

Unsupervised Learning with Gaussian Processes

Neil D. Lawrence

GPSS
17th September 2014



Outline

Motivating Example

Linear Dimensionality Reduction

Non-linear Dimensionality Reduction

Outline

Motivating Example

Linear Dimensionality Reduction

Non-linear Dimensionality Reduction

Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

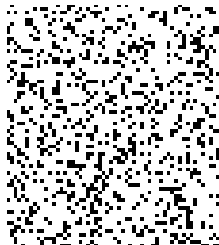
- ▶ 3648 Dimensions
 - ▶ 64 rows by 57 columns



Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

- ▶ 3648 Dimensions
 - ▶ 64 rows by 57 columns
 - ▶ Space contains more than just this digit.



Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

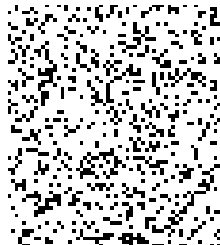
- ▶ 3648 Dimensions
 - ▶ 64 rows by 57 columns
 - ▶ Space contains more than just this digit.
 - ▶ Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

- ▶ 3648 Dimensions
 - ▶ 64 rows by 57 columns
 - ▶ Space contains more than just this digit.
 - ▶ Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



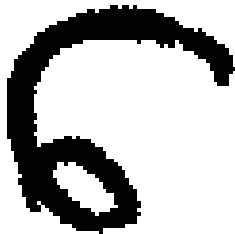
Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'



Simple Model of Digit

Rotate a 'Prototype'

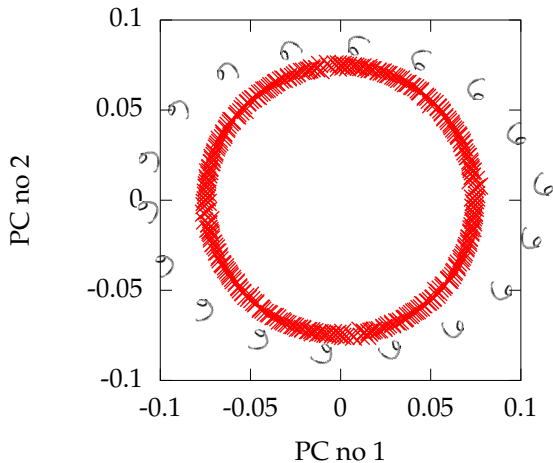


MATLAB Demo

```
demDigitsManifold([1 2], 'all')
```

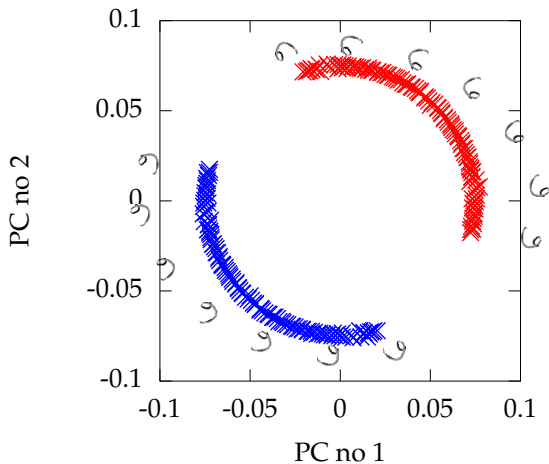
MATLAB Demo

```
demDigitsManifold([1 2], 'all')
```



MATLAB Demo

```
demDigitsManifold([1 2], 'sixnine')
```



Pure Rotation is too Simple

- ▶ In practice the data may undergo several distortions.
 - ▶ *e.g.* digits undergo 'thinning', translation and rotation.
- ▶ For data with 'structure':
 - ▶ we expect fewer distortions than dimensions;
 - ▶ we therefore expect the data to live on a lower dimensional manifold.
- ▶ Conclusion: deal with high dimensional data by looking for lower dimensional non-linear embedding.

Outline

Motivating Example

Linear Dimensionality Reduction

Non-linear Dimensionality Reduction

Notation

q — dimension of latent/embedded space

p — dimension of data space

n — number of data points

data, $\mathbf{Y} = [\mathbf{y}_{1,:}, \dots, \mathbf{y}_{n,:}]^T = [\mathbf{y}_{:,1}, \dots, \mathbf{y}_{:,p}] \in \mathcal{R}^{n \times p}$

centred data, $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_{1,:}, \dots, \hat{\mathbf{y}}_{n,:}]^T = [\hat{\mathbf{y}}_{:,1}, \dots, \hat{\mathbf{y}}_{:,p}] \in \mathcal{R}^{n \times p}$,

$$\hat{\mathbf{y}}_{i,:} = \mathbf{y}_{i,:} - \boldsymbol{\mu}$$

latent variables, $\mathbf{X} = [\mathbf{x}_{1,:}, \dots, \mathbf{x}_{n,:}]^T = [\mathbf{x}_{:,1}, \dots, \mathbf{x}_{:,q}] \in \mathcal{R}^{n \times q}$

mapping matrix, $\mathbf{W} \in \mathcal{R}^{p \times q}$

$\mathbf{a}_{i,:}$ is a vector from the i th row of a given matrix \mathbf{A}

$\mathbf{a}_{:,j}$ is a vector from the j th row of a given matrix \mathbf{A}

Reading Notation

\mathbf{X} and \mathbf{Y} are *design matrices*

- ▶ Data covariance given by $\frac{1}{n} \hat{\mathbf{Y}}^\top \hat{\mathbf{Y}}$

$$\text{cov}(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{y}}_{i,:} \hat{\mathbf{y}}_{i,:}^\top = \frac{1}{n} \hat{\mathbf{Y}}^\top \hat{\mathbf{Y}} = \mathbf{S}.$$

- ▶ Inner product matrix given by $\mathbf{Y}\mathbf{Y}^\top$

$$\mathbf{K} = (k_{i,j})_{i,j}, \quad k_{i,j} = \mathbf{y}_{i,:}^\top \mathbf{y}_{j,:}$$

Linear Dimensionality Reduction

- ▶ Find a lower dimensional plane embedded in a higher dimensional space.
- ▶ The plane is described by the matrix $\mathbf{W} \in \mathbb{R}^{p \times q}$.

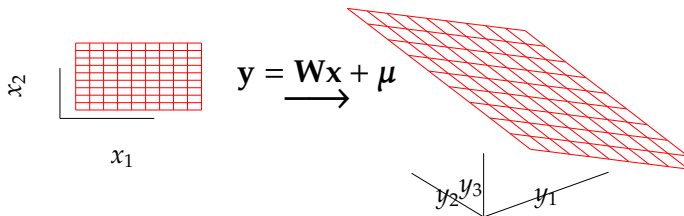


Figure : Mapping a two dimensional plane to a higher dimensional space in a linear way. Data are generated by corrupting points on the plane with noise.

Linear Latent Variable Model

- ▶ Represent data, \mathbf{Y} , with a lower dimensional set of latent variables \mathbf{X} .
- ▶ Assume a linear relationship of the form

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:},$$

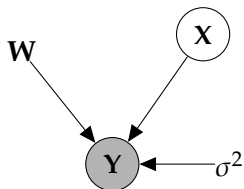
where

$$\boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

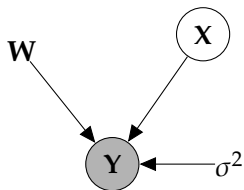


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Standard** Latent variable approach:

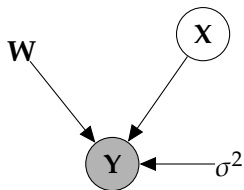


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Standard** Latent variable approach:
 - ▶ Define Gaussian prior over *latent space*, \mathbf{X} .



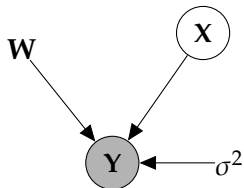
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{X}) = \prod_{i=1}^n \mathcal{N}(\mathbf{x}_{i,:} | \mathbf{0}, \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Standard Latent variable approach:**
 - ▶ Define Gaussian prior over *latent space*, \mathbf{X} .
 - ▶ Integrate out *latent variables*.



$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{X}) = \prod_{i=1}^n \mathcal{N}(\mathbf{x}_{i,:} | \mathbf{0}, \mathbf{I})$$

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I})$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\mathbf{W}\mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\top),$$

Computation of the Marginal Likelihood

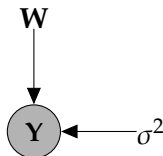
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\mathbf{W}\mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\top),$$

$$\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I})$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)



$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i:}|\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i:}|\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2} \log |\mathbf{C}| - \frac{1}{2} \text{tr}(\mathbf{C}^{-1}\mathbf{Y}^T\mathbf{Y}) + \text{const.}$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i:}|\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2} \log |\mathbf{C}| - \frac{1}{2} \text{tr}(\mathbf{C}^{-1}\mathbf{Y}^T\mathbf{Y}) + \text{const.}$$

If \mathbf{U}_q are first q principal eigenvectors of $n^{-1}\mathbf{Y}^T\mathbf{Y}$ and the corresponding eigenvalues are Λ_q ,

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2} \log |\mathbf{C}| - \frac{1}{2} \text{tr}(\mathbf{C}^{-1}\mathbf{Y}^\top\mathbf{Y}) + \text{const.}$$

If \mathbf{U}_q are first q principal eigenvectors of $n^{-1}\mathbf{Y}^\top\mathbf{Y}$ and the corresponding eigenvalues are Λ_q ,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^\top, \quad \mathbf{L} = (\Lambda_q - \sigma^2\mathbf{I})^{\frac{1}{2}}$$

where \mathbf{R} is an arbitrary rotation matrix.

Outline

Motivating Example

Linear Dimensionality Reduction

Non-linear Dimensionality Reduction

Difficulty for Probabilistic Approaches

- ▶ Propagate a probability distribution through a non-linear mapping.
- ▶ Normalisation of distribution becomes intractable.

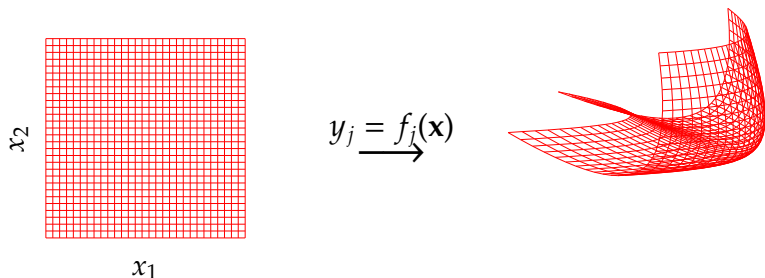


Figure : A three dimensional manifold formed by mapping from a two dimensional space to a three dimensional space.

Difficulty for Probabilistic Approaches

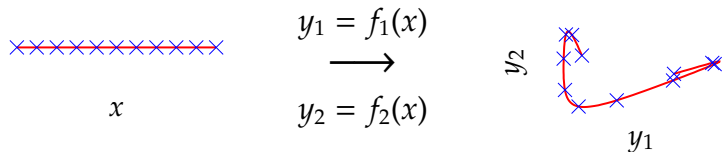


Figure : A string in two dimensions, formed by mapping from one dimension, x , line to a two dimensional space, $[y_1, y_2]$ using nonlinear functions $f_1(\cdot)$ and $f_2(\cdot)$.

Difficulty for Probabilistic Approaches

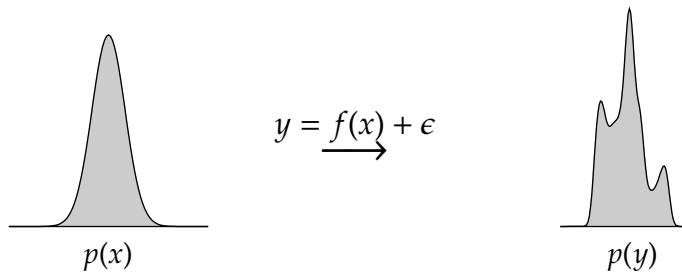
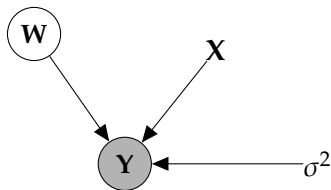


Figure : A Gaussian distribution propagated through a non-linear mapping. $y_i = f(x_i) + \epsilon_i$. $\epsilon \sim \mathcal{N}(0, 0.2^2)$ and $f(\cdot)$ uses RBF basis, 100 centres between -4 and 4 and $\ell = 0.1$. New distribution over y (right) is multimodal and difficult to normalize.

Linear Latent Variable Model III

Dual Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.

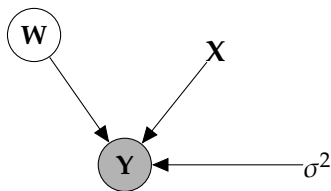


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_i; \mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I})$$

Linear Latent Variable Model III

Dual Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Novel** Latent variable approach:

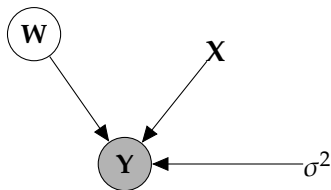


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_i; \mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I})$$

Linear Latent Variable Model III

Dual Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Novel** Latent variable approach:
 - ▶ Define Gaussian prior over *parameters*, \mathbf{W} .



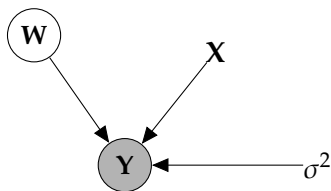
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{W}) = \prod_{i=1}^p \mathcal{N}(\mathbf{w}_{i,:} | \mathbf{0}, \mathbf{I})$$

Linear Latent Variable Model III

Dual Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Novel** Latent variable approach:
 - ▶ Define Gaussian prior over *parameters*, \mathbf{W} .
 - ▶ Integrate out *parameters*.



$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{W}) = \prod_{i=1}^p \mathcal{N}(\mathbf{w}_{i,:} | \mathbf{0}, \mathbf{I})$$

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

Computation of the Marginal Likelihood

$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\mathbf{X}\mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\mathbf{X}^\top),$$

Computation of the Marginal Likelihood

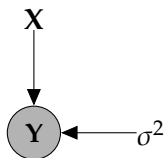
$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\mathbf{X}\mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\mathbf{X}^\top),$$

$$\mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})$$

Linear Latent Variable Model IV

Dual Probabilistic PCA Max. Likelihood Soln (Lawrence, 2004, 2005)



$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})$$

Linear Latent Variable Model IV

Dual PPCA Max. Likelihood Soln (Lawrence, 2004, 2005)

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K}), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}$$

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K}), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{X}) = -\frac{p}{2} \log |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^\top) + \text{const.}$$

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{X}) = -\frac{p}{2} \log |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^\top) + \text{const.}$$

If \mathbf{U}'_q are first q principal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ and the corresponding eigenvalues are Λ_q ,

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{X}) = -\frac{p}{2} \log |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^\top) + \text{const.}$$

If \mathbf{U}'_q are first q principal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ and the corresponding eigenvalues are Λ_q ,

$$\mathbf{X} = \mathbf{U}'_q \mathbf{L} \mathbf{R}^\top, \quad \mathbf{L} = (\Lambda_q - \sigma^2 \mathbf{I})^{\frac{1}{2}}$$

where \mathbf{R} is an arbitrary rotation matrix.

Linear Latent Variable Model IV

Dual PPCA Max. Likelihood Soln (Lawrence, 2004, 2005)

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}), \quad \mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{X}) = -\frac{p}{2} \log |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^\top) + \text{const.}$$

If \mathbf{U}'_q are first q principal eigenvectors of $p^{-1}\mathbf{Y}\mathbf{Y}^\top$ and the corresponding eigenvalues are Λ_q ,

$$\mathbf{X} = \mathbf{U}'_q \mathbf{L} \mathbf{R}^\top, \quad \mathbf{L} = (\Lambda_q - \sigma^2 \mathbf{I})^{\frac{1}{2}}$$

where \mathbf{R} is an arbitrary rotation matrix.

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2} \log |\mathbf{C}| - \frac{1}{2} \text{tr}(\mathbf{C}^{-1}\mathbf{Y}^\top\mathbf{Y}) + \text{const.}$$

If \mathbf{U}_q are first q principal eigenvectors of $n^{-1}\mathbf{Y}^\top\mathbf{Y}$ and the corresponding eigenvalues are Λ_q ,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^\top, \quad \mathbf{L} = (\Lambda_q - \sigma^2\mathbf{I})^{\frac{1}{2}}$$

where \mathbf{R} is an arbitrary rotation matrix.

Equivalence of Formulations

The Eigenvalue Problems are equivalent

- ▶ Solution for Probabilistic PCA (solves for the mapping)

$$\mathbf{Y}^\top \mathbf{Y} \mathbf{U}_q = \mathbf{U}_q \mathbf{\Lambda}_q \quad \mathbf{W} = \mathbf{U}_q \mathbf{L} \mathbf{R}^\top$$

- ▶ Solution for Dual Probabilistic PCA (solves for the latent positions)

$$\mathbf{Y} \mathbf{Y}^\top \mathbf{U}'_q = \mathbf{U}'_q \mathbf{\Lambda}_q \quad \mathbf{X} = \mathbf{U}'_q \mathbf{L} \mathbf{R}^\top$$

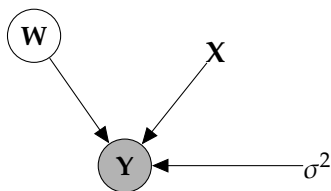
- ▶ Equivalence is from

$$\mathbf{U}_q = \mathbf{Y}^\top \mathbf{U}'_q \mathbf{\Lambda}_q^{-\frac{1}{2}}$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- ▶ Define *linear-Gaussian relationship* between latent variables and data.
- ▶ **Novel** Latent variable approach:
 - ▶ Define Gaussian prior over *parameters*, \mathbf{W} .
 - ▶ Integrate out *parameters*.



$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

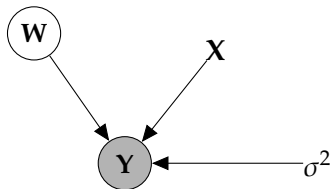
$$p(\mathbf{W}) = \prod_{i=1}^p \mathcal{N}(\mathbf{w}_{i,:} | \mathbf{0}, \mathbf{I})$$

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- ▶ Inspection of the marginal likelihood shows ...

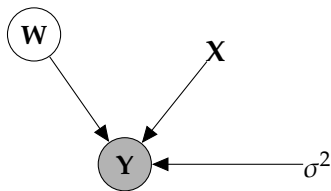


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(y_{:,j} | \mathbf{0}, \mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I})$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- ▶ Inspection of the marginal likelihood shows ...
 - ▶ The covariance matrix is a covariance function.



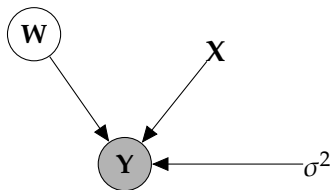
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

$$\mathbf{K} = \mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I}$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- ▶ Inspection of the marginal likelihood shows ...
 - ▶ The covariance matrix is a covariance function.
 - ▶ We recognise it as the 'linear kernel'.



$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(y_{:,j} | \mathbf{0}, \mathbf{K})$$

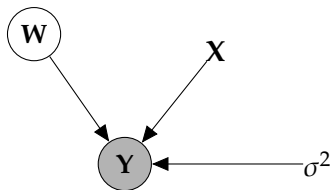
$$\mathbf{K} = \mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I}$$

This is a product of Gaussian processes with linear kernels.

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- ▶ Inspection of the marginal likelihood shows ...
 - ▶ The covariance matrix is a covariance function.
 - ▶ We recognise it as the 'linear kernel'.
 - ▶ We call this the Gaussian Process Latent Variable model (GP-LVM).



$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

$\mathbf{K} = ?$

Replace linear kernel with non-linear kernel for non-linear model.

Non-linear Latent Variable Models

Exponentiated Quadratic (EQ) Covariance

- ▶ The EQ covariance has the form $k_{i,j} = k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:})$, where

$$k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}) = \alpha \exp\left(-\frac{\|\mathbf{x}_{i,:} - \mathbf{x}_{j,:}\|_2^2}{2\ell^2}\right).$$

- ▶ No longer possible to optimise wrt \mathbf{X} via an eigenvalue problem.
- ▶ Instead find gradients with respect to \mathbf{X} , α , ℓ and σ^2 and optimise using conjugate gradients.

Applications

Style Based Inverse Kinematics

- ▶ Facilitating animation through modeling human motion (Grochow et al., 2004)

Tracking

- ▶ Tracking using human motion models (Urtasun et al., 2005, 2006)

Assisted Animation

- ▶ Generalizing drawings for animation (Baxter and Anjyo, 2006)

Shape Models

- ▶ Inferring shape (e.g. pose from silhouette). (Ek et al., 2008b,a; Priacuriu and Reid, 2011a,b)

Generalization with less Data than Dimensions

- ▶ Powerful uncertainty handling of GPs leads to surprising properties.
- ▶ Non-linear models can be used where there are fewer data points than dimensions *without overfitting*.
- ▶ Example: Modelling a stick man in 102 dimensions with 55 data points!

Stick Man II

demStick1

Figure : The latent space for the stick man motion capture data.

Stick Man II

demStick1

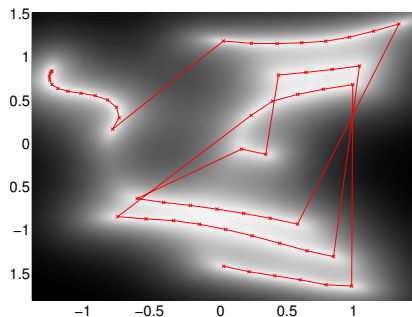


Figure : The latent space for the stick man motion capture data.

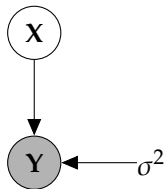
Selecting Data Dimensionality

- ▶ GP-LVM Provides probabilistic non-linear dimensionality reduction.
- ▶ How to select the dimensionality?
- ▶ Need to estimate marginal likelihood.
- ▶ In standard GP-LVM it increases with increasing q .

Integrate Mapping Function and Latent Variables

Bayesian GP-LVM

- ▶ Start with a standard GP-LVM.

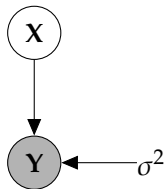


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

Integrate Mapping Function and Latent Variables

Bayesian GP-LVM

- ▶ Start with a standard GP-LVM.
- ▶ Apply standard latent variable approach:
 - ▶ Define Gaussian prior over *latent space*, \mathbf{X} .

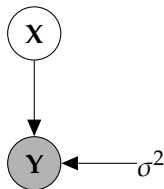


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

Integrate Mapping Function and Latent Variables

Bayesian GP-LVM

- ▶ Start with a standard GP-LVM.
- ▶ Apply standard latent variable approach:
 - ▶ Define Gaussian prior over *latent space*, \mathbf{X} .
 - ▶ Integrate out *latent variables*.



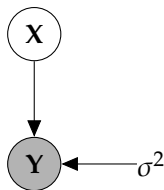
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(y_{:,j} | \mathbf{0}, \mathbf{K})$$

$$p(\mathbf{X}) = \prod_{j=1}^q \mathcal{N}(\mathbf{x}_{:,j} | \mathbf{0}, \alpha_i^{-2} \mathbf{I})$$

Integrate Mapping Function and Latent Variables

Bayesian GP-LVM

- ▶ Start with a standard GP-LVM.
- ▶ Apply standard latent variable approach:
 - ▶ Define Gaussian prior over *latent space*, \mathbf{X} .
 - ▶ Integrate out *latent variables*.
 - ▶ Unfortunately integration is intractable.



$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

$$p(\mathbf{X}) = \prod_{j=1}^q \mathcal{N}(\mathbf{x}_{:,j} | \mathbf{0}, \alpha_j^{-2} \mathbf{I})$$

$$p(\mathbf{Y}|\boldsymbol{\alpha}) = ??$$

Standard Variational Approach Fails

- ▶ Standard variational bound has the form:

$$\mathcal{L} = \langle \log p(\mathbf{y}|\mathbf{X}) \rangle_{q(\mathbf{X})} + \text{KL}(q(\mathbf{X}) \| p(\mathbf{X}))$$

Standard Variational Approach Fails

- ▶ Standard variational bound has the form:

$$\mathcal{L} = \langle \log p(\mathbf{y}|\mathbf{X}) \rangle_{q(\mathbf{X})} + \text{KL}(q(\mathbf{X}) \| p(\mathbf{X}))$$

- ▶ Requires expectation of $\log p(\mathbf{y}|\mathbf{X})$ under $q(\mathbf{X})$.

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top (\mathbf{K}_{f,f} + \sigma^2\mathbf{I})^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_{f,f} + \sigma^2\mathbf{I}| - \frac{n}{2} \log 2\pi$$

Standard Variational Approach Fails

- ▶ Standard variational bound has the form:

$$\mathcal{L} = \langle \log p(\mathbf{y}|\mathbf{X}) \rangle_{q(\mathbf{X})} + \text{KL}(q(\mathbf{X}) \| p(\mathbf{X}))$$

- ▶ Requires expectation of $\log p(\mathbf{y}|\mathbf{X})$ under $q(\mathbf{X})$.

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top (\mathbf{K}_{f,f} + \sigma^2\mathbf{I})^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_{f,f} + \sigma^2\mathbf{I}| - \frac{n}{2} \log 2\pi$$

- ▶ Extremely difficult to compute because $\mathbf{K}_{f,f}$ is dependent on \mathbf{X} and appears in the inverse.

Variational Bayesian GP-LVM

- ▶ Consider collapsed variational bound,

$$p(\mathbf{y}) \geq \prod_{i=1}^n c_i \int \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle, \sigma^2 \mathbf{I}) p(\mathbf{u}) d\mathbf{u}$$

Variational Bayesian GP-LVM

- ▶ Consider collapsed variational bound,

$$p(\mathbf{y}|\mathbf{X}) \geq \prod_{i=1}^n c_i \int \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u}, \mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{u}) d\mathbf{u}$$

Variational Bayesian GP-LVM

- ▶ Consider collapsed variational bound,

$$\int p(\mathbf{y}|\mathbf{X})p(\mathbf{X})d\mathbf{X} \geq \int \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{X}) d\mathbf{X} p(\mathbf{u}) d\mathbf{u}$$

Variational Bayesian GP-LVM

- ▶ Consider collapsed variational bound,

$$\int p(\mathbf{y}|\mathbf{X})p(\mathbf{X})d\mathbf{X} \geq \int \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{X}) d\mathbf{X} p(\mathbf{u}) d\mathbf{u}$$

- ▶ Apply variational lower bound to the inner integral.

Variational Bayesian GP-LVM

- ▶ Consider collapsed variational bound,

$$\int p(\mathbf{y}|\mathbf{X})p(\mathbf{X})d\mathbf{X} \geq \int \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{X}) d\mathbf{X} p(\mathbf{u}) d\mathbf{u}$$

- ▶ Apply variational lower bound to the inner integral.

$$\begin{aligned} \int \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{X}) d\mathbf{X} \\ \geq \left\langle \sum_{i=1}^n \log c_i \right\rangle_{q(\mathbf{X})} \\ + \left\langle \log \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) \right\rangle_{q(\mathbf{X})} \\ + \text{KL}(q(\mathbf{X}) \| p(\mathbf{X})) \end{aligned}$$

Variational Bayesian GP-LVM

- ▶ Consider collapsed variational bound,

$$\int p(\mathbf{y}|\mathbf{X})p(\mathbf{X})d\mathbf{X} \geq \int \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{X}) d\mathbf{X} p(\mathbf{u}) d\mathbf{u}$$

- ▶ Apply variational lower bound to the inner integral.

$$\begin{aligned} \int \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) p(\mathbf{X}) d\mathbf{X} \\ \geq \left\langle \sum_{i=1}^n \log c_i \right\rangle_{q(\mathbf{X})} \\ + \left\langle \log \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{X})}, \sigma^2 \mathbf{I}) \right\rangle_{q(\mathbf{X})} \\ + \text{KL}(q(\mathbf{X}) \| p(\mathbf{X})) \end{aligned}$$

- ▶ Which is analytically tractable for Gaussian $q(\mathbf{X})$ and some covariance functions.

Required Expectations

- ▶ Need expectations under $q(\mathbf{X})$ of:

$$\log c_i = \frac{1}{2\sigma^2} \left[k_{i,i} - \mathbf{k}_{i,\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{k}_{i,\mathbf{u}} \right]$$

and

$$\log \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u},\mathbf{Y})}, \sigma^2 \mathbf{I}) = -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y}_i - \mathbf{K}_{\mathbf{f},\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u})^2$$

- ▶ This requires the expectations

$$\langle \mathbf{K}_{\mathbf{f},\mathbf{u}} \rangle_{q(\mathbf{X})}$$

and

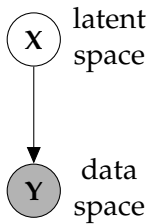
$$\langle \mathbf{K}_{\mathbf{f},\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}} \rangle_{q(\mathbf{X})}$$

which can be computed analytically for some covariance functions.

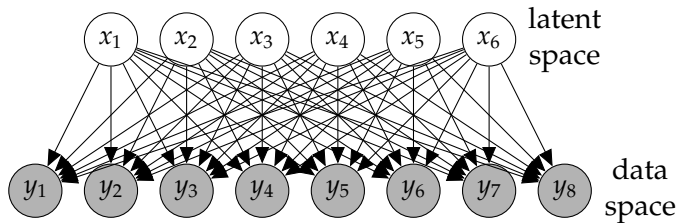
Titsias and Lawrence (2010)

- ▶ Variational marginalization of \mathbf{X} allows us to learn parameters of $p(\mathbf{X})$.
- ▶ Standard GP-LVM where \mathbf{X} learnt by MAP, this is not possible (see e.g. Wang et al., 2008).
- ▶ First example: learn the dimensionality of latent space.

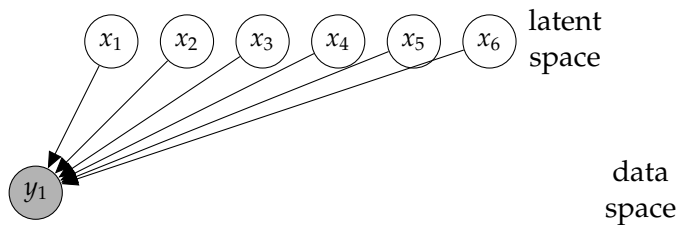
Graphical Representations of GP-LVM



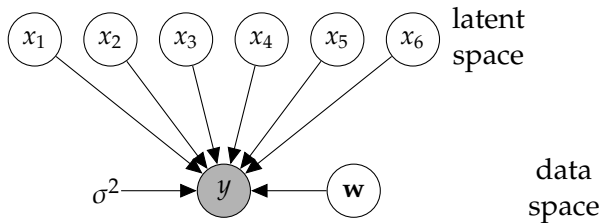
Graphical Representations of GP-LVM



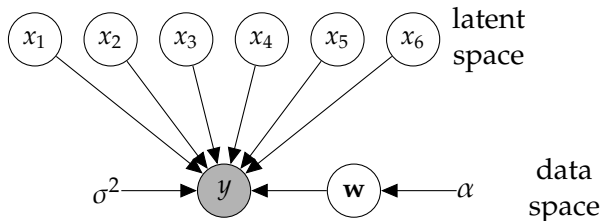
Graphical Representations of GP-LVM



Graphical Representations of GP-LVM



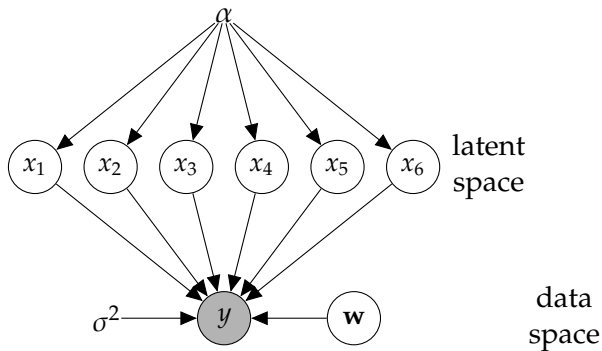
Graphical Representations of GP-LVM



$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I}) \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$$

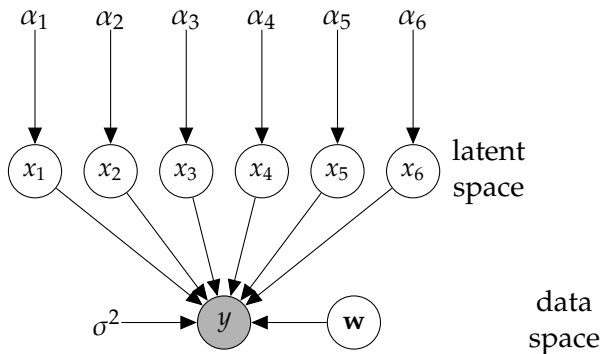
Graphical Representations of GP-LVM



$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$$

$$y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$$

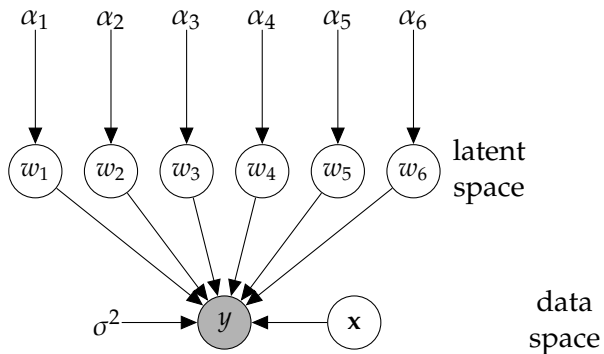
Graphical Representations of GP-LVM



$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad x_i \sim \mathcal{N}(0, \alpha_i)$$

$$y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$$

Graphical Representations of GP-LVM



$$w_i \sim \mathcal{N}(0, \alpha_i) \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$$

Non-linear $f(\mathbf{x})$

- ▶ In linear case equivalence because $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$

$$p(w_i) \sim \mathcal{N}(\mathbf{0}, \alpha_i)$$

- ▶ In non linear case, need to scale columns of \mathbf{X} in prior for $f(\mathbf{x})$.
- ▶ This implies scaling columns of \mathbf{X} in covariance function

$$k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}) = \exp\left(-\frac{1}{2}(\mathbf{x}_{:,i} - \mathbf{x}_{:,j})^\top \mathbf{A}(\mathbf{x}_{:,i} - \mathbf{x}_{:,j})\right)$$

\mathbf{A} is diagonal with elements α_i^2 . Now keep prior spherical

$$p(\mathbf{X}) = \prod_{j=1}^q \mathcal{N}(\mathbf{x}_{:,j} | \mathbf{0}, \mathbf{I})$$

- ▶ Covariance functions of this type are known as ARD (see e.g. Neal, 1996; MacKay, 2003; Rasmussen and Williams, 2006).

Other Priors on X

- ▶ Dynamical prior gives us Gaussian process dynamical system (Wang et al., 2006; Damianou et al., 2011)
- ▶ Structured learning prior gives us (soft) manifold sharing (Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008b,a; Damianou et al., 2012)
- ▶ Gaussian process prior gives us Deep Gaussian Processes (Lawrence and Moore, 2007; Damianou and Lawrence, 2013)

References I

- W. V. Baxter and K.-I. Anjyo. Latent doodle space. In *EUROGRAPHICS*, volume 25, pages 477–485, Vienna, Austria, September 4-8 2006.
- A. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In J. Langford and J. Pineau, editors, *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kaufman. [PDF].
- A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31. [PDF].
- A. Damianou, M. K. Titsias, and N. D. Lawrence. Variational Gaussian process dynamical systems. In P. Bartlett, F. Peirre, C. Williams, and J. Lafferty, editors, *Advances in Neural Information Processing Systems*, volume 24, Cambridge, MA, 2011. MIT Press. [PDF].
- C. H. Ek, J. Rihan, P. Torr, G. Rogez, and N. D. Lawrence. Ambiguity modeling in latent spaces. In A. Popescu-Belis and R. Stiefelhagen, editors, *Machine Learning for Multimodal Interaction (MLMI 2008)*, LNCS, pages 62–73. Springer-Verlag, 28–30 June 2008a. [PDF].
- C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction (MLMI 2007)*, volume 4892 of LNCS, pages 132–143, Brno, Czech Republic, 2008b. Springer-Verlag. [PDF].
- K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popovic. Style-based inverse kinematics. In *ACM Transactions on Graphics (SIGGRAPH 2004)*, pages 522–531, 2004.
- 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 and A. J. Moore. Hierarchical Gaussian process latent variable models. In Z. Ghahramani, editor, *Proceedings of the International Conference in Machine Learning*, volume 24, pages 481–488. Omnipress, 2007. [Google Books]. [PDF].

References II

- D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, U.K., 2003. [[Google Books](#)].
- R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society Press, 2007.
- R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996. Lecture Notes in Statistics 118.
- V. Priacuriu and I. D. Reid. Nonlinear shape manifolds as shape priors in level set segmentation and tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011a.
- V. Priacuriu and I. D. Reid. Shared shape spaces. In *IEEE International Conference on Computer Vision (ICCV)*, 2011b.
- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. [[Google Books](#)].
- A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In Weiss et al. (2006).
- M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999. [[PDF](#)]. [[DOI](#)].
- M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. In Y. W. Teh and D. M. Titterton, editors, *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, volume 9, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13–16 May 2010. JMLR W&CP 9. [[PDF](#)].
- R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 238–245, New York, U.S.A., 17–22 Jun. 2006. IEEE Computer Society Press.
- R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *IEEE International Conference on Computer Vision (ICCV)*, pages 403–410, Beijing, China, 17–21 Oct. 2005. IEEE Computer Society Press.
- J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In Weiss et al. (2006).
- J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008. ISSN 0162-8828. [[DOI](#)].
- Y. Weiss, B. Schölkopf, and J. C. Platt, editors. *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.