THE LAPLACIAN EIGENMAPS LATENT VARIABLE MODEL.

Miguel Á. Carreira-Perpiñán and Zhengdong Lu.
Dept. of CSEE, OGI, Oregon Health & Science University.

1 Abstract

We introduce the Laplacian Eigenmaps Latent Variable Model (LELVM), a probabilistic method for nonlinear dimensionality reduction that combines the advantages of spectral methods (global optimisation, ability to learn convoluted manifolds of high intrinsic dimensionality) with those of latent variable models (dimensionality reduction and reconstruction mappings, a density model). We derive LELVM by defining a natural out-of-sample mapping for Laplacian eigenmaps with a semi-supervised learning argument. LELVM is simple, nonparametric and computationally not very costly, and is shown to perform well with motion-capture data.

4 The Laplacian Eigenmaps Latent Variable Model (LELVM)

- Assume a LE embedding $\mathbf{X_s} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ of seen points $\mathbf{Y_s} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$. Consider a set of M unseen (out-of-sample) points in observed space $\mathbf{Y_u} = (\mathbf{y}_{N+1}, \dots, \mathbf{y}_{N+M})$.
- Natural way to embed new points without perturbing the old embedding: solve the LE problem again but keeping the old $\mathbf{X}$ fixed:

$$\min_{\mathbf{X_u} \in \mathbb{R}^{L \times M}} \operatorname{tr}\left(\begin{pmatrix} \mathbf{X_s} & \mathbf{X_u} \end{pmatrix} \begin{pmatrix} \mathbf{L_{ss}} & \mathbf{L_{su}} \\ \mathbf{L_{us}} & \mathbf{L_{uu}} \end{pmatrix} \begin{pmatrix} \mathbf{X_s}^T \\ \mathbf{X_u}^T \end{pmatrix} \right)$$

with solution $\mathbf{X_u} = -\mathbf{X_s}\mathbf{L_{su}}\mathbf{L_{uu}}^{-1}$.

- Semi-supervised learning: embedding $\mathbf{X_s}$ as (real-valued) labels for $\mathbf{Y_s}$, want to label $\mathbf{Y_u}$ by using a graph prior.
- In particular, for a single out-of-sample point: $\mathbf{y} = \mathbf{Y_u} \in \mathbb{R}^D$, $\mathbf{x} = \mathbf{X_u} \in \mathbb{R}^L$, $\mathbf{L_{su}} = -\mathbf{W_{su}} = -\mathbf{K}(\mathbf{y}) \in \mathbb{R}^N$ and $l_{\text{uu}} = d_{\text{u}} - w_{\text{uu}} = \mathbf{1}^T\mathbf{K}(\mathbf{y})$,

$$\mathbf{x} = \mathbf{F}(\mathbf{y}) = -\frac{1}{l_{\text{uu}}}\mathbf{X_s}\mathbf{L_{su}} = \frac{\mathbf{X_s}\mathbf{K}(\mathbf{y})}{\mathbf{1}^T\mathbf{K}(\mathbf{y})} = \sum_{n=1}^{N} \frac{K(\mathbf{y}, \mathbf{y}_n)}{\sum_{n'=1}^{N} K(\mathbf{y}, \mathbf{y}_{n'})} \mathbf{x}_n$$

which defines an out-of-sample dimensionality reduction mapping $\mathbf{x} = \mathbf{F}(\mathbf{y})$ applicable to any point $\mathbf{y}$ (new or old).

- We can augment $\mathbf{F}$ with a probabilistic model that is consistent with the mapping $\mathbf{F}$:

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} K_{\mathbf{y}}\left(\frac{\mathbf{y} - \mathbf{y}_n}{\sigma_{\mathbf{y}}}\right) K_{\mathbf{x}}\left(\frac{\mathbf{x} - \mathbf{x}_n}{\sigma_{\mathbf{x}}}\right)$$

$$p(\mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} K_{\mathbf{y}}\left(\frac{\mathbf{y} - \mathbf{y}_n}{\sigma_{\mathbf{y}}}\right) \qquad p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} K_{\mathbf{x}}\left(\frac{\mathbf{x} - \mathbf{x}_n}{\sigma_{\mathbf{x}}}\right)$$

$$\mathbf{F}(\mathbf{y}) = \sum_{n=1}^{N} \frac{K_{\mathbf{y}}((\mathbf{y} - \mathbf{y}_n)/\sigma_{\mathbf{y}})}{\sum_{n'=1}^{N} K_{\mathbf{y}}((\mathbf{y} - \mathbf{y}_{n'})/\sigma_{\mathbf{y}})} \mathbf{x}_n = \sum_{n=1}^{N} p(n|\mathbf{y})\mathbf{x}_n = \mathrm{E}\{\mathbf{x}|\mathbf{y}\}$$

$$\mathbf{f}(\mathbf{x}) = \sum_{n=1}^{N} \frac{K_{\mathbf{x}}((\mathbf{x} - \mathbf{x}_n)/\sigma_{\mathbf{x}})}{\sum_{n'=1}^{N} K_{\mathbf{x}}((\mathbf{x} - \mathbf{x}_{n'})/\sigma_{\mathbf{x}})} \mathbf{y}_n = \sum_{n=1}^{N} p(n|\mathbf{x})\mathbf{y}_n = \mathrm{E}\{\mathbf{y}|\mathbf{x}\}$$

This defines the Laplacian Eigenmaps Latent Variable Model (LELVM):
  - densities are kernel density estimates
  - mappings are Nadaraya-Watson estimators

  all nonparametric.
- We consider Gaussian kernels $K_{\mathbf{x}}$ and $K_{\mathbf{y}}$ with bandwidths $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$ that control the smoothness of densities and mappings ($\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$ can be set with a statistics rule for bandwidths).
- For a graph (e.g. k-nearest-neighbours): same $\mathbf{F}$ but using $\sum_{\mathbf{y}_n \sim \mathbf{y}}$ (where $\mathbf{y}_n \sim \mathbf{y}$ is given by the graph construction rule).
- Advantages: those of latent variable models and spectral methods:
  - yields mappings (nonlinear, infinitely differentiable and based on a global coordinate system)
  - yields densities (potentially multimodal), can deal with missing data
  - no local optima, yet succeeds with convoluted manifolds
  - can use any dimension L
  - computational efficiency $\mathcal{O}(N^3)$ or $\mathcal{O}(N^2)$ (sparse graph)
- Disadvantages: it relies on the success of Laplacian eigenmaps (which depends on the graph).
- Other properties:
  - mappings interpolate $(\mathbf{X}, \mathbf{Y})$ only for $\sigma_{\mathbf{x}}, \sigma_{\mathbf{y}} \to 0$
  - ranges of $\mathbf{f}$ and $\mathbf{F}$ are the convex hulls of $\mathbf{Y}$ and $\mathbf{X}$
  - noise model $p(\mathbf{y}|\mathbf{x})$ can be skewed or even multimodal
- What the user needs to do:
  1. Give graph parameters (k, σ) and run LE.
  2. Give bandwidths $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$ and use the LELVM equations to compute mappings and densities.
- Other out-of-sample extension (Bengio et al 2004): Nyström's formula yields (in matrix notation) $\mathbf{F}(\mathbf{y}) = \sqrt{N}\boldsymbol{\Lambda}^{-1}\mathbf{X}\mathbf{K}(\mathbf{y})/\sqrt{\mathbf{1}^T\mathbf{K}(\mathbf{y})}$ where $\mathbf{X} = \mathbf{V}^T\mathbf{D}^{-\frac{1}{2}}$ are the latent points defined by LE; LELVM yields $\mathbf{F}(\mathbf{y}) = \mathbf{X}\mathbf{K}(\mathbf{y})/\mathbf{1}^T\mathbf{K}(\mathbf{y})$. Crucial difference: the LELVM mapping is a convex sum, which allows augmenting it to a kde.

7 Future work

- Can apply same out-of-sample argument to LLE, etc.
- Asymptotic convergence ($N \to \infty$)
- Applications: priors for tracking in high-dimensional state spaces, value function approximation for MDPs, visualisation, etc.

Partially supported by NSF CAREER award IIS–0546857.

2 Latent variable models (LVMs)

They define a joint density $p(\mathbf{x}, \mathbf{y})$, with latent variables $\mathbf{x} \in \mathbb{R}^L$ and observed variables $\mathbf{y} \in \mathbb{R}^D$, usually by defining:

- Prior in latent space $p(\mathbf{x})$
- Reconstruction mapping $\mathbf{f}: \mathbf{x} \to \mathbf{y}$
- Noise model $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{f}(\mathbf{x}))$

From $p(\mathbf{x}, \mathbf{y})$ we can obtain:

- Marginal in observed space $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$. Necessary for ML estimation: $\max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \log p(\mathbf{y}_n; \boldsymbol{\Theta})$.
- Dimensionality reduction mapping as posterior mean or modes in latent space: $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})$ (Bayes' th.).

Examples:

- Linear LVMs: probabilistic PCA, factor analysis, ICA. . .
- Nonlinear LVMs: Generative Topographic Mapping (GTM) (Bishop et al 1998), Generalised Elastic Net (GEN) (Carreira-Perpiñán et al 2005)

Advantages:

- can represent nonlinear mappings
- can represent multimodal densities
- can deal with missing data

Disadvantages:

- the objective function has many local optima, most of which yield very poor manifolds
- computational cost grows exponentially with latent dimension L, so this limits $L \lesssim 3$ (reason: need to discretise the latent space to compute $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$)

3 Spectral methods: Laplacian Eigenmaps

They obtain a set of points $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ in latent space that correspond to the observed data $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$, such that distances in $\mathbf{X}$ correlate well with distances in $\mathbf{Y}$. Laplacian eigenmaps (LE) (Belkin & Niyogi 2003):

1. Construct a neighbourhood graph on dataset $\mathbf{y}_1, \dots, \mathbf{y}_N$ with weighted edges $w_{mn} = \exp\left(-\frac{1}{2}\|(\mathbf{y}_m - \mathbf{y}_n)/\sigma\|^2\right)$.
2. Set up a quadratic optimisation problem

$$\min_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}\mathbf{L}\mathbf{X}^T\right) \quad \text{s.t.} \quad \mathbf{X}\mathbf{D}\mathbf{X}^T = \mathbf{I},\ \mathbf{X}\mathbf{D}\mathbf{1} = \mathbf{0}$$

$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, affinity matrix $\mathbf{W} = (w_{mn})$, degree matrix $\mathbf{D} = \operatorname{diag}\left(\sum_{n=1}^{N} w_{nm}\right)$, graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

3. Solution: eigenvectors $\mathbf{V} = (\mathbf{v}_2, \dots, \mathbf{v}_{L+1})$ of $\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$, which yield the low-dimensional points $\mathbf{X} = \mathbf{V}^T\mathbf{D}^{-\frac{1}{2}}$.

Other methods (MDS, LLE, Isomap, etc.) use matrices other than $\mathbf{L}$ and possibly different constraints.

Advantages:

- no local optima (unique solution)
- yet succeeds with convoluted manifolds (if the neighbourhood graph is good)
- computational cost $\mathcal{O}(N^3)$ or, for sparse graphs, $\mathcal{O}(N^2)$
- can use any latent space dimension L (just use L eigenvectors)

Disadvantages:

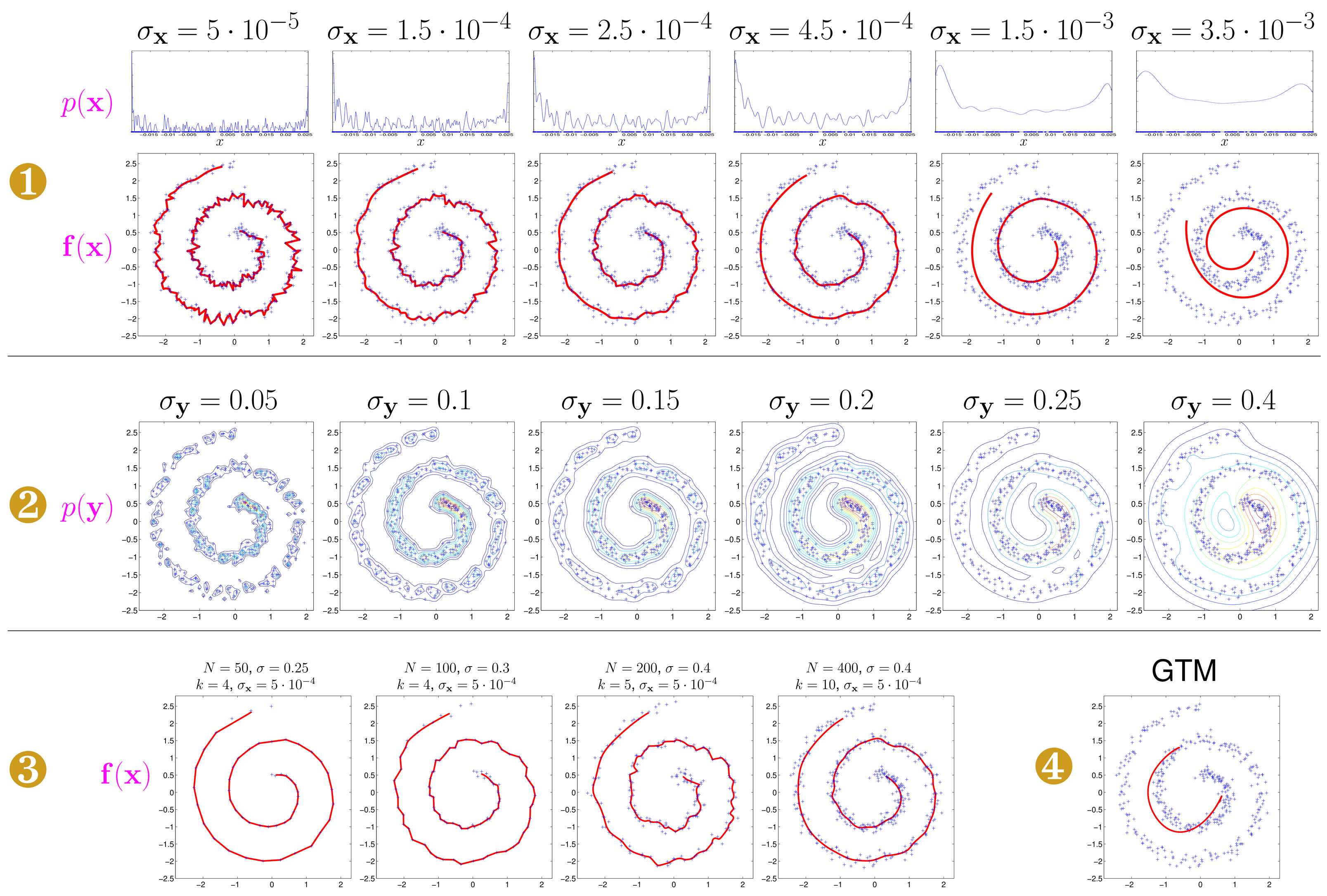
- no mapping for points not in $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ or $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ (out-of-sample mapping)
- no density $p(\mathbf{x}, \mathbf{y})$

☞ Is it possible to combine the advantages of LVMs and spectral methods?

5 Experiments: spiral

Results for the spiral dataset ($N = 400$ points in 2D). 1–2 LELVM results obtained from the LE embedding using a k-nearest-neighbour graph ($k = 10$) and Gaussian affinities ($\sigma = 0.4$). 1 Latent density $p(\mathbf{x})$ and dimensionality reduction mapping $\mathbf{f}(\mathbf{x})$ for different values of $\sigma_{\mathbf{x}}$. 2 Observed density $p(\mathbf{y})$ for different values of $\sigma_{\mathbf{y}}$. 3 LELVM mapping $\mathbf{f}(\mathbf{x})$ obtained for different numbers of training samples N (appropriate k, σ, $\sigma_{\mathbf{x}}$ were used in each case). 4 GTM result ($N = 400$ points) when initialised from PCA.

- LELVM recovers mappings and densities for convoluted manifolds if the LE embedding is good.
- Large bandwidths can unfold loops mistakenly produced by the LE embedding.
- Cross-validation rule for bandwidths yields relatively rough mappings.
- Note the shortening of mappings at the boundaries (convex sum).



6 Experiments: motion-capture data

Results for a mocap dataset recorded from several cycles of a running sequence ($N = 217$ points in 10D, corresponding to 34 3D markers, normalised for translation but not for rotation). A 2D latent space suffices to capture the periodic nature of the running motion. LELVM gives a principled way to interpolate between poses (animation) or deal with missing data (occlusion). 1–2 LELVM results obtained from the LE embedding using a k-nearest-neighbour graph ($k = 40$) and Gaussian affinities ($\sigma = 1.5$), and with $\sigma_{\mathbf{x}} = 0.005$ and $\sigma_{\mathbf{y}} = 0.3$. 1 (left) Latent space. We connect data points $\mathbf{x}_n$ in the sequential order of their corresponding data points $\mathbf{y}_n$, and for some of those we plot $\mathbf{y}_n$ as a stick man. The loop is travelled clockwise; the contours indicate $p(\mathbf{x})$. 1 (right) interpolation (animation) by designing trajectories in latent space and mapping them to pose space with $\mathbf{f}(\mathbf{x})$ (lower plot). For each trajectory, only the initial and final points were in the dataset; the rest are smoothly produced by the mapping. 2 Reconstruction of missing data with LELVM. Given a partially observed stick man (black: observed, red: missing) we show the contours of $p(\mathbf{x}|\mathbf{y}_{\text{obs}})$ and the reconstructed stick men. When the legs are missing, $p(\mathbf{x}|\mathbf{y}_{\text{obs}})$ is unimodal, but when only the forearm is observed, it is multimodal. 3 (left): latent space using GTM initialised from the LE embedding. 3 (middle & right): latent space using GPLVM without/with back constraints (note that, unlike the contours in panels 1–2, the greyscales do not represent $p(\mathbf{x}|\mathbf{y})$).