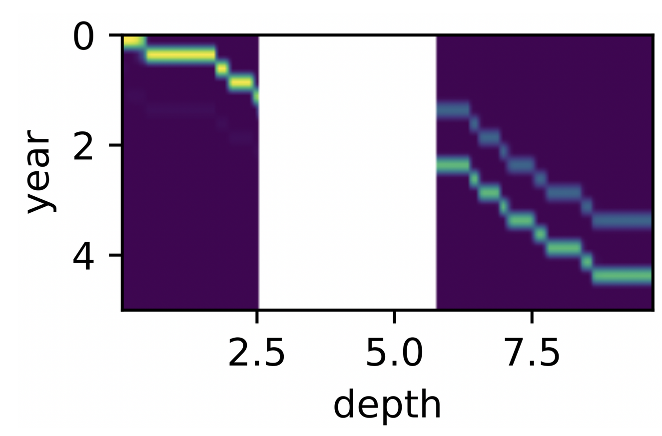
enable efficient reformulations
enable prior specification
e.g. RE models are GPs



aid communication
enable software cross-utilization
e.g. GAMs are GPs



enable reasoning about uncertainties and methods like bayesopt
e.g. graph kernels are deep (gnn) kernels



enable model extension
enable alternative inference methods
e.g. DTW is (cts-)HMM inference