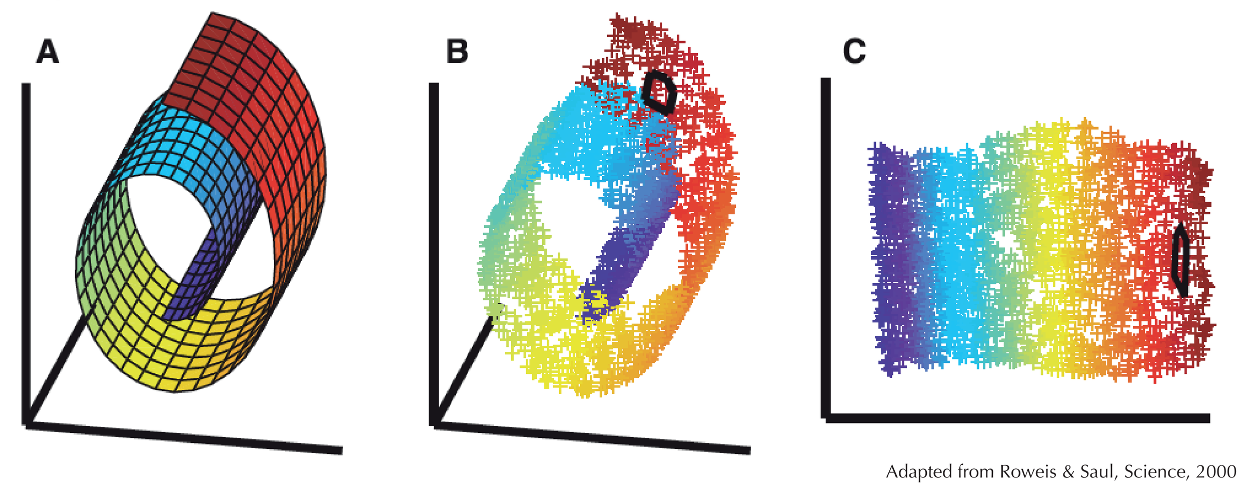
# Bayesian Manifold Learning : Locally Linear Latent Variable Model (LL-LVM)

**Mijung Park, Zoltán Szabó, Ahmad Qamar, Lars Buesing, Maneesh Sahani**
Gatsby Computational Neuroscience Unit, University College London

## Manifold Learning

- Problems with high-dimensional data
  - optimisation in high-d parameter space is computationally expensive and hard to find a global optimum
- Good news: in many cases, the intrinsic dimensionality is actually low
  - datapoints are sampled from a low-dimensional manifold embedded in a high-dimensional space
  - example: swiss roll




- Manifold learning : attempts to uncover the manifold structure

## Non-probabilistic prior work

- idea: preserve geometric properties of local neighbourhoods
- limits:
  - sensitive to noise due to lack of explicit model
  - heuristic methods to evaluate manifold dimensionality
  - no measure of uncertainties in the estimates
  - out-of-sample extension requires extra approximations

## Gaussian process latent variable model (GP-LVM)

- idea: define a functional mapping from latent space to data space using GP [1, 2]
- for data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{d_y}] \in \mathbb{R}^{n \times d_y}$ and latents $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{d_x}] \in \mathbb{R}^{n \times d_x}$,

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{k=1}^{d_y} \mathcal{N}(\mathbf{y}_k|\mathbf{0}, \mathbf{K}_{nn} + \beta^{-1}\mathbf{I}_n),$$

where the $i, j$th element of the covariance matrix is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left[-\tfrac{1}{2}\sum_{q=1}^{d_x} \alpha_q (x_{i,q} - x_{j,q})^2\right],$$

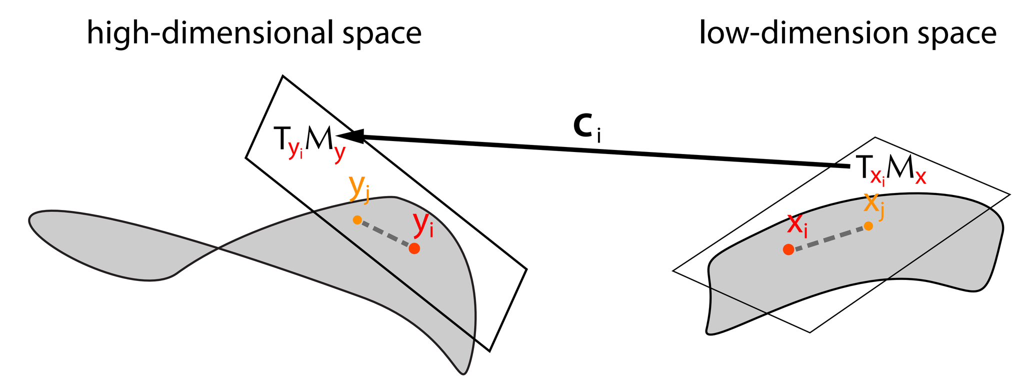
where $\alpha_q$'s determine dimensionality of latent space.
- limits:
  - no intuitive preservation of local neighbourhood properties
  - smoothness of manifold constrained by pre-chosen covariance function
  - auxiliary variable for variational inference (also restricts choice of cov func)

## Question

Can we learn a manifold in a probabilistic and possibly *Bayesian* way, while preserving geometric properties of local neighbourhoods?

## Our approach: LL-LVM

- Key idea: there is a *locally linear* mapping between tangent spaces in low and high dimensional spaces




$$\mathbf{y}_j - \mathbf{y}_i \approx \mathbf{C}_i(\mathbf{x}_j - \mathbf{x}_i)$$

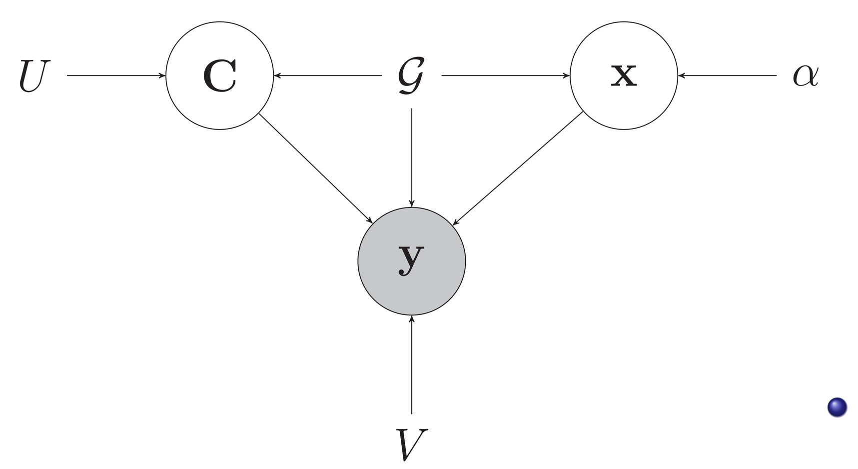
- given a graph $\mathcal{G}$ of neighbours with adjacency indicator $\eta_{ij} = 1$ if $j \in N(i)$, find the distribution over the linear maps $\mathbf{C} = [\mathbf{C}_1, \cdots, \mathbf{C}_n] \in \mathbb{R}^{d_y \times n d_x}$ and the latent variables $\mathbf{x} = [\mathbf{x}_1{}^\top, \cdots, \mathbf{x}_n{}^\top]^\top \in \mathbb{R}^{n d_x}$ that best describe the data

$$\log p(\mathbf{y}|\mathcal{G}) = \log \int\int p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathcal{G}) d\mathbf{x} d\mathbf{C}.$$

where

$$p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathcal{G}) = p(\mathbf{y}|\mathbf{C}, \mathbf{x}, \mathcal{G}) p(\mathbf{C}|\mathcal{G}) p(\mathbf{x}|\mathcal{G}).$$

## Essential quantities




- **prior on latents**: assuming the neighbouring latent variables are similar

$$-\tfrac{1}{2}\sum_{i=1}^{n}(\alpha||\mathbf{x}_i||^2 + \sum_{j=1}^{n}\eta_{ij}||\mathbf{x}_i - \mathbf{x}_j||^2)$$
$$\Longrightarrow\ p(\mathbf{x}|\mathbf{G}, \alpha) = \mathcal{N}(\mathbf{0}, \mathbf{\Pi})$$

where $\alpha$ controls the expected scale, $\mathbf{\Omega}^{-1} = 2\mathbf{L} \otimes \mathbf{I}_{d_x}$ and $\mathbf{\Pi}^{-1} = \alpha \mathbf{I}_{nd_x} + \mathbf{\Omega}^{-1}$.
- **prior on linear maps**: similarly

$$p(\mathbf{C}|\mathbf{G}, \mathbf{U}) = \mathcal{MN}(\mathbf{0}, \mathbf{U}, \mathbf{\Omega}),$$

where $\mathbb{E}[\mathbf{C}\mathbf{C}^\top] \propto \mathbf{U}$.
- **likelihood**: penalising the approximation error yields

$$-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\eta_{ij}((\mathbf{y}_j - \mathbf{y}_i) - \mathbf{C}_i(\mathbf{x}_j - \mathbf{x}_i))^\top \mathbf{V}^{-1}((\mathbf{y}_j - \mathbf{y}_i) - \mathbf{C}_i(\mathbf{x}_j - \mathbf{x}_i))$$
$$\Longrightarrow\ p(\mathbf{y}|\mathbf{C}, \mathbf{x}, \mathbf{V}, \mathbf{G}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y})$$

assuming $\mathbf{V}^{-1} = \gamma I$ and $\gamma$ is a parameter.

## Variational EM

- maximizing log marginal likelihood is intractable, instead maximise lower bound

$$\log p(\mathbf{y}|\mathbf{G}, \boldsymbol{\theta}) \geq \int\int q(\mathbf{C}, \mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}, \boldsymbol{\theta})}{q(\mathbf{C}, \mathbf{x})} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{C} = \mathcal{F}(q(\mathbf{C}, \mathbf{x}), \boldsymbol{\theta}),$$

- for computational tractability, assume $q(\mathbf{C}, \mathbf{x}) = q(\mathbf{x})q(\mathbf{C})$.
- variational expectation maximization algorithm
  - expectation step for computing $q(\mathbf{C}, \mathbf{x})$ by

$$q(\mathbf{x}) \propto \exp\left[\int q(\mathbf{C}) \log p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}, \boldsymbol{\theta}) \mathrm{d}\mathbf{C}\right] = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x}),$$
$$q(\mathbf{C}) \propto \exp\left[\int q(\mathbf{x}) \log p(\mathbf{y}, \mathbf{C}, \mathbf{x}|\mathbf{G}, \boldsymbol{\theta}) \mathrm{d}\mathbf{x}\right] = \mathcal{N}(\mathbf{c}|\boldsymbol{\mu}_\mathbf{c}, \boldsymbol{\Sigma}_\mathbf{c}),$$

  - maximization step for estimating $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q(\mathbf{C}, \mathbf{x}), \boldsymbol{\theta}).$$

## Relation to GP-LVM

Integrating out $\mathbf{C}$ from likelihood yields

$$p(\mathbf{y}|\mathbf{x}, \mathbf{G}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{C}, \mathbf{x}, \mathbf{G}, \boldsymbol{\theta}) p(\mathbf{C}|\mathbf{G}, \boldsymbol{\theta}) d\mathbf{C},$$
$$= \frac{1}{Z_{Y_y}} \exp\left[-\frac{1}{2}\mathbf{y}^\top \mathbf{K}_{LL}^{-1}\mathbf{y}\right].$$

- In contrast to GP-LVM, the precision matrix $\mathbf{K}_{LL}^{-1}$ depends on the Laplacian matrix.
- The functional form of precision is *directly* determined by the *graph structure* given the observations.

$$\mathbf{K}_{LL}^{-1} = (2\mathbf{L} \otimes \mathbf{V}^{-1}) - (\mathbf{W} \otimes \mathbf{V}^{-1})\,\mathbf{\Lambda}\,(\mathbf{W}^\top \otimes \mathbf{V}^{-1}),$$

where $\mathbf{W}$ is a function in $\mathbf{x}$ and $\mathbf{L}$ and $\mathbf{\Lambda}$ is a function in $\mathbf{x}^\top\mathbf{x}$ and $\mathbf{L}$.

## Illustration

- Mitigating short-circuiting problems

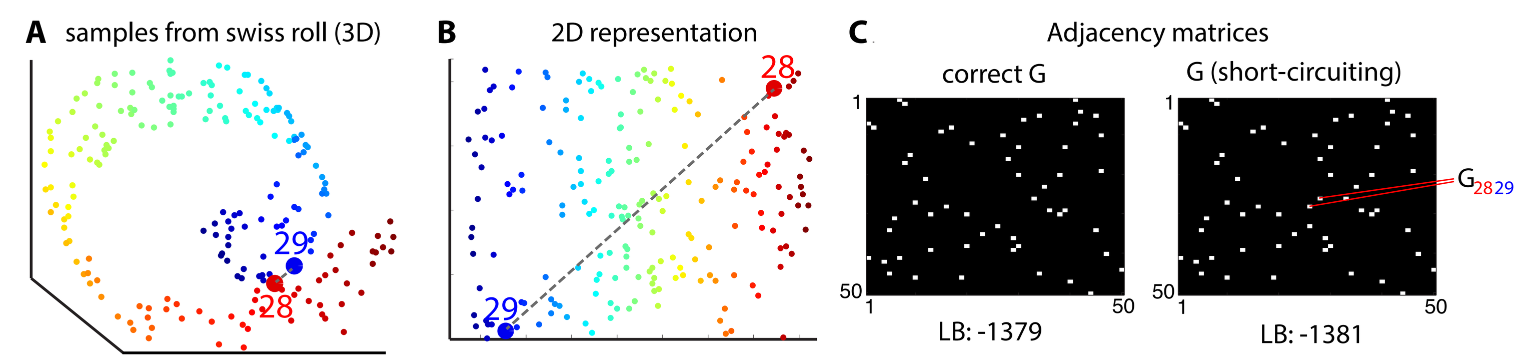



Figure : Two datapoints seem close to each other (**A**) but actually far in 2D space (**B**). Short-circuiting the two datapoints lower the lower bound (**C**)

- Finding the optimal number of neighbours using variational lower bound

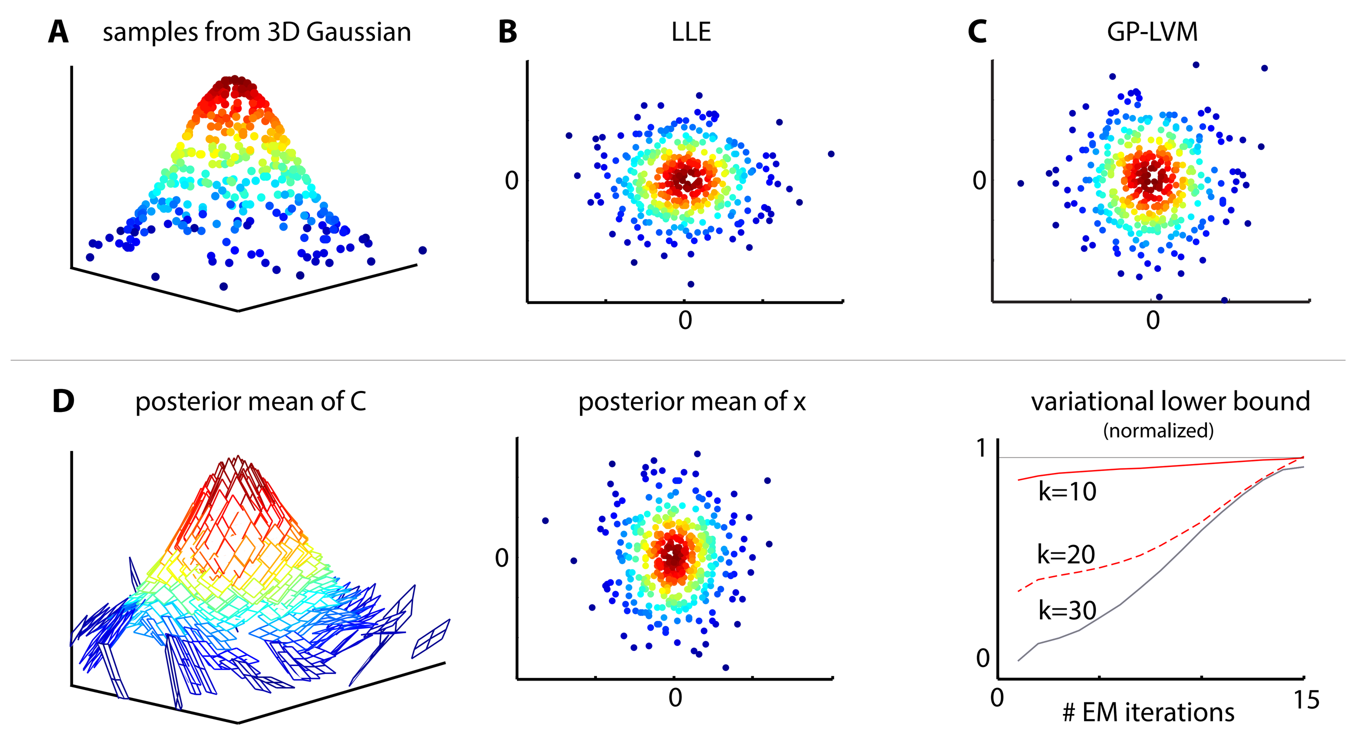



Figure : **A**: 400 samples drawn from 3D Gaussian. **B**: LLE. **C**:GP-LVM. **D (Left)**: The posterior mean of C. **D (Middle)**: posterior mean of x. **D (Right)**: Normalized variational lower bound.

## Conclusion

A new probabilistic approach to manifold learning preserving local geometries in data and equipped with straightforward variational inference for learning the manifold.

## References

[1] N.D. Lawrence, GP-LVM *NIPS 2003*
[2] M.K. Titsias, N.D. Lawrence, Bayesian GP-LVM *AIStats*, 2010

This work was funded by the Gatsby Charitable Foundation.

Contact: mijung@gatsby.ucl.ac.uk