A Unifying Probabilistic Perspective on Spectral Approaches to Dimensionality Reduction

Neil D. Lawrence

University of Sheffield

5th September 2013

Outline

Notation

p	data dimensionality	
q	latent dimensionality	
n	number of data points	
$\mathbf{Y}$	*design matrix* containing our data	$n \times p$
$\mathbf{X}$	matrix of latent variables	$n \times q$
$\mathbf{D}$	matrix of interpoint squared distances	$n \times n$
$\mathbf{K}$	similarities/covariance/kernel	$n \times n$
$\mathbf{L}$	Laplacian matrix	$n \times n$

Row vector from matrix $\mathbf{A}$ given by $\mathbf{a}_{i,:}$ column vector $\mathbf{a}_{:,j}$ and element given by $a_{i,j}$.

Outline

Review

Relation to Laplacian Eigenmaps

Relation to Locally Linear Embedding

Relation to Isomap

Relation to GP-LVM

Experiments

Discussion and Conclusions

Distances and Similarities

- Typical scenario, a data set, **Y** stored in a matrix of dimension $n \times p$.
- Proximity data: a data set in form of distances, **D**, or similarities **K**. These matrices are dimension $n \times n$.
  - Similarity matrices have large entries when data points are close.
  - Distance matrices have large entries when points are far apart.

Multidimensional Scaling

- Multidimensional scaling (MDS) algorithms are dimensionality reduction for proximity matrices.
- We can move between similarity and squared distance as follows $d_{i,j} = k_{i,i} - 2k_{i,j} + k_{j,j}$.
    - In MDS this is known as the standard transformation (Mardia et al., 1979).
    - If $k_{i,j} = k(\mathbf{y}_{i,:}, \mathbf{y}_{j,:})$ is a "kernel" this is the "distance in feature space" (Schölkopf and Smola, 2001).
    - If $k_{i,j}$ is an element from a covariance matrix $\mathbf{K}$, it is the *expected squared distance* between two samples with that covariance.

Note: Centering and Squared Distances

- Consider matrix form of squared distance,

$$\mathbf{D} = \text{diag}\left(\mathbf{Y}\mathbf{Y}^\top\right)\mathbf{1}^\top - 2\mathbf{Y}\mathbf{Y}^\top + \mathbf{1}\text{diag}\left(\mathbf{Y}\mathbf{Y}^\top\right)^\top.$$

- A Centering matrix has the form

$$\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top : \quad \mathbf{H}\mathbf{1} = \mathbf{0}$$

- This implies:

$$-\frac{1}{2}\mathbf{H}\mathbf{D}\mathbf{H} = \mathbf{H}\mathbf{Y}\mathbf{Y}^\top\mathbf{H} = \hat{\mathbf{Y}}\hat{\mathbf{Y}}^\top.$$

- i.e. centered square distance matrix is closely related to centred similarity/kernel.

Spectral Dimensionality Reduction in Machine Learning

- Spectral approach to dimensionality reduction.
  1. Convert data to a matrix of dimension $n \times n$.
  2. Visualize data with eigenvectors of matrix.
- Examples:
  - isomap (Tenenbaum et al., 2000),
  - locally linear embeddings (LLE, Roweis and Saul, 2000),
  - Laplacian eigenmaps (LE, Belkin and Niyogi, 2003) and
  - maximum variance unfolding (MVU, Weinberger et al., 2004).
  - Also kernel PCA (Schölkopf et al., 1998; Ham et al., 2004).

Classical Multidimensional Scaling Perspective

- Classical multidimensional scaling (CMDS)
  1. Compute an $n \times n$ squared distance matrix, $\mathbf{D}$.
  2. Form the centered "similarity matrix" $\mathbf{HKH} = -\frac{1}{2}\mathbf{HDH}$.
  3. Visualize through q principal eigenvectors (as latent matrix $\mathbf{X}$).
- This algorithm matches squared distances computed in $\mathbf{X}$ to those computed in $\mathbf{Y}$ through an L1 error.
- Our Argument:
  - Main innovation in ML work: how to compute the squared distance matrix $\mathbf{D}$.

Unifying Perspective

- Introduce probabilistic approach to constructing squared distance matrices.
- Relate isomap, LLE, LE and MVU to the approach.
- Wrap spectral methods in a unifying perspective of *Gaussian random fields* and CMDS.

Maximum Variance Unfolding

Learn a "Kernel" for Dimensionality Reduction

- In maximum variance unfolding (MVU Weinberger et al., 2004): learn a "kernel matrix" that will allow for dimensionality reduction.

Maximum Variance Unfolding

Learn a "Kernel" for Dimensionality Reduction

- In maximum variance unfolding (MVU Weinberger et al., 2004): learn a "kernel matrix" that will allow for dimensionality reduction.
- Preserve only *local* proximity relationships in the data.

Maximum Variance Unfolding

Learn a “Kernel” for Dimensionality Reduction

- In maximum variance unfolding (MVU Weinberger et al., 2004): learn a “kernel matrix” that will allow for dimensionality reduction.
- Preserve only *local* proximity relationships in the data.
  - Take a set of neighbors.

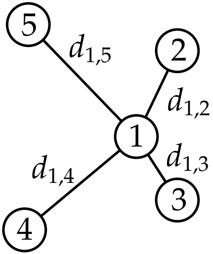
Maximum Variance Unfolding

Learn a "Kernel" for Dimensionality Reduction

- In maximum variance unfolding (MVU Weinberger et al., 2004): learn a "kernel matrix" that will allow for dimensionality reduction.
- Preserve only *local* proximity relationships in the data.
  - Take a set of neighbors.
  - Construct a kernel matrix where only distances between neighbors match data distances.

Maximum Variance Unfolding

- Optimize elements of $\mathbf{K}$ by maximizing[1] $\text{tr}\left(\mathbf{K}\right)$.



- Subject to squared distance constraints between neighbors

$$d_{i,j} = k_{i,i} - 2k_{i,j} + k_{j,j}$$

Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).

Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).
- Entropy and variance are closely related.

Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).
- Entropy and variance are closely related.
- Maximum entropy leads to a *probabilistic model*.

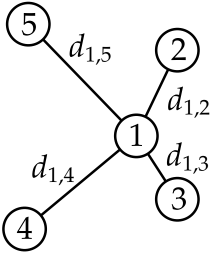
Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).
- Entropy and variance are closely related.
- Maximum entropy leads to a *probabilistic model*.
- Each spectral approach approximates MEU in some way.

Maximum Entropy Unfolding

- Find distribution with maximum entropy subject to constraints on *moments*.

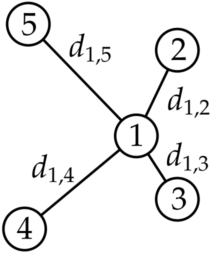


- MEU constraints are on expected distances between neighbors.

$$d_{i,j} = \left\langle \mathbf{y}_{i,:}^\top \mathbf{y}_{i,:} \right\rangle - 2 \left\langle \mathbf{y}_{i,:}^\top \mathbf{y}_{j,:} \right\rangle + \left\langle \mathbf{y}_{j,:}^\top \mathbf{y}_{j,:} \right\rangle$$

Maximum Entropy Unfolding

- Find distribution with maximum entropy subject to constraints on *moments*.



- MEU constraints are on expected distances between neighbors.

$$d_{i,j} = k_{i,i} - 2k_{i,j} + k_{j,j}$$

which can be written in terms of the covariance.

Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).

Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).
- Entropy and variance are closely related.

Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).
- Entropy and variance are closely related.
- Maximum entropy leads to a *probabilistic model*.

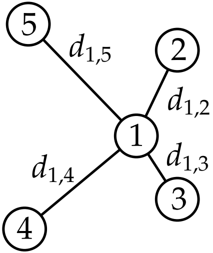
Maximum Entropy Unfolding

New Contribution

- Maximize *entropy* instead of variance (Jaynes, 1986): MEU (Lawrence, 2011, 2010).
- Entropy and variance are closely related.
- Maximum entropy leads to a *probabilistic model*.
- Each spectral approach approximates MEU in some way.

Maximum Entropy Unfolding

- Find distribution with maximum entropy subject to constraints on *moments*.

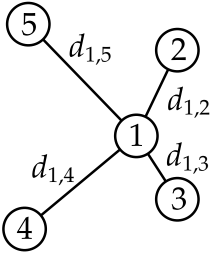


- MEU constraints are on expected distances between neighbors.

$$d_{i,j} = \left\langle \mathbf{y}_{i,:}^{\top}\mathbf{y}_{i,:} \right\rangle - 2\left\langle \mathbf{y}_{i,:}^{\top}\mathbf{y}_{j,:} \right\rangle + \left\langle \mathbf{y}_{j,:}^{\top}\mathbf{y}_{j,:} \right\rangle$$

Maximum Entropy Unfolding

- Find distribution with maximum entropy subject to constraints on *moments*.



- MEU constraints are on expected distances between neighbors.

$$d_{i,j} = k_{i,i} - 2k_{i,j} + k_{j,j}$$

which can be written in terms of the covariance.

Maximum Entropy

- Maximum entropy distribution.

$$p(\mathbf{Y}) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\gamma\mathbf{Y}\mathbf{Y}^\top\right)\right)\exp\left(-\frac{1}{2}\sum_i \sum_{j\in\mathcal{N}(i)} \lambda_{i,j}d_{i,j}\right)$$

$\mathcal{N}(i)$ is neighborhood, $\{\lambda_{i,j}\}$, Lagrange multipliers.

Maximum Entropy

- Maximum entropy distribution.

$$p(\mathbf{Y}) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(\gamma\mathbf{Y}\mathbf{Y}^\top\right) - \frac{1}{4}\mathrm{tr}\left(\mathbf{\Lambda}\mathbf{D}\right)\right)$$

$\mathcal{N}(i)$ is neighborhood, $\{\lambda_{i,j}\}$, Lagrange multipliers. Lagrange multipliers in sparse matrix $\mathbf{\Lambda}$.

Maximum Entropy

- Maximum entropy distribution.

$$p(\mathbf{Y}) = \frac{\left|\mathbf{L} + \gamma\mathbf{I}\right|^{\frac{1}{2}}}{(2\pi)^{\frac{np}{2}}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{L} + \gamma\mathbf{I})\mathbf{Y}\mathbf{Y}^\top\right)\right)$$

$\mathcal{N}(i)$ is neighborhood, $\{\lambda_{i,j}\}$, Lagrange multipliers.
Introduce Laplacian: $\ell_{i,j} = -\lambda_{i,j}$, $\ell_{i,i} = \sum_{j\in\mathcal{N}(i)} \lambda_{i,j}$, $\mathbf{L1} = \mathbf{0}$.

Details: Moving to the Laplacian

- $\mathbf{D}$ has a zero diagonal.
- $\mathrm{tr}\left(\mathbf{L}\mathbf{D}\right)$ is unaffected by diagonal of $\mathbf{L}$.
- Constrain $\mathbf{L}\mathbf{1} = \mathbf{0}$ giving

$$-\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbf{D}\right) = \mathrm{tr}\left(\mathbf{L}\mathbf{D}\right)$$

Details: Moving to the Laplacian

- $\mathbf{D}$ has a zero diagonal.
- $\text{tr}\left(\mathbf{LD}\right)$ is unaffected by diagonal of $\mathbf{L}$.
- Constrain $\mathbf{L1} = \mathbf{0}$ giving

$$-\text{tr}\left(\mathbf{\Lambda D}\right) = \text{tr}\left(\mathbf{LD}\right)$$

Details: Moving to the Laplacian

- $\mathbf{D}$ has a zero diagonal.
- $\mathrm{tr}\left(\mathbf{L}\mathbf{D}\right)$ is unaffected by diagonal of $\mathbf{L}$.
- Constrain $\mathbf{L}\mathbf{1} = \mathbf{0}$ giving

$$-\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbf{D}\right) = \mathrm{tr}\left(\mathbf{L}\mathbf{1}\mathrm{diag}\left(\mathbf{Y}\mathbf{Y}^{\top}\right)^{\top} - 2\mathbf{L}\mathbf{Y}\mathbf{Y}^{\top} + \mathrm{diag}\left(\mathbf{Y}\mathbf{Y}^{\top}\right)\mathbf{1}^{\top}\mathbf{L}\right)$$

Details: Moving to the Laplacian

- $\mathbf{D}$ has a zero diagonal.
- $\text{tr}\left(\mathbf{LD}\right)$ is unaffected by diagonal of $\mathbf{L}$.
- Constrain $\mathbf{L1} = \mathbf{0}$ giving

$$-\text{tr}\left(\boldsymbol{\Lambda}\mathbf{D}\right) = \text{tr}\left(\cancel{\mathbf{L1}\text{diag}\left(\mathbf{YY}^{\top}\right)^{\top}} - 2\mathbf{LYY}^{\top} + \cancel{\text{diag}\left(\mathbf{YY}^{\top}\right)\mathbf{1}^{\top}\mathbf{L}}\right)$$

Details: Moving to the Laplacian

- $\mathbf{D}$ has a zero diagonal.
- $\mathrm{tr}\left(\mathbf{L}\mathbf{D}\right)$ is unaffected by diagonal of $\mathbf{L}$.
- Constrain $\mathbf{L}\mathbf{1} = \mathbf{0}$ giving

$$-\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbf{D}\right) = -2\mathrm{tr}\left(\mathbf{L}\mathbf{Y}\mathbf{Y}^{\top}\right).$$

Gaussian Random Field

- The maximum entropy probability distribution is a *Gaussian random field*

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Covariance matrix is

$$\mathbf{K} = (\mathbf{L} + \gamma\mathbf{I})^{-1}$$

.

- Where $\mathbf{L}$ is the *Laplacian* matrix associated with the neighborhood graph.
- Off diagonal elements of the Laplacian are Lagrange multipliers from moment constraints.
- On diagonal elements given by negative sum of off-diagonal ($\mathbf{L1} = \mathbf{0}$).

Data Feature Independence

- The GRF specifying independence across data *features*.
- Most applications of Gaussian models are applied independently across data *points*.
    - Notable exceptions include Zhu et al. (2003); Lawrence (2004, 2005); Kemp and Tenenbaum (2008).
- Maximum likelihood in this model is equivalent maximizing entropy under distance constraints.

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)
  - As we increase data points parameters become better determined.

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)
  - As we increase data points parameters become better determined.
  - **Not** in this model.

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)
  - As we increase data points parameters become better determined.
  - **Not** in this model.
  - As we increase data features parameters become better determined.

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)
  - As we increase data points parameters become better determined.
  - **Not** in this model.
  - As we increase data features parameters become better determined.
- This turns the large p small n problem on its head.

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{j=1}^{p} \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{:,j}^{\top}\mathbf{K}^{-1}\mathbf{y}_{:,j}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)
  - As we increase data points parameters become better determined.
  - **Not** in this model.
  - As we increase data features parameters become better determined.
- This turns the large p small n problem on its head.
- There is a "Blessing of Dimensionality" in this model.

Blessing of Dimensionality

$$p(\mathbf{Y}) = \prod_{i=1}^{n} \frac{1}{|\mathbf{C}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_{i,:}^{\top}\mathbf{C}^{-1}\mathbf{y}_{i,:}\right),$$

- Maximum likelihood is consistent: (see e.g. Wasserman, 2003, pg 126)
  - As we increase data points parameters become better determined.
  - **Not** in this model.
  - As we increase data features parameters become better determined.
- This turns the large p small n problem on its head.
- There is a "Blessing of Dimensionality" in this model.

Inverse Covariance

- From the "covariance interpretation" we think of the similarity matrix as a covariance matrix.
  - Each element of the covariance is a function of two data points.
- For LE, LLE and MVU the stiffness matrix is like an *inverse covariance*.
  - This is a *conditional independence* assumption.
  - Describes how points are connected.

Conditional Independence

- A covariance matrix specifies correlation between two variables. If elements are zero those variables are *truly* independent.
  - In a marginal Gaussian those correlations don't change.
- The inverse covariance (precision, or information matrix) specifies conditional independencies.
  - If elements are zero those variables are *conditionally* independent.

Outline

Relationship to Laplacian Eigenmaps

- Laplacian eigenmaps (Belkin and Niyogi, 2003): graph Laplacian is specified across the data points.

Relationship to Laplacian Eigenmaps

- Laplacian eigenmaps (Belkin and Niyogi, 2003): graph Laplacian is specified across the data points.
- Laplacian has exactly the same form as our matrix **L**.

Relationship to Laplacian Eigenmaps

- Laplacian eigenmaps (Belkin and Niyogi, 2003): graph Laplacian is specified across the data points.
- Laplacian has exactly the same form as our matrix **L**.
- Parameters of the Laplacian are set either as constant or according to the distance between two points.

Relationship to Laplacian Eigenmaps

- Laplacian eigenmaps (Belkin and Niyogi, 2003): graph Laplacian is specified across the data points.
- Laplacian has exactly the same form as our matrix $\mathbf{L}$.
- Parameters of the Laplacian are set either as constant or according to the distance between two points.
- Smallest eigenvectors of this Laplacian are then used for visualizing the data.

Smallest Eigenvalues of Laplacian

- Eigendecomposition of the covariance is

$$\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$$

Smallest Eigenvalues of Laplacian

- Eigendecomposition of the covariance is

$$\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\top}$$

- Eigendecomposition of the Laplacian is

$$\mathbf{L} = \mathbf{U}\left(\boldsymbol{\Lambda}^{-1} - \gamma\mathbf{I}\right)\mathbf{U}^{\top}$$

Smallest Eigenvalues of Laplacian

- Eigendecomposition of the covariance is

$$\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$$

- Eigendecomposition of the Laplacian is

$$\mathbf{L} = \mathbf{U}\left(\boldsymbol{\Lambda}^{-1} - \gamma\mathbf{I}\right)\mathbf{U}^\top$$

- Principal eigenvalues of $\mathbf{K}$ are smallest eigenvalues of $\mathbf{L}$.
  - (smallest eigenvalue of $\mathbf{L}$ is zero, but this is removed by the centering operation on $\mathbf{K}$, or discarded in LE)

Laplacian Eigenmaps

- Set parameters of Laplacian.

Laplacian Eigenmaps

- Set parameters of Laplacian.
- Perform CMDS on the implied matrix $\mathbf{K}$.

Laplacian Eigenmaps

- Set parameters of Laplacian.
- Perform CMDS on the implied matrix $\mathbf{K}$.
    1. No constraints are imposed in Laplacian eigenmaps so distances will not be preserved.

Laplacian Eigenmaps

- Set parameters of Laplacian.
- Perform CMDS on the implied matrix $\mathbf{K}$.
  1. No constraints are imposed in Laplacian eigenmaps so distances will not be preserved.
  2. LE gains significant computational advantage by not representing the covariance matrix explicitly.

Laplacian Eigenmaps

- Set parameters of Laplacian.
- Perform CMDS on the implied matrix $\mathbf{K}$.
  1. No constraints are imposed in Laplacian eigenmaps so distances will not be preserved.
  2. LE gains significant computational advantage by not representing the covariance matrix explicitly.
  3. No matrix inverses required, eigenvalue problem sparse.

Outline

Locally Linear Embedding

- The Laplacian should be constrained positive definite.

Locally Linear Embedding

- The Laplacian should be constrained positive definite.
- This constraint can be imposed by factorizing it as

$$\mathbf{L} = \mathbf{M}\mathbf{M}^{\top}$$

- The Laplacian should be constrained positive definite.
- This constraint can be imposed by factorizing it as

$$\mathbf{L} = \mathbf{M}\mathbf{M}^{\top}$$

- To ensure it is a Laplacian, we need to constrain $\mathbf{M}^{\top}\mathbf{1} = \mathbf{0}$ giving $\mathbf{L}\mathbf{1} = \mathbf{0}$.

Locally Linear Embedding

- The Laplacian should be constrained positive definite.
- This constraint can be imposed by factorizing it as

$$\mathbf{L} = \mathbf{M}\mathbf{M}^\top$$

- To ensure it is a Laplacian, we need to constrain $\mathbf{M}^\top\mathbf{1} = \mathbf{0}$ giving $\mathbf{L}\mathbf{1} = \mathbf{0}$.
  - i.e. $m_{i,i} = -\sum_{j \in \mathcal{N}(i)} m_{j,i}$

Locally Linear Embedding

- The Laplacian should be constrained positive definite.
- This constraint can be imposed by factorizing it as

$$\mathbf{L} = \mathbf{M}\mathbf{M}^\top$$

- To ensure it is a Laplacian, we need to constrain $\mathbf{M}^\top\mathbf{1} = \mathbf{0}$ giving $\mathbf{L}\mathbf{1} = \mathbf{0}$.
  - i.e. $m_{i,i} = -\sum_{j \in \mathcal{N}(i)} m_{j,i}$
  - Set $m_{j,i} = 0$ if $j \notin \mathcal{N}(i)$.

Locally Linear Embedding

- Locally linear embeddings (Roweis and Saul, 2000) are then a specific case of MEU where

Locally Linear Embedding

- Locally linear embeddings (Roweis and Saul, 2000) are then a specific case of MEU where
  1. The diagonal sums, $m_{i,i}$, are further constrained to unity.

Locally Linear Embedding

- Locally linear embeddings (Roweis and Saul, 2000) are then a specific case of MEU where
  1. The diagonal sums, $m_{i,i}$, are further constrained to unity.
  2. Model parameters found by maximizing *pseudolikelihood* of the data.

Point One

- For unit diagonals we have $\mathbf{M} = \mathbf{I} - \mathbf{W}$.

Point One

- For unit diagonals we have $\mathbf{M} = \mathbf{I} - \mathbf{W}$.
- Here the off diagonal sparsity pattern of $\mathbf{W}$ matches $\mathbf{M}$.

Point One

- For unit diagonals we have $\mathbf{M} = \mathbf{I} - \mathbf{W}$.
- Here the off diagonal sparsity pattern of $\mathbf{W}$ matches $\mathbf{M}$.
- Thus
$$(\mathbf{I} - \mathbf{W})^\top \mathbf{1} = \mathbf{0}.$$

Point One

- For unit diagonals we have $\mathbf{M} = \mathbf{I} - \mathbf{W}$.
- Here the off diagonal sparsity pattern of $\mathbf{W}$ matches $\mathbf{M}$.
- Thus

$$(\mathbf{I} - \mathbf{W})^\top \mathbf{1} = \mathbf{0}.$$

- LLE proscribes that the smallest eigenvectors of

$$(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^\top = \mathbf{M}\mathbf{M}^\top = \mathbf{L}$$

(like Laplacian Eigenmaps).

Point One

- For unit diagonals we have $\mathbf{M} = \mathbf{I} - \mathbf{W}$.
- Here the off diagonal sparsity pattern of $\mathbf{W}$ matches $\mathbf{M}$.
- Thus
$$(\mathbf{I} - \mathbf{W})^\top \mathbf{1} = \mathbf{0}.$$
- LLE proscribes that the smallest eigenvectors of
$$(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^\top = \mathbf{M}\mathbf{M}^\top = \mathbf{L}$$
(like Laplacian Eigenmaps).
- Equivalent to CMDS on the GRF described by $\mathbf{L}$.

Second Point

- Pseudolikelihood approximation (see e.g. Koller and Friedman, 2009, pg 970): product of the conditional densities:

$$p(\mathbf{Y}) \approx \prod_{i=1}^{n} p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}),$$

$\mathbf{Y}_{\backslash i}$ represents data other than the ith point.

Second Point

- Pseudolikelihood approximation (see e.g. Koller and Friedman, 2009, pg 970): product of the conditional densities:

$$p(\mathbf{Y}) \approx \prod_{i=1}^{n} p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}),$$

  $\mathbf{Y}_{\backslash i}$ represents data other than the ith point.
- True likelihood is proportional to this but requires renormalization.

Second Point

- Pseudolikelihood approximation (see e.g. Koller and Friedman, 2009, pg 970): product of the conditional densities:

$$p(\mathbf{Y}) \approx \prod_{i=1}^{n} p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}),$$

  $\mathbf{Y}_{\backslash i}$ represents data other than the ith point.
- True likelihood is proportional to this but requires renormalization.
- In pseudolikelihood normalization is ignored.

Conditionals

- Factors in the GRF are the conditionals,

$$p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}) = \left(\frac{m_{i,i}^2}{2\pi}\right)^{\frac{p}{2}} \exp\left(-\frac{m_{i,i}^2}{2}\left\|\mathbf{y}_{i,:} - \sum_{j\in\mathcal{N}(i)} \frac{w_{j,i}}{m_{i,i}}\mathbf{y}_{j,:}\right\|_2^2\right).$$

Conditionals

- Factors in the GRF are the conditionals,

$$p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}) = \left(\frac{m_{i,i}^2}{2\pi}\right)^{\frac{p}{2}} \exp\left(-\frac{m_{i,i}^2}{2}\left\|\mathbf{y}_{i,:} - \sum_{j\in\mathcal{N}(i)}\frac{w_{j,i}}{m_{i,i}}\mathbf{y}_{j,:}\right\|_2^2\right).$$

- Maximizing each conditional is equivalent to optimizing LLE objective.

Conditionals

- Factors in the GRF are the conditionals,

$$p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}) = \left(\frac{m_{i,i}^2}{2\pi}\right)^{\frac{p}{2}} \exp\left(-\frac{m_{i,i}^2}{2}\left\|\mathbf{y}_{i,:} - \sum_{j\in\mathcal{N}(i)} \frac{w_{j,i}}{m_{i,i}}\mathbf{y}_{j,:}\right\|_2^2\right).$$

- Maximizing each conditional is equivalent to optimizing LLE objective.
- Constraint that LLE weights sum to one arises naturally because $w_{j,i}/m_{i,i}$ and $m_{i,i} = \sum_{j\in\mathcal{N}(i)} w_{j,i}$.

Conditionals

- Factors in the GRF are the conditionals,

$$p(\mathbf{y}_{i,:}|\mathbf{Y}_{\backslash i}) = \left(\frac{m_{i,i}^2}{2\pi}\right)^{\frac{p}{2}} \exp\left(-\frac{m_{i,i}^2}{2}\left\|\mathbf{y}_{i,:} - \sum_{j\in\mathcal{N}(i)} \frac{w_{j,i}}{m_{i,i}}\mathbf{y}_{j,:}\right\|_2^2\right).$$

- Maximizing each conditional is equivalent to optimizing LLE objective.
- Constraint that LLE weights sum to one arises naturally because $w_{j,i}/m_{i,i}$ and $m_{i,i} = \sum_{j\in\mathcal{N}(i)} w_{j,i}$.
- In LLE a *further* constraint is imposed $m_{i,i} = 1$.

LLE Approximates MEU

- LLE is an approximation to maximum likelihood.

LLE Approximates MEU

- LLE is an approximation to maximum likelihood.
- Laplacian has factorized form.

LLE Approximates MEU

- LLE is an approximation to maximum likelihood.
- Laplacian has factorized form.
- Pseudolikelihood also allows for relatively quick parameter estimation.

LLE Approximates MEU

- LLE is an approximation to maximum likelihood.
- Laplacian has factorized form.
- Pseudolikelihood also allows for relatively quick parameter estimation.
  - ignoring the partition function removes the need to invert to recover the covariance matrix.

LLE Approximates MEU

- LLE is an approximation to maximum likelihood.
- Laplacian has factorized form.
- Pseudolikelihood also allows for relatively quick parameter estimation.
  - ignoring the partition function removes the need to invert to recover the covariance matrix.
  - LLE can be applied to larger data sets than MEU or MVU.

Note: The sparsity pattern in the Laplacian for LLE will not match that used in the Laplacian for the other algorithms due to the factorized representation.

LLE and PCA

- LLE is motivated by considering local linear embeddings of the data.

LLE and PCA

- LLE is motivated by considering local linear embeddings of the data.
- Interestingly, as we increase the neighborhood size to $K = n - 1$ we do not recover PCA.

LLE and PCA

- LLE is motivated by considering local linear embeddings of the data.
- Interestingly, as we increase the neighborhood size to $K = n - 1$ we do not recover PCA.
- But PCA is the "optimal" linear embedding!!

LLE and PCA

- LLE is motivated by considering local linear embeddings of the data.
- Interestingly, as we increase the neighborhood size to $K = n - 1$ we do not recover PCA.
- But PCA is the "optimal" linear embedding!!
- LLE is optimizing a pseudolikelihood: in contrast the MEU algorithm, which LLE approximates, does recover PCA when $K = n - 1$.

Acyclic Locally Linear Embedding

- The pseudolikelihood is an approximation.
- Unless neighborhood in **M** is forced *acyclic*.
- Then **M** is a *Cholesky* factor and pseudolikelihood approximation is *exact*.
- Normalizer of Gaussian model *is*

$$\left(\frac{\left|\mathbf{M}\mathbf{M}^\top\right|}{2\pi}\right)^{\frac{p}{2}} = \left(\frac{m_{i,i}^2}{2\pi}\right)^{\frac{p}{2}}$$

- This gives a *very fast* approach to fitting MEU.
- We call this acyclic LLE.
- It does include PCA as special case.

Outline

- Review
- Relation to Laplacian Eigenmaps
- Relation to Locally Linear Embedding
- **Relation to Isomap**
- Relation to GP-LVM
- Experiments
- Discussion and Conclusions

Isomap and MEU

- Both MVU and MEU can be thought of as starting with a sparse graph of (squared) distances.

Isomap and MEU

- Both MVU and MEU can be thought of as starting with a sparse graph of (squared) distances.
- Fill in other distances by maximizing the total variance/entropy.

Isomap and MEU

- Both MVU and MEU can be thought of as starting with a sparse graph of (squared) distances.
- Fill in other distances by maximizing the total variance/entropy.
- Interneighbor distances in this graph are preserved just like in isomap.

Isomap and MEU

- Both MVU and MEU can be thought of as starting with a sparse graph of (squared) distances.
- Fill in other distances by maximizing the total variance/entropy.
- Interneighbor distances in this graph are preserved just like in isomap.
    1. For isomap the implied covariance can have negative eigenvalues (see Weinberger et al., 2004).

Isomap and MEU

- Both MVU and MEU can be thought of as starting with a sparse graph of (squared) distances.
- Fill in other distances by maximizing the total variance/entropy.
- Interneighbor distances in this graph are preserved just like in isomap.
  1. For isomap the implied covariance can have negative eigenvalues (see Weinberger et al., 2004).
  2. Isomap is slower than LLE and LE: requires a dense eigenvalue problem and a shortest path algorithm.

Outline

Relationship to GP-LVM

- Gaussian Process latent variable models (Lawrence, 2005) also define Gaussian densities independently over the features.

Relationship to GP-LVM

- Gaussian Process latent variable models (Lawrence, 2005) also define Gaussian densities independently over the features.
- GP-LVMs construct a Gaussian process by specifying a covariance function (Mercer kernel) in $\mathbf{X}$.
- A Gauss Markov random field can be specified by a Gaussian process through appropriate covariance functions
$$k(x, x') = \exp(-\|x - x\prime\|_1)$$

Relationship to GP-LVM

- Gaussian Process latent variable models (Lawrence, 2005) also define Gaussian densities independently over the features.
- GP-LVMs construct a Gaussian process by specifying a covariance function (Mercer kernel) in **X**.
- A Gauss Markov random field can be specified by a Gaussian process through appropriate covariance functions
$$k(x, x') = \exp(-\|x - x\prime\|_1)$$
- Inverse covariance will be sparse and based on neighborhood.

Relationship to GP-LVM

- Gaussian Process latent variable models (Lawrence, 2005) also define Gaussian densities independently over the features.
- GP-LVMs construct a Gaussian process by specifying a covariance function (Mercer kernel) in **X**.
- A Gauss Markov random field can be specified by a Gaussian process through appropriate covariance functions
$$k(x, x') = \exp(-\|x - x\prime\|_1)$$
- Inverse covariance will be sparse and based on neighborhood.
- In the GP-LVM the neighborhood is learnt by optimization of **X**.

Outline

Review

Relation to Laplacian Eigenmaps

Relation to Locally Linear Embedding

Relation to Isomap

Relation to GP-LVM

Experiments

Discussion and Conclusions

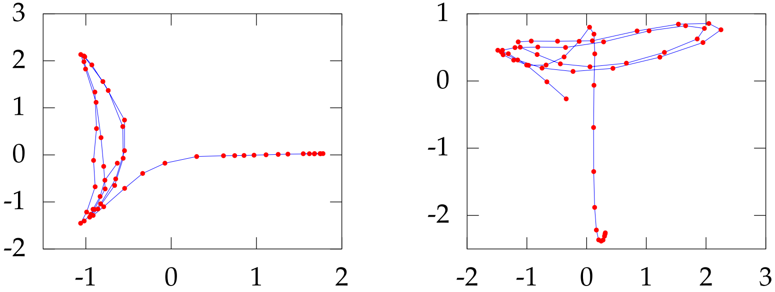
Simple Experiments

- Consider two real data sets.
- We apply each of the spectral methods we have reviewed.
- Apply the MEU framework.
- Follow the suggestion of Harmeling (Harmeling, 2007) and use the GPLVM likelihood (Lawrence, 2005) for embedding quality.
- The higher the likelihood the better the embedding.

Motion Capture Data

- Data consists of a 3-dimensional point cloud of the location of 34 points from a subject performing a run.
- 102 dimensional data set containing 55 frames of motion capture.
- Subject begins the motion from stationary and takes approximately three strides of run.
- Should see this structure in the visualization: a starting position followed by a series of loops.
- Data was made available by Ohio State University.
- The two dominant eigenvectors are visualized in following figures.

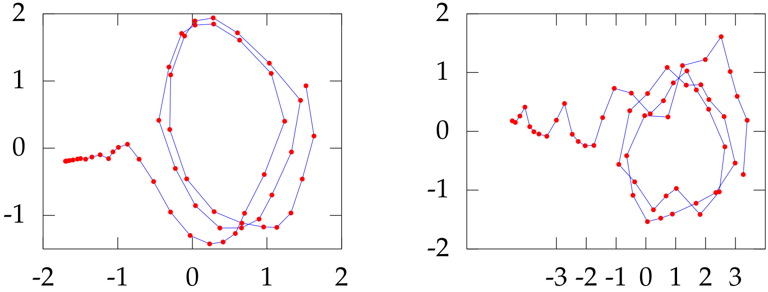
Laplacian Eigenmaps and LLE



(a) Laplacian Eigenmaps (b) Locally Linear Embedding

Figure: Models capture either the cyclic structure or the structure associated with the start of the run or both parts.

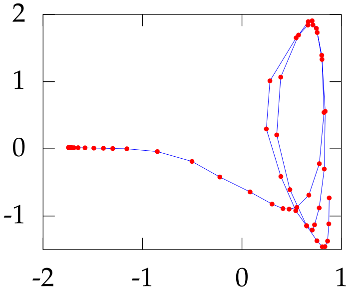
Isomap and MVU



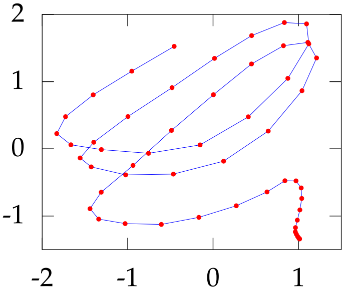
(a) Isomap (b) MVU

Figure: Models capture either the cyclic structure or the structure associated with the start of the run or both parts.

MEU and ALLE



(a) MEU



(b) ALLE

Figure: Models capture either the cyclic structure or the structure associated with the start of the run or both parts.

Motion Capture: Model Scores

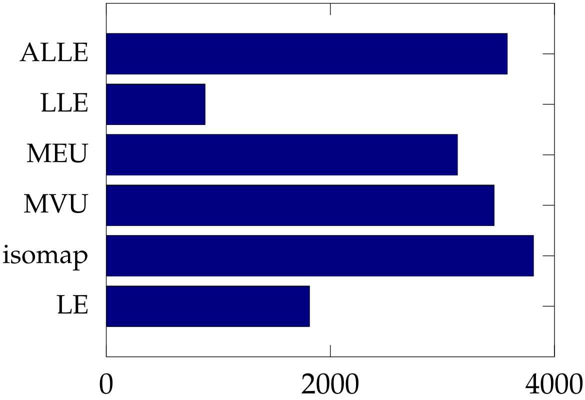


Figure: Model score for the different spectral approaches.

Robot Navigation Example

- Second data set: series of recordings from a robot as it traces a square path in a building.

Robot Navigation Example

- Second data set: series of recordings from a robot as it traces a square path in a building.
- It records the strength of WiFi signals (see Ferris et al., 2007, for an application).

Robot Navigation Example

- Second data set: series of recordings from a robot as it traces a square path in a building.
- It records the strength of WiFi signals (see Ferris et al., 2007, for an application).
- Robot only in two dimensions, the inherent dimensionality of the data should be two.

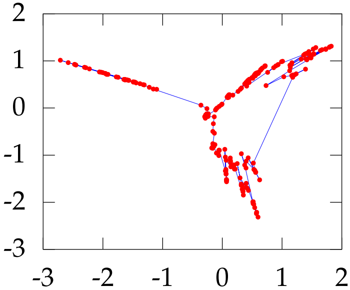
Robot Navigation Example

- Second data set: series of recordings from a robot as it traces a square path in a building.
- It records the strength of WiFi signals (see Ferris et al., 2007, for an application).
- Robot only in two dimensions, the inherent dimensionality of the data should be two.
- Robot completes a single circuit after entry: it is expected to exhibit "loop closure".

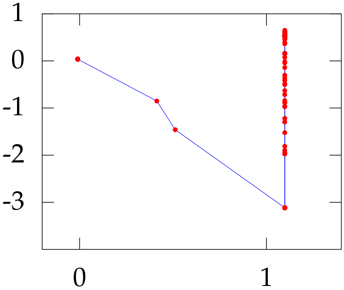
Robot Navigation Example

- Second data set: series of recordings from a robot as it traces a square path in a building.
- It records the strength of WiFi signals (see Ferris et al., 2007, for an application).
- Robot only in two dimensions, the inherent dimensionality of the data should be two.
- Robot completes a single circuit after entry: it is expected to exhibit "loop closure".
- Data consists of 215 frames of measurement of WiFi signal strength of 30 access points.

Laplacian Eigenmaps and LLE



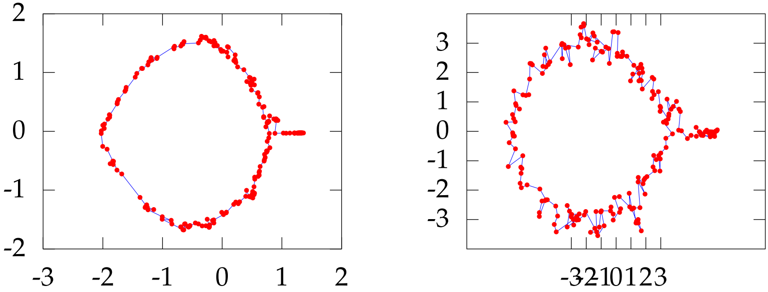
(a) Laplacian Eigenmaps



(b) Locally Linear Embedding

Figure: Models show loop closure but smooth the trace to different degrees.

Isomap and MVU



(a) Isomap (b) MVU

Figure: Models show loop closure but smooth the trace to different degrees.

MEU and DRILL

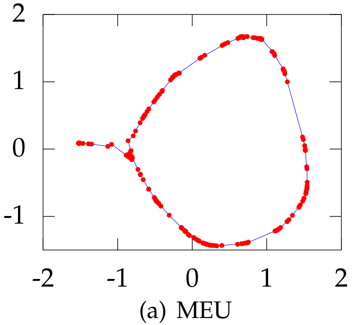


Figure: Models show loop closure but smooth the trace to different degrees.

Robot Navigation: Model Scores

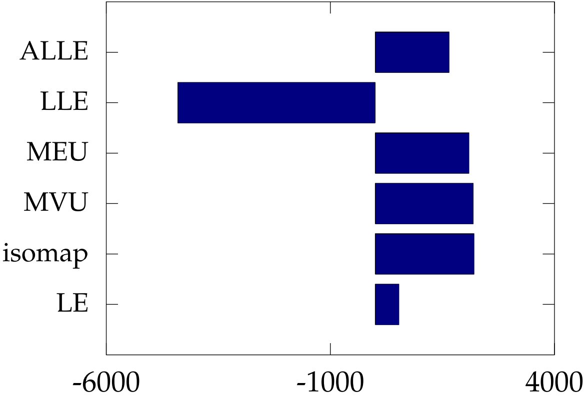


Figure: Model score for the different spectral approaches.

Outline

Review

Relation to Laplacian Eigenmaps

Relation to Locally Linear Embedding

Relation to Isomap

Relation to GP-LVM

Experiments

Discussion

- New perspective on dimensionality reduction algorithms based around maximum entropy.
- Start with MVU and end with GRFs.
- Hope that this perspective on dimensionality reduction will encourage new strands of research at the interface of these areas.

Stages of Spectral Dimensionality Reduction

- Our perspective shows there are three separate stages used in existing spectral dimensionality algorithms.
  1. A neighborhood between data points is selected. Normally k-nearest neighbors or similar algorithms are used.
  2. Interpoint distances between neighbors are fed to the algorithms which provide a similarity matrix. The way the entries in the similarity matrix are computed is the main difference between the different algorithms.
  3. The relationship between points in the similarity matrix is visualized using the eigenvectors of the similarity matrix.

Our Perspective

- Each step is somewhat orthogonal.
- Neighborhood relations need not come from nearest neighbors: can use structure learning.
- Main difference between approaches is how similarity matrix entries are determiend.
- Final step attempts to visualize the similarity using eigenvectors. This is just one possible approach.
- There is an entire field of graph visualization proposing different approaches to visualizing such graphs.

Advantages of Existing Approaches

- Conflating the three steps allows faster complete algorithms.
- E.g. mixing 2nd & 3rd allows speed ups by never computing the similarity matrix.
- We still can understand the algorithm from the unifying perspective while exploiting the computational advantages offered by this neat shortcut.

Other Points

- ALLE may provide a good compromise in speed vs accuracy.
- Also looked at structural learning.
  - See Lawrence (2012) for more details.

Acknowledgements

Conversations with John Kent, Chris Williams, Brenden Lake, Joshua Tenenbaum and John Lafferty have influenced the thinking in this work.

References I

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. [DOI].

B. D. Ferris, D. Fox, and N. D. Lawrence. WiFi-SLAM using Gaussian process latent variable models. In M. M. Veloso, editor, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 2480–2485, 2007. [PDF].

R. Greiner and D. Schuurmans, editors. *Proceedings of the International Conference in Machine Learning*, volume 21, 2004. Omnipress.

J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of dimensionality reduction of manifolds. In Greiner and Schuurmans (2004). [ǫlorbluePDF].

S. Harmeling. Exploring model selection techniques for nonlinear dimensionality reduction. Technical Report EDI-INF-RR-0960, University of Edinburgh,

E. T. Jaynes. Bayesian methods: General background. In J. H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25. Cambridge University Press, 1986.

C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proc. Natl. Acad. Sci. USA*, 105(31), 2008.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. [Google Books] .

N. D. Lawrence. Gaussian process models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 329–336, Cambridge, MA, 2004. MIT Press.

N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 11 2005.

N. D. Lawrence. A unifying probabilistic perspective for spectral dimensionality reduction. Technical report, University of Sheffield, [PDF].

N. D. Lawrence. Spectral dimensionality reduction via maximum entropy. In G. Gordon and D. Dunson, editors, *Proceedings of the Fourteenth International Workshop on Artificial Intelligence and Statistics*, volume 15, pages 51–59, Fort Lauderdale, FL, USA, 11-13 April 2011. JMLR W&CP 15. [PDF]. Notable Paper.

N. D. Lawrence. A unifying probabilistic perspective for spectral dimensionality reduction: Insights and new models. *Journal of Machine Learning Research*, 13, 2012. [URL]. [PDF].

References II

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, London, 1979. [Google Books] .

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, 2000. [DOI].

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998. [DOI].

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2001.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. [DOI].

L. A. Wasserman. *All of Statistics*. Springer-Verlag, New York, 2003. [Google Books] .

K. Q. Weinberger, F. Sha, and L. K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In Greiner and Schuurmans (2004), pages 839–846.

X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning: From Gaussian fields to Gaussian processes. Technical Report CMU-CS-03-175, Carnegie Mellon University, [PDF].

Outline

Learning the Neighborhood

Final Experiment: Structure Learning

- Test the ability of L1 regularization of the random field to learn the neighborhood.
- Considered the motion capture data and used the DRILL with a neighborhood size of 20 and full connectivity.
- L1 regularization on the parameters: vary regularization size and seek a maximum under the GPLVM.

Structure Learning from Neighborhood of 20

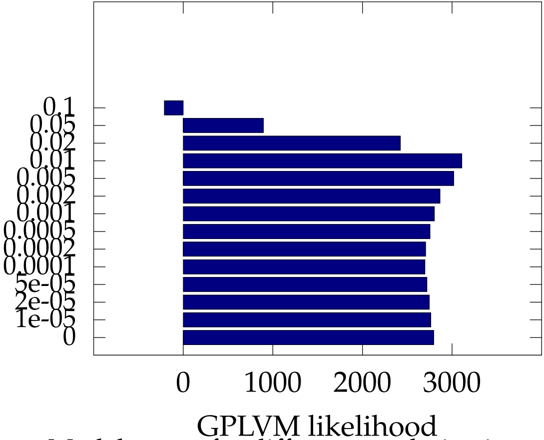


Figure: Model scores for different regularization coefficients.

Structure Learning from Neighborhood of 20

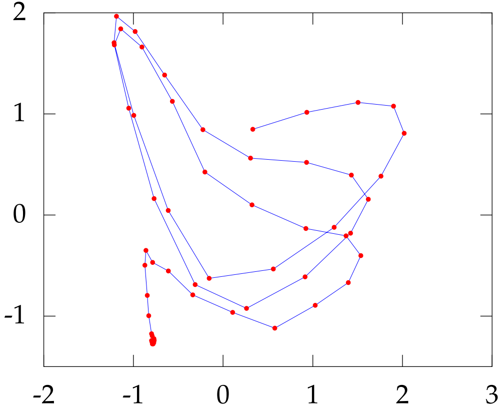


Figure: Visualization associated with highest model score.

Full Structure Learning

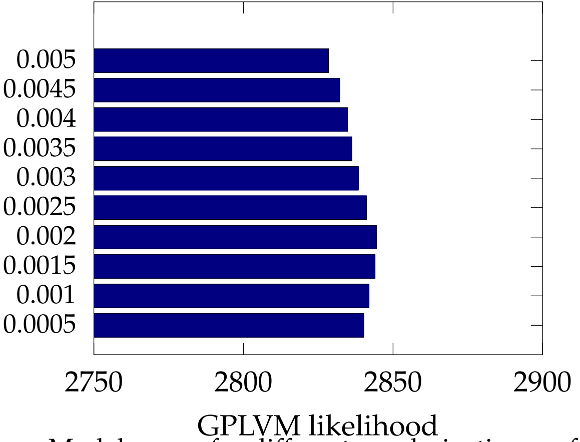


Figure: Model scores for different regularization coefficients.

Full Structure Learning

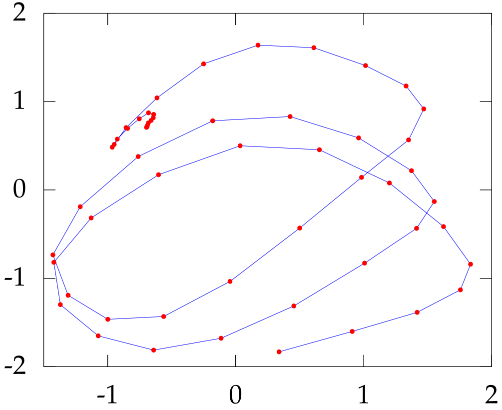


Figure: Visualization associated with highest model score.

Different Neighborhood Scores

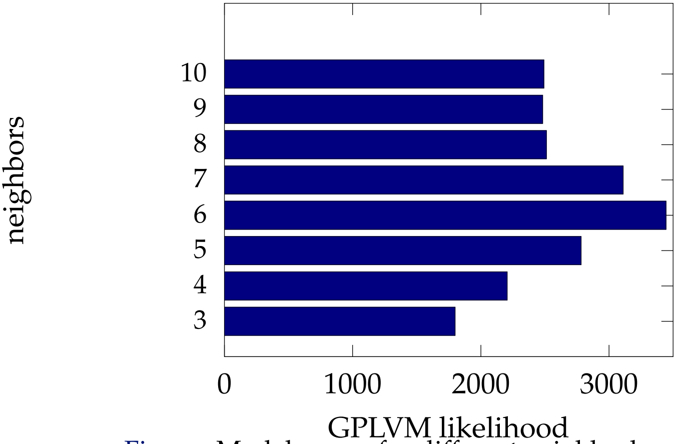


Figure: Model scores for different neighborhood sizes.

Different Neighborhood Scores

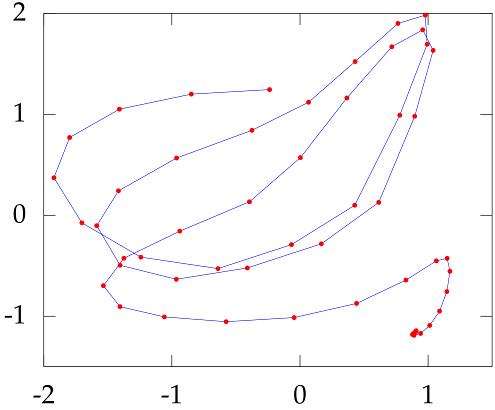


Figure: Visualization associated with highest model score.

Structure Learning from Neighborhood of 6

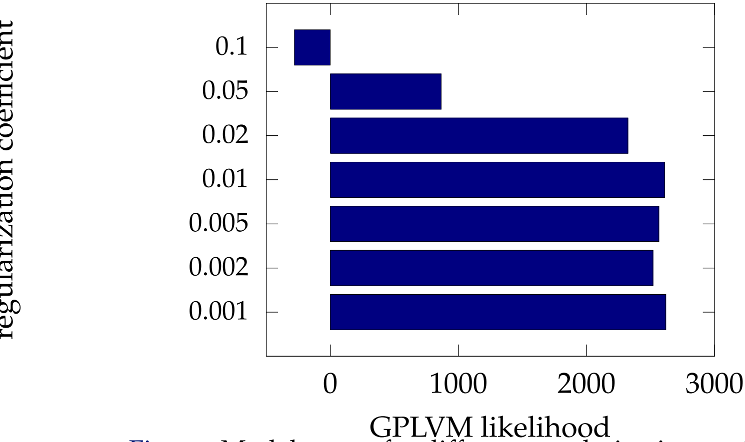


Figure: Model scores for different regularization coefficients.

Structure Learning from Neighborhood of 6

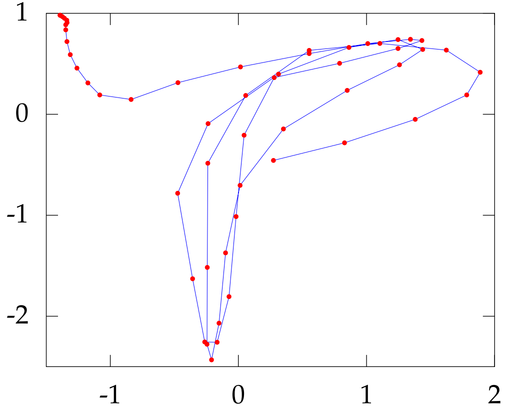


Figure: Visualization associated with highest model score.