

Probabilistic Dimensionality Reduction

Neil D. Lawrence
University of Sheffield

Facebook, London
14th April 2016



Outline

Probabilistic Linear Dimensionality Reduction

Non Linear Probabilistic Dimensionality Reduction

Conclusions

Conclusions

Notation

q — dimension of latent/embedded space

p — dimension of data space

n — number of data points

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

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

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

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

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

$\mathbf{a}_{i,:}$ 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 $\mathbf{W} \in \Re^{p \times q}$.

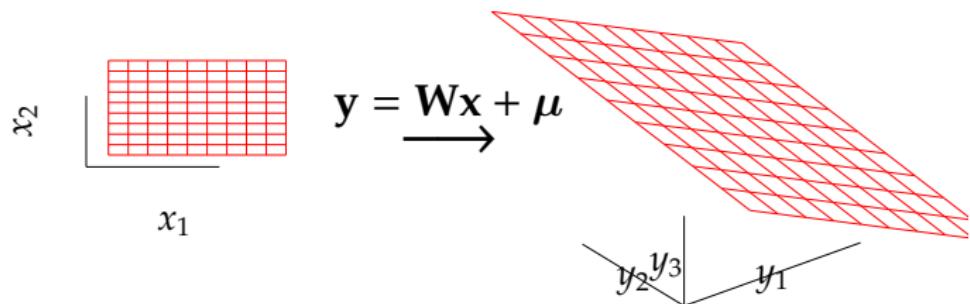


Figure: Mapping a two dimensional plane to a higher dimensional space in a linear way. Data are generated by corrupting points on the plane with noise.

Linear Dimensionality Reduction

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

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

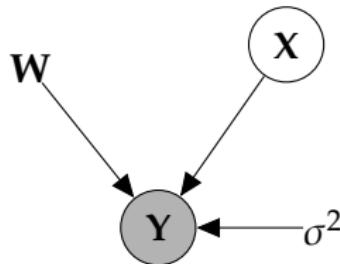
where

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

Linear Latent Variable Model

Probabilistic PCA

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

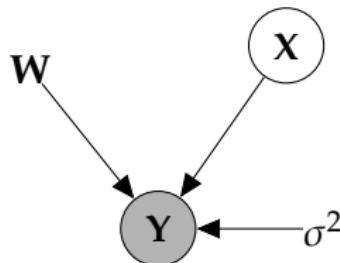


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

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

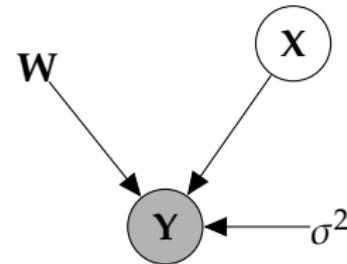


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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} .



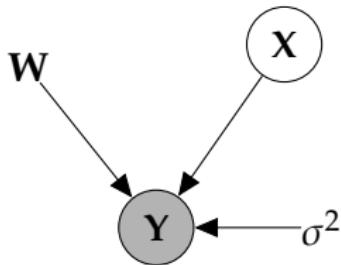
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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

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*.



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

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

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

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

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

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

Computation of the Marginal Likelihood

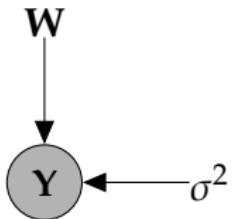
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

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

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

Linear Latent Variable Model II

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



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

Linear Latent Variable Model II

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

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

Linear Latent Variable Model II

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

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

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

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 ,

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.

Linear Latent Variable Model

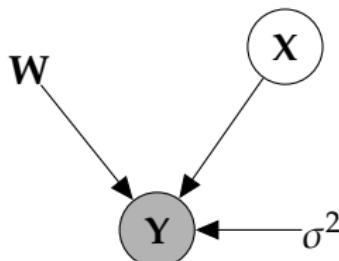
Factor Analysis

- ▶ Linear-Gaussian relationship between latent variables and data,

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}.$$

- ▶ Now each $\eta_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ has a separate variance.

1. Optimize likelihood wrt \mathbf{W} .



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

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

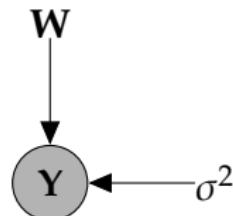
Linear Latent Variable Model

Factor Analysis

- Linear-Gaussian relationship between latent variables and data,

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

- Now each $\eta_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ has a separate variance.
 1. Optimize likelihood wrt \mathbf{W} .



$$p(\hat{\mathbf{Y}}|\mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\hat{\mathbf{y}}_{i,:} | \mathbf{0}, \mathbf{W}\mathbf{W}^\top + \mathbf{D})$$

where \mathbf{D} is diagonal with elements given by σ_j^2 .

Factor Analysis Optimization

- ▶ Optimization is more difficult: no longer an eigenvalue problem.

Linear Latent Variable Model

Independent Component Analysis

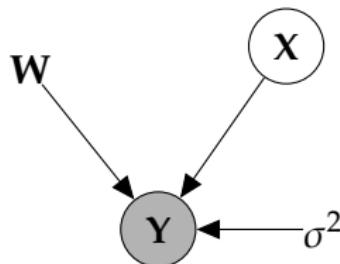
- ▶ Linear-Gaussian relationship between latent variables and data,

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\eta}_{i,:}.$$

- ▶ Now latent variables are independent and non-Gaussian:

$$x_{i,:} \sim \prod_{j=1}^q p(x_{i,j}).$$

1. Optimize likelihood wrt \mathbf{W} .



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

$$p(\mathbf{X}) = \prod_{i=1}^n \prod_{j=1}^p p(x_{i,j})$$

Independent Component Analysis Samples

- Rotational symmetry of Gaussian is removed.

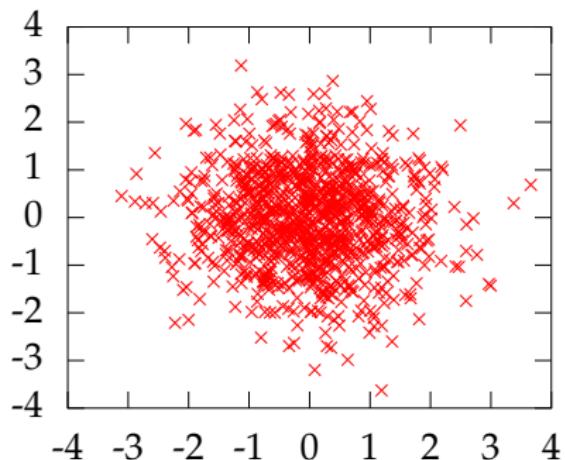


Figure: Independent variables which are Gaussian.

Independent Component Analysis Samples

- Rotational symmetry of Gaussian is removed.

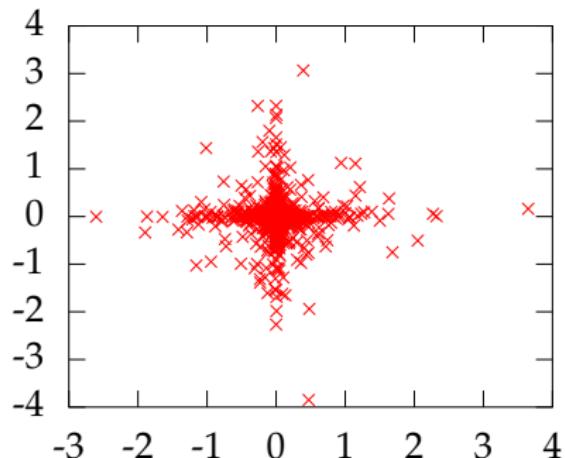


Figure: Independent variables which are super-Gaussian.

Independent Component Analysis Samples

- Rotational symmetry of Gaussian is removed.

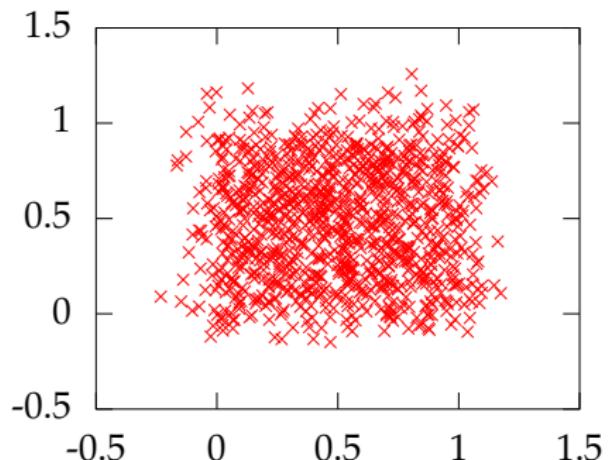


Figure: Independent variables which are sub-Gaussian.

Outline

Probabilistic Linear Dimensionality Reduction

Non Linear Probabilistic Dimensionality Reduction

Conclusions

Conclusions

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'

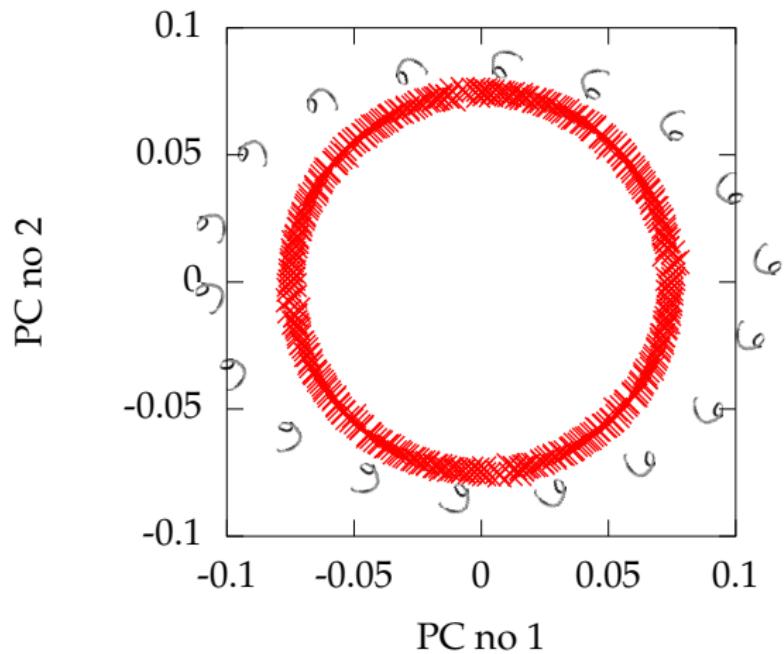


MATLAB Demo

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

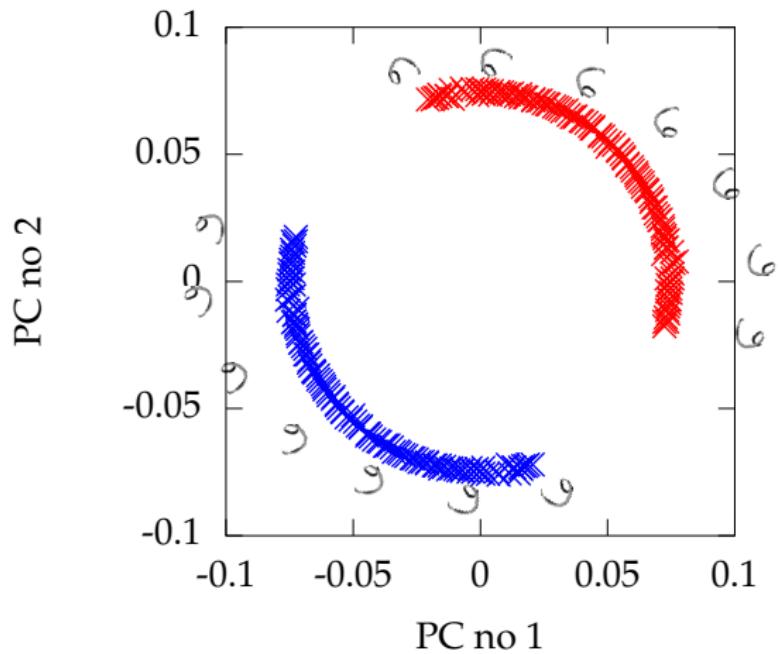
MATLAB Demo

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



MATLAB Demo

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



Low Dimensional Manifolds

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.

Linear Dimensionality Reduction

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

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

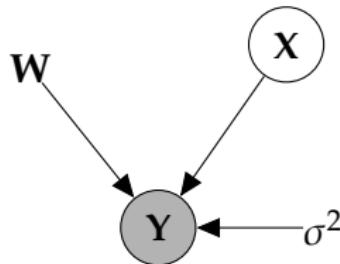
where

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

Linear Latent Variable Model

Probabilistic PCA

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

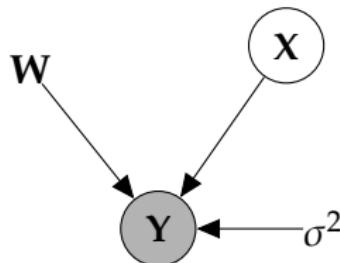


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

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

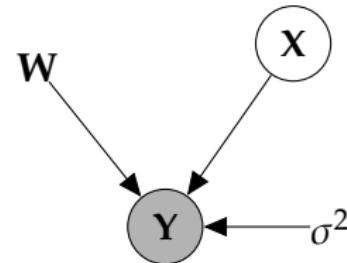


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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} .



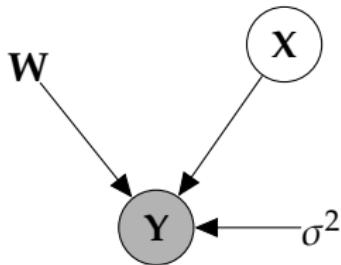
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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

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*.



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

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

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

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

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

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

Computation of the Marginal Likelihood

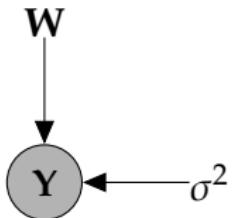
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:}, \quad \mathbf{x}_{i,:} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

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

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

Linear Latent Variable Model II

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



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

Linear Latent Variable Model II

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

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

Linear Latent Variable Model II

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

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

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

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 ,

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 ,

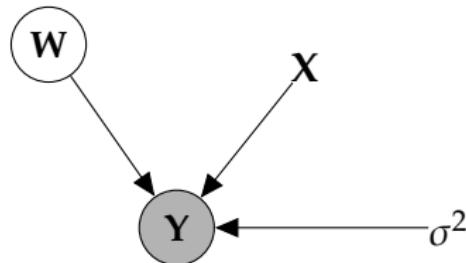
$$\mathbf{W} = \mathbf{U}_q \mathbf{L} \mathbf{R}^\top, \quad \mathbf{L} = (\Lambda_q - \sigma^2 \mathbf{I})^{\frac{1}{2}}$$

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

Linear Latent Variable Model III

Dual Probabilistic PCA

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

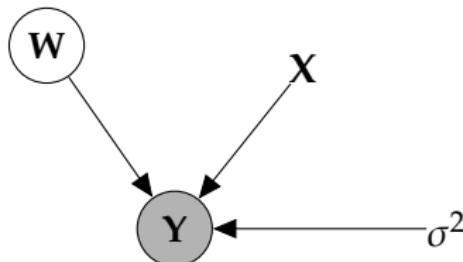


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model III

Dual Probabilistic PCA

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

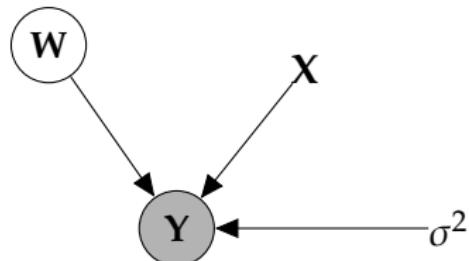


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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} .



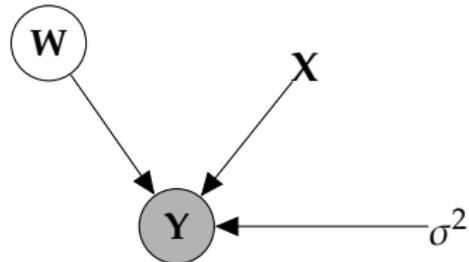
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n N(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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

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

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

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

Computation of the Marginal Likelihood

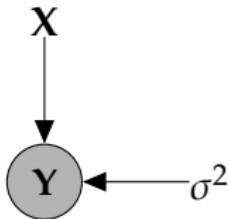
$$\mathbf{y}_{:,j} = \mathbf{X}\mathbf{w}_{:,j} + \boldsymbol{\epsilon}_{:,j}, \quad \mathbf{w}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

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

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

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)

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

Linear Latent Variable Model IV

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

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

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

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln

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

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

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

Linear Latent Variable Model IV

PPCA Max. Likelihood Soln

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

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

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

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

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

Linear Latent Variable Model IV

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

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

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

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

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

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)

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

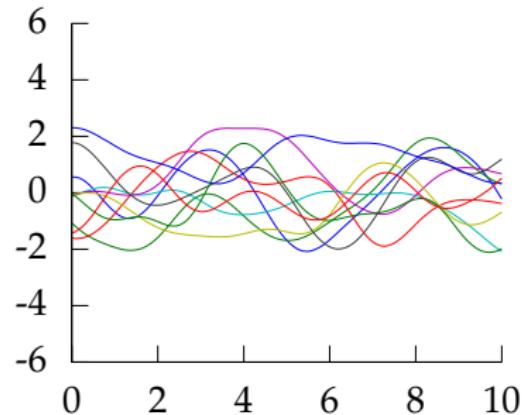
- ▶ Solution for Dual Probabilistic PCA (solves for the latent positions)

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

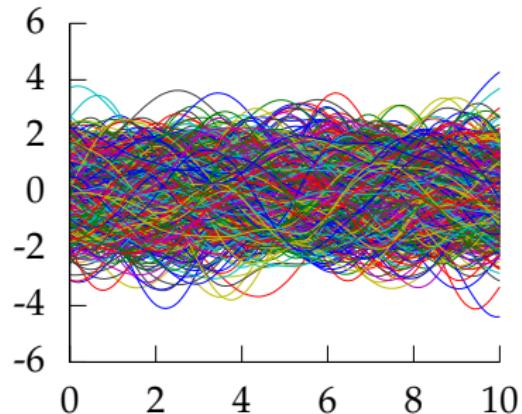
- ▶ Equivalence is from

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

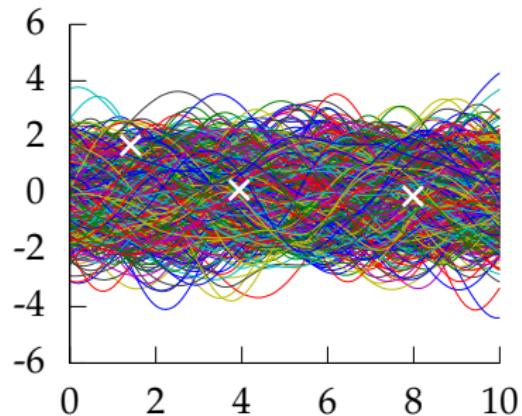
Gaussian Processes: Extremely Short Overview



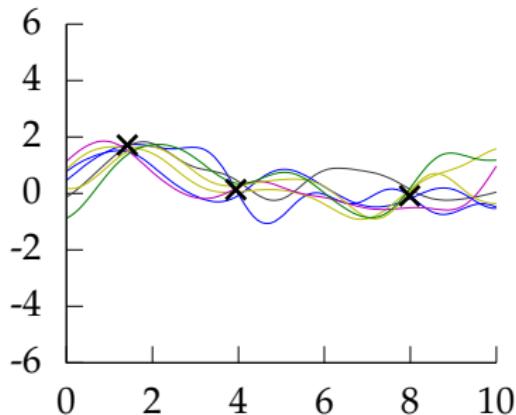
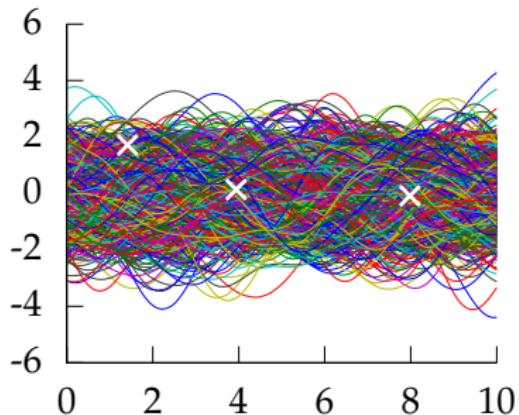
Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



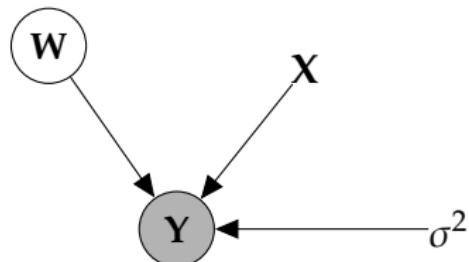
Gaussian Processes: Extremely Short Overview



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*.



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

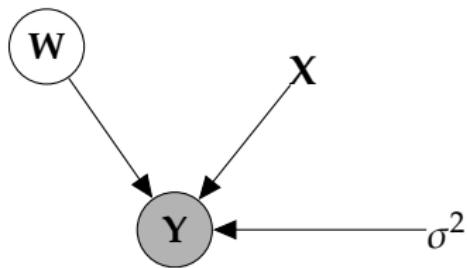
$$p(\mathbf{W}) = \prod_{i=1}^p \mathcal{N}(\mathbf{w}_{i,:} | \mathbf{0}, \mathbf{I})$$

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

Non-Linear Latent Variable Model

Dual Probabilistic PCA

- ▶ Inspection of the marginal likelihood shows ...

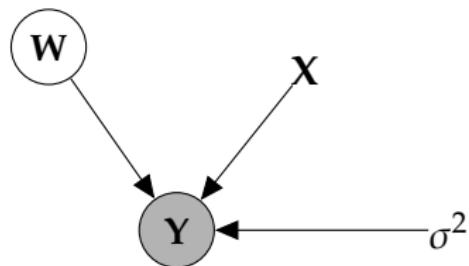


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})$$

Non-Linear Latent Variable Model

Dual Probabilistic PCA

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



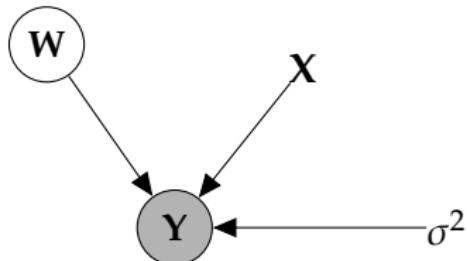
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

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

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'.



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

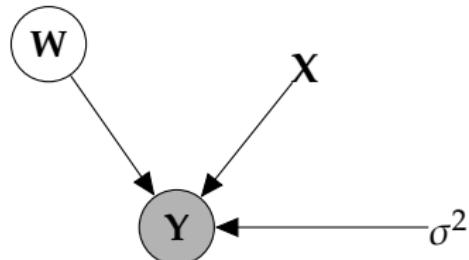
$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}$$

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)

Example: Latent Doodle Space

(Baxter and Anjyo, 2006)

Example: Latent Doodle Space

(Baxter and Anjyo, 2006)

Generalization with much 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*.

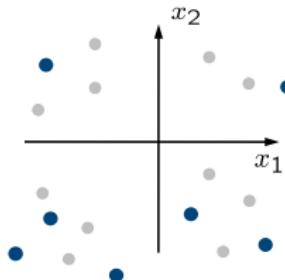
Prior for Supervised Learning

(Urtasun and Darrell, 2007)

- ▶ We introduce a prior that is based on the Fisher criteria

$$p(\mathbf{X}) \propto \exp \left\{ -\frac{1}{\sigma_d^2} \text{tr} (\mathbf{S}_w^{-1} \mathbf{S}_b) \right\},$$

with \mathbf{S}_b the between class matrix and \mathbf{S}_w the within class matrix



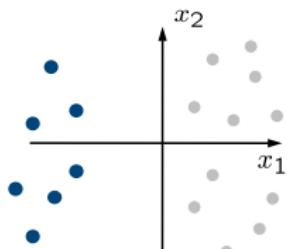
Prior for Supervised Learning

(Urtasun and Darrell, 2007)

- ▶ We introduce a prior that is based on the Fisher criteria

$$p(\mathbf{X}) \propto \exp \left\{ -\frac{1}{\sigma_d^2} \text{tr} (\mathbf{S}_w^{-1} \mathbf{S}_b) \right\},$$

with \mathbf{S}_b the between class matrix and \mathbf{S}_w the within class matrix



$$\mathbf{S}_w = \sum_{i=1}^L \frac{n_i}{n} (\mathbf{M}_i - \mathbf{M}_0)(\mathbf{M}_i - \mathbf{M}_0)^{\top}$$

where $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{n_i}^{(i)}]$ are the n_i training points of class i , \mathbf{M}_i is the mean of the elements of class i , and \mathbf{M}_0 is the mean of all the training points of all classes.

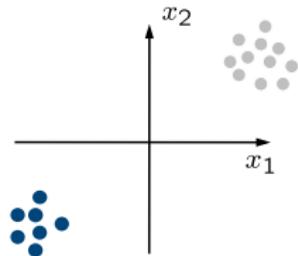
Prior for Supervised Learning

(Urtasun and Darrell, 2007)

- ▶ We introduce a prior that is based on the Fisher criteria

$$p(\mathbf{X}) \propto \exp \left\{ -\frac{1}{\sigma_d^2} \text{tr} (\mathbf{S}_w^{-1} \mathbf{S}_b) \right\},$$

with \mathbf{S}_b the between class matrix and \mathbf{S}_w the within class matrix



$$\mathbf{S}_w = \sum_{i=1}^L \frac{n_i}{n} (\mathbf{M}_i - \mathbf{M}_0)(\mathbf{M}_i - \mathbf{M}_0)^\top$$

$$\mathbf{S}_b = \sum_{i=1}^L \frac{n_i}{n} \left[\frac{1}{n_i} \sum_{k=1}^{n_i} (\mathbf{x}_k^{(i)} - \mathbf{M}_i)(\mathbf{x}_k^{(i)} - \mathbf{M}_i)^\top \right]$$

where $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{n_i}^{(i)}]$ are the n_i training points of class i , \mathbf{M}_i is the mean of the elements of class i , and \mathbf{M}_0 is the

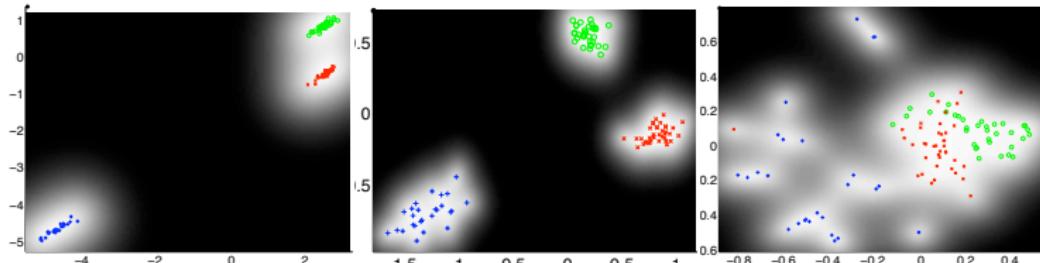
Prior for Supervised Learning

(Urtasun and Darrell, 2007)

- ▶ We introduce a prior that is based on the Fisher criteria

$$p(\mathbf{X}) \propto \exp \left\{ -\frac{1}{\sigma_d^2} \text{tr} (\mathbf{S}_w^{-1} \mathbf{S}_b) \right\},$$

with \mathbf{S}_b the between class matrix and \mathbf{S}_w the within class matrix



GaussianFace

(Lu and Tang, 2014)

- ▶ First system to surpass human performance on cropped Learning Faces in Wild Data.
<http://tinyurl.com/nkt9a38>
- ▶ Lots of feature engineering, followed by a Discriminative GP-LVM.

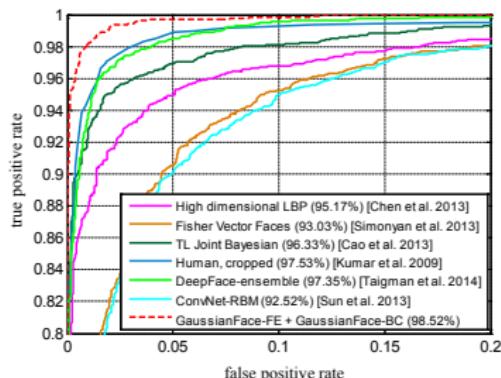


Figure 4: The ROC curve on LFW. Our method achieves the best performance, beating human-level performance.



Figure 5: The two rows present examples of matched and mismatched pairs respectively from LFW that were incorrectly classified by the GaussianFace model.

Conclusion and Future Work

This paper presents a principled Multi-Task Learning ap-

Continuous Character Control

(Levine et al., 2012)

- ▶ Graph diffusion prior for enforcing connectivity between motions.

$$\log p(\mathbf{X}) = w_c \sum_{i,j} \log K_{ij}^d$$

with the graph diffusion kernel \mathbf{K}^d obtain from

$$K_{ij}^d = \exp(\beta \mathbf{H}) \quad \text{with} \quad \mathbf{H} = -\mathbf{T}^{-1/2} \mathbf{L} \mathbf{T}^{-1/2}$$

the graph Laplacian, and \mathbf{T} is a diagonal matrix with $T_{ii} = \sum_j w(\mathbf{x}_i, \mathbf{x}_j)$,

$$L_{ij} = \begin{cases} \sum_k w(\mathbf{x}_i, \mathbf{x}_k) & \text{if } i = j \\ -w(\mathbf{x}_i, \mathbf{x}_j) & \text{otherwise.} \end{cases}$$

and $w(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|^{-p}$ measures similarity.

Character Control: Results

GPLVM for Character Animation

- ▶ Learn a GPLVM from a small mocap sequence
- ▶ Pose synthesis by solving an optimization problem

$$\begin{aligned} & \arg \min \mathbf{x}, \mathbf{y} - \log p(\mathbf{y}|\mathbf{x}) \\ & \text{such that } C(\mathbf{y}) = 0 \end{aligned}$$

- ▶ These handle constraints may come from a user in an interactive session, or from a mocap system.
- ▶ Smooth the latent space by adding noise in order to reduce the number of local minima.
- ▶ Optimization in an annealed fashion over different anneal version of the latent space.

Application: Replay same motion

(Grochow et al., 2004)

Application: Keyframing joint trajectories

(Grochow et al., 2004)

Application: Deal with missing data in mocap

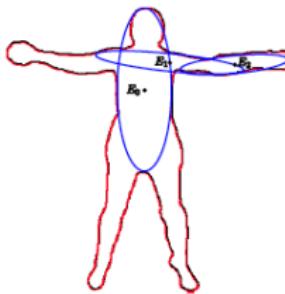
(Grochow et al., 2004)

Application: Style Interpolation

(Grochow et al., 2004)

Shape Priors in Level Set Segmentation

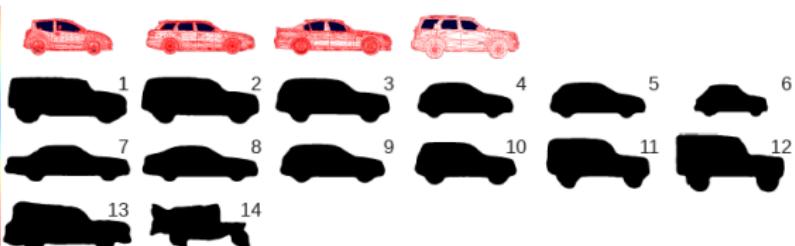
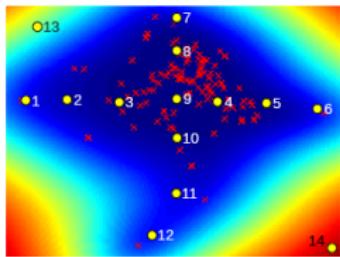
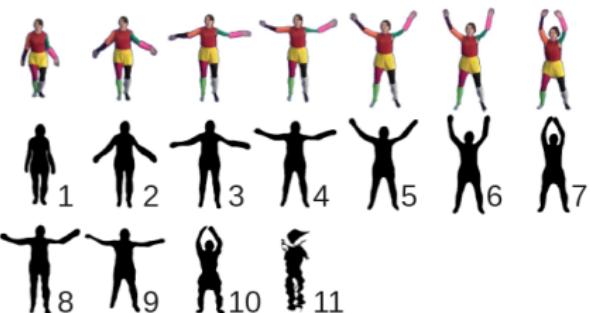
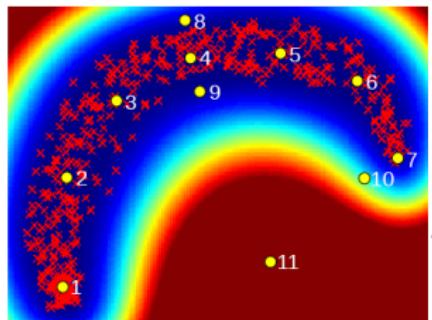
- ▶ Represent contours with elliptic Fourier descriptors



- ▶ Learn a GPLVM on the parameters of those descriptors
- ▶ We can now generate close contours from the latent space
- ▶ Segmentation is done by non-linear minimization of an image-driven energy which is a function of the latent space

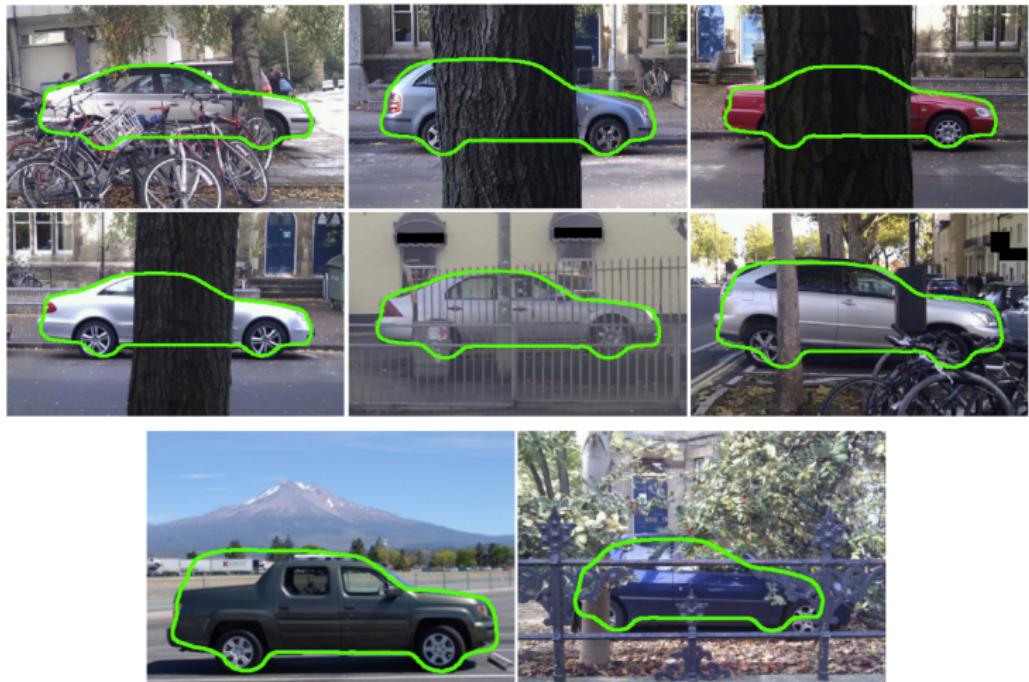
GPLVM on Contours

[V. Prisacariu and I. Reid, ICCV 2011]



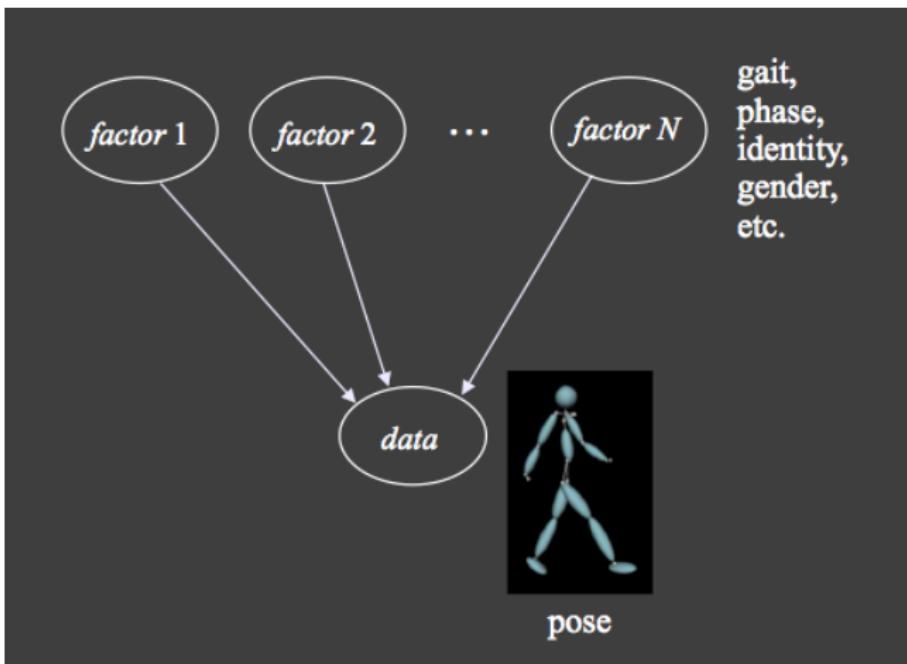
Segmentation Results

[V. Prisacariu and I. Reid, ICCV 2011]



5) Style Content Separation and Multi-linear models

Multiple aspects that affect the input signal, interesting to factorize them



Multilinear models

- ▶ Style-Content Separation (Tenenbaum and Freeman, 2000)

$$\mathbf{y} = \sum_{i,j} w_{i,j} a_i b_j + \epsilon$$

- ▶ Multi-linear analysis (Vasilescu and Terzopoulos, 2002)

$$\mathbf{y} = \sum_{i,j,k,\dots} w_{i,j,k,\dots} a_i b_j c_k \dots + \epsilon$$

- ▶ Non-linear basis functions (Elgammal and Lee, 2004)

$$\mathbf{y} = \sum_{i,j} w_{i,j} a_i \phi_j(b) + \epsilon$$

Multi (non)-linear models with GPs

- ▶ In the GPLVM

$$\mathbf{y} = \sum_j w_j \phi_j(\mathbf{x}) + \epsilon = \mathbf{w}^\top \Phi(\mathbf{x}) + \epsilon$$

with

$$E[\mathbf{y}, \mathbf{y}'] = \Phi(\mathbf{x})^\top \Phi(\mathbf{y}) + \beta^{-1} \delta = k(\mathbf{x}, \mathbf{x}') + \beta^{-1} \delta$$

- ▶ Multifactor Gaussian process

$$\mathbf{y} = \sum_{i,j,k,\dots} w_{ijk\dots} \phi_i^{(1)} \phi_j^{(1)} \phi_k^{(1)} \dots + \epsilon$$

with

$$E[\mathbf{y}, \mathbf{y}'] = \prod_i \Phi^{(i)\top} \Phi^{(i)} + \beta^{-1} \delta = \prod_i k_i(\mathbf{x}^{(i)}, \mathbf{x}^{(i)})' + \beta^{-1} \delta$$

- ▶ Learning in this model is the same, just the kernel changes.

Training Data

Each training motion is a collection of poses, sharing the same combination of subject (s) and gait (g).

Stylistic factors

subject 1

subject 2

subject 3

stride



run



walk



Character Animation

(Wang et al., 2007)

Training data, 6 sequences, 314 frames in total

Generating new styles for a subject

(Wang et al., 2007)

Interpolating Gaits

(Wang et al., 2007)

Generating Different Styles

(Wang et al., 2007)

Other Topics

- ▶ Dynamical models [► Details](#)
- ▶ Hierarchical models [► Details](#)
- ▶ Bayesian GP-LVM [► Details](#)
- ▶ Deep GPs [► Details](#)

Hierarchical GP-LVM

(Lawrence and Moore, 2007)

Stacking Gaussian Processes

- ▶ Regressive dynamics provides a simple hierarchy.
 - ▶ The input space of the GP is governed by another GP.
- ▶ By stacking GPs we can consider more complex hierarchies.
- ▶ Ideally we should marginalise latent spaces
 - ▶ In practice we seek MAP solutions.

Two Correlated Subjects

(Lawrence and Moore, 2007)

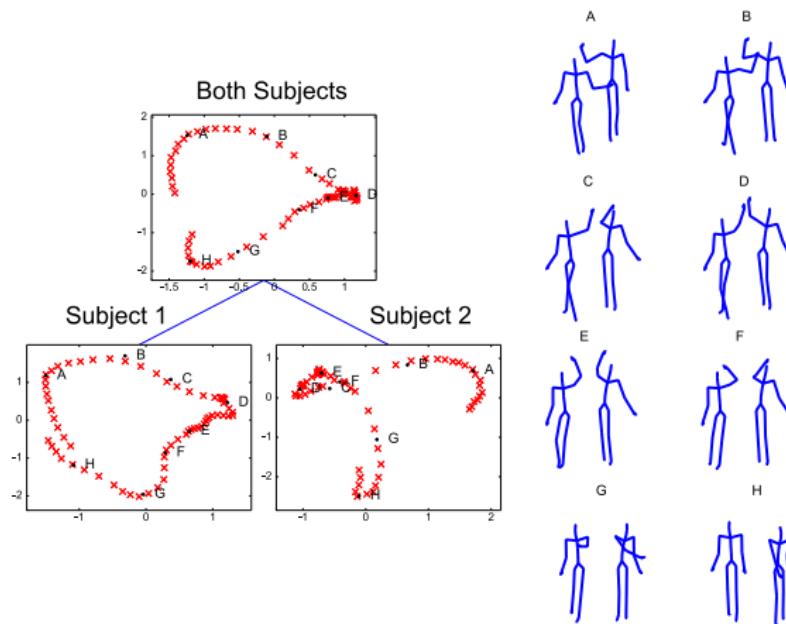


Figure: Hierarchical model of a 'high five'.

Within Subject Hierarchy

(Lawrence and Moore, 2007)

Decomposition of Body

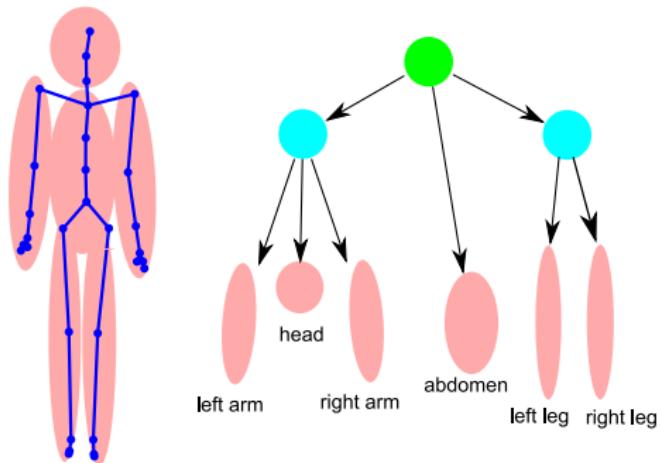


Figure: Decomposition of a subject.

Single Subject Run/Walk

(Lawrence and Moore, 2007)

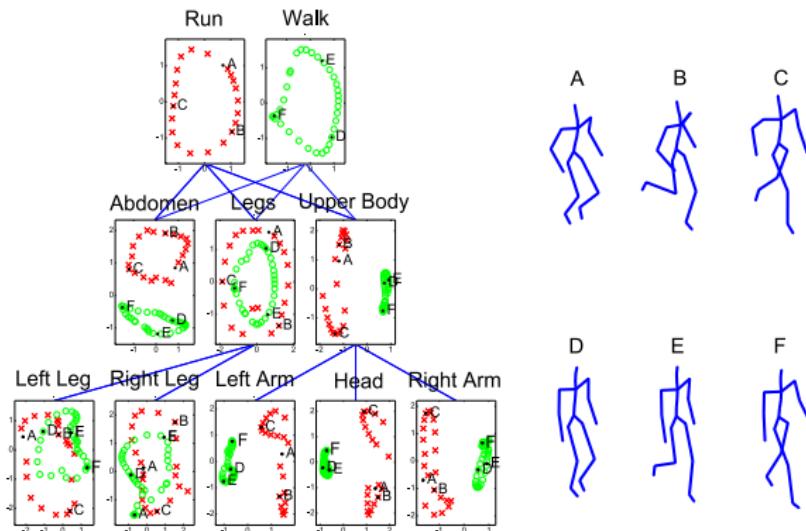


Figure: Hierarchical model of a walk and a run.

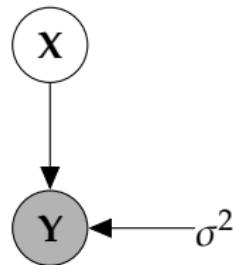
Selecting Data Dimensionality

- ▶ 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.

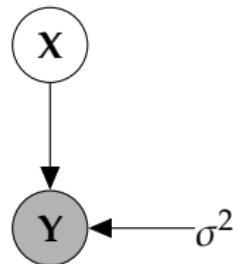


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

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} .

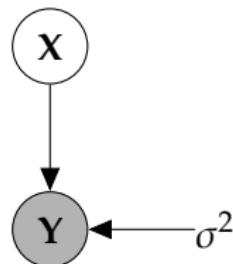


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

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*.



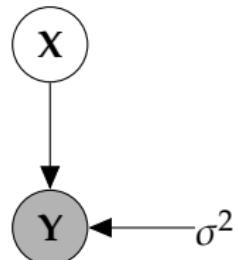
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K})$$

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

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.



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

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

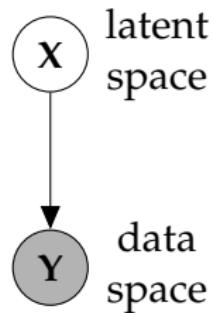
$$p(\mathbf{Y}|\boldsymbol{\alpha}) = ??$$

Priors for Latent Space

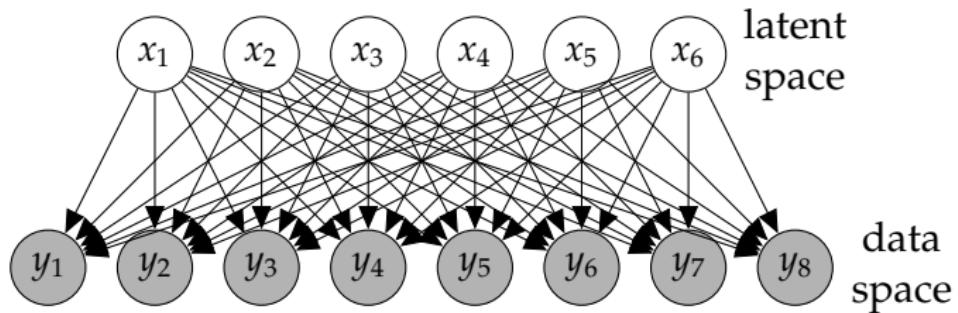
Titsias and Lawrence (2010)

- ▶ 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.

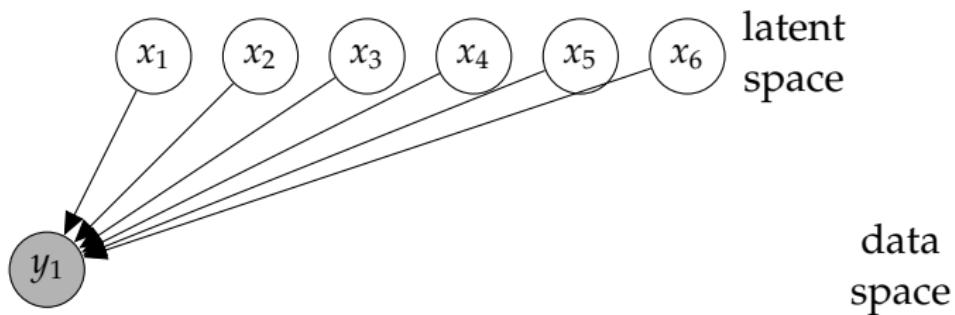
Graphical Representations of GP-LVM



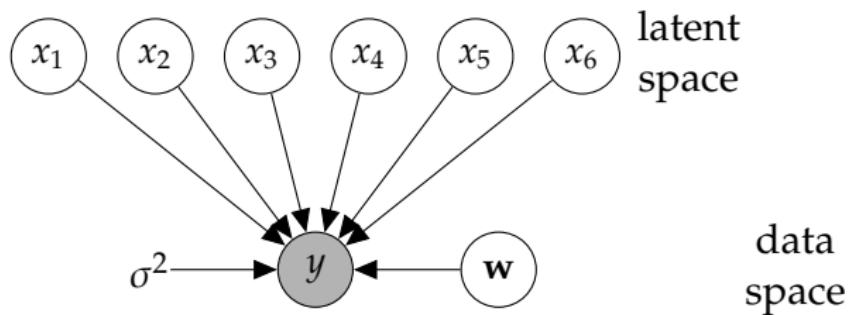
Graphical Representations of GP-LVM



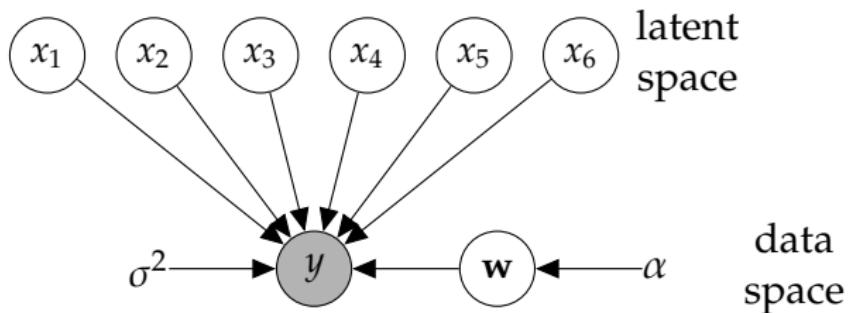
Graphical Representations of GP-LVM



Graphical Representations of GP-LVM



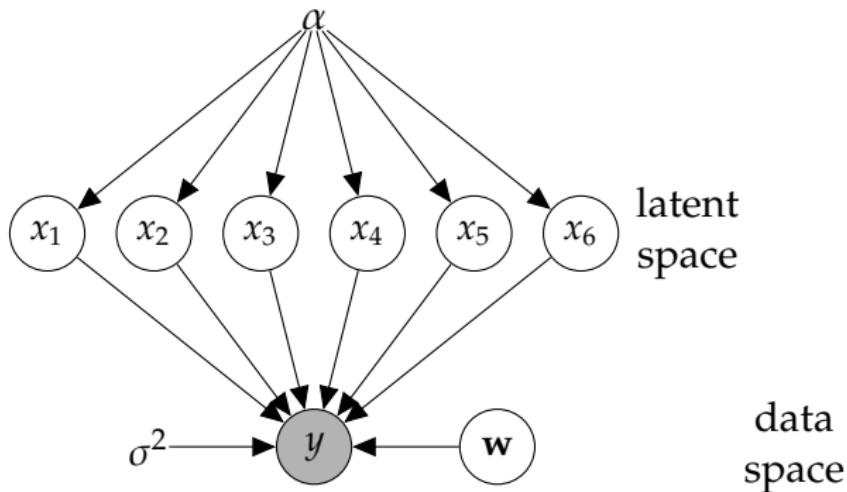
Graphical Representations of GP-LVM



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

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

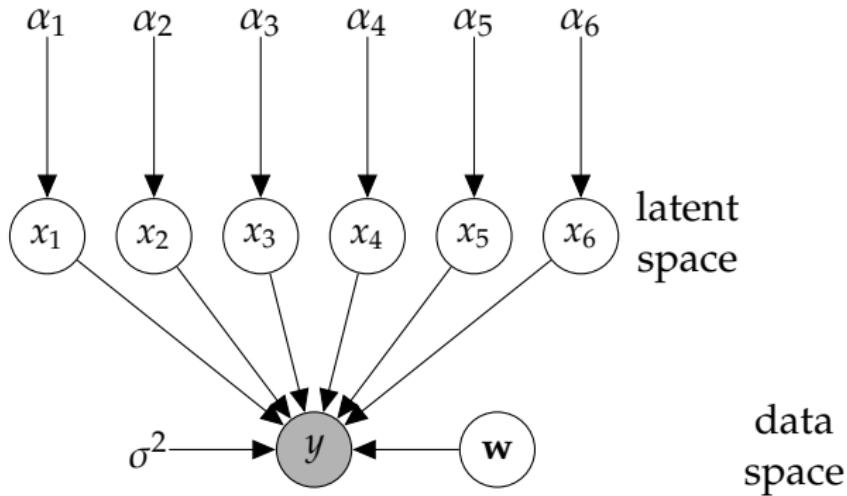
Graphical Representations of GP-LVM



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

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

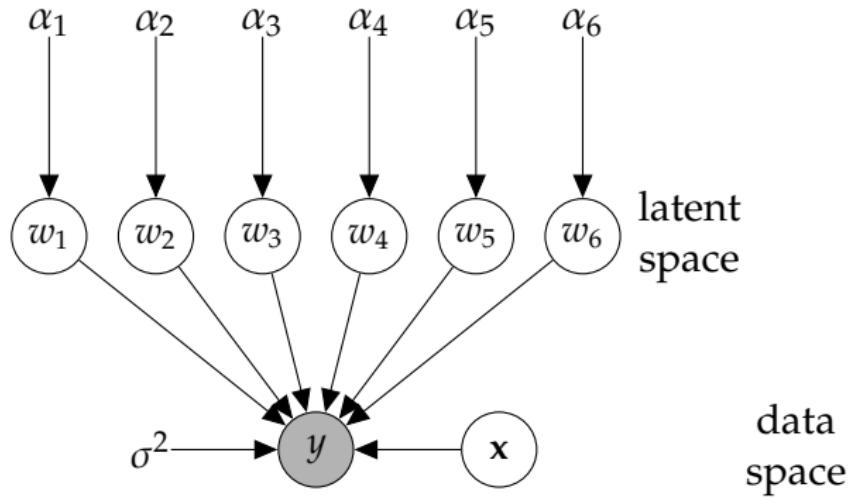
Graphical Representations of GP-LVM



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

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

Graphical Representations of GP-LVM



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

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

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

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

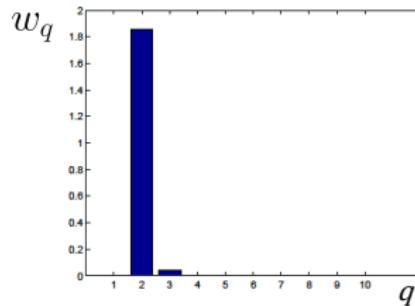
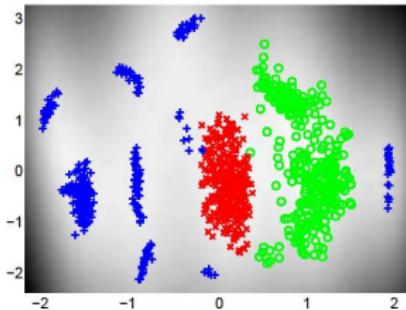
- ▶ Covariance functions of this type are known as ARD (see e.g. Neal, 1996; MacKay, 2003; Rasmussen and Williams, 2006).

Automatic dimensionality detection

- Achieved by employing an *Automatic Relevance Determination (ARD)* covariance function for the prior on the GP mapping
- $f \sim GP(\mathbf{0}, k_f)$ with

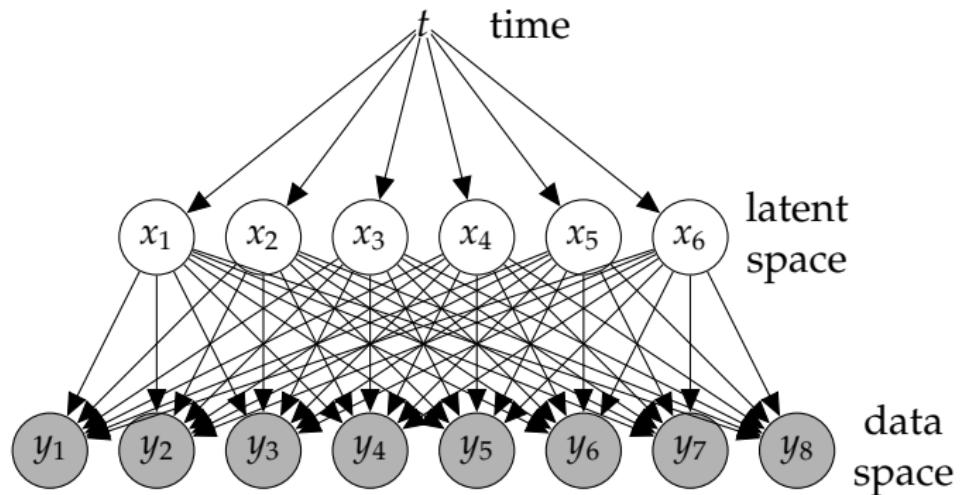
$$k_f(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{1}{2} \sum_{q=1}^Q w_q (x_{i,q} - x_{j,q})^2\right)$$

- Example



Gaussian Process Dynamical Systems

(Damianou et al., 2011)



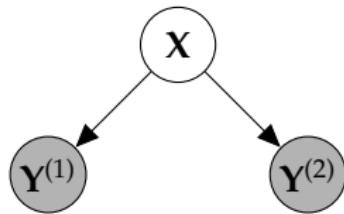
Gaussian Process over Latent Space

- ▶ Assume a GP prior for $p(\mathbf{X})$.
- ▶ Input to the process is time, $p(\mathbf{X}|t)$.

Interpolation of HD Video

Modeling Multiple ‘Views’

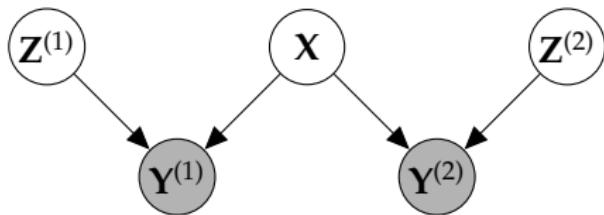
- ▶ Single space to model correlations between two different data sources, e.g., images & text, image & pose.
- ▶ Shared latent spaces: (Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008b)



- ▶ Effective when the ‘views’ are correlated.
- ▶ But not all information is shared between both ‘views’.
- ▶ PCA applied to concatenated data vs CCA applied to data.

Shared-Private Factorization

- ▶ In real scenarios, the ‘views’ are neither fully independent, nor fully correlated.
- ▶ Shared models
 - ▶ either allow information relevant to a single view to be mixed in the shared signal,
 - ▶ or are unable to model such private information.
- ▶ Solution: Model shared and private information (Virtanen et al., 2011; Ek et al., 2008a; Leen and Fyfe, 2006; Klami and Kaski, 2007, 2008; Tucker, 1958)

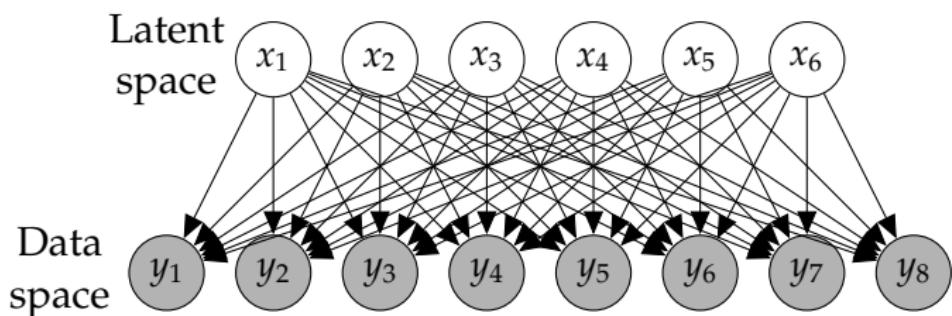


- ▶ Probabilistic CCA is case when dimensionality of \mathbf{Z} matches $\mathbf{Y}^{(i)}$ (cf Inter Battery Factor Analysis (Tucker, 1958)).

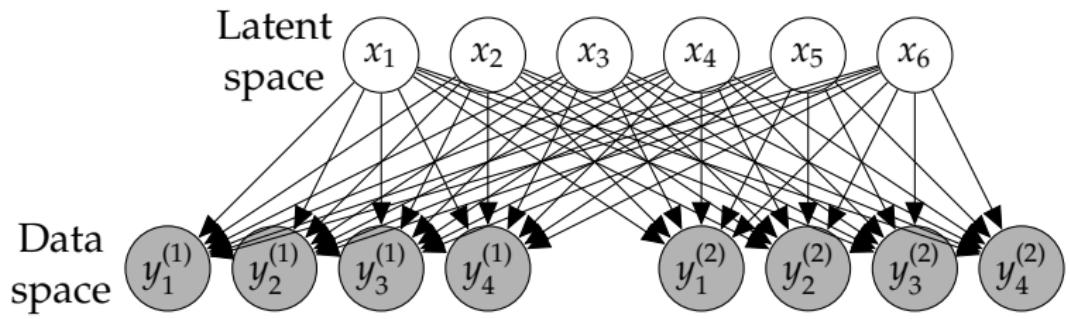
Manifold Relevance Determination



Damianou et al. (2012)



Shared GP-LVM



Separate ARD parameters for mappings to $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$.

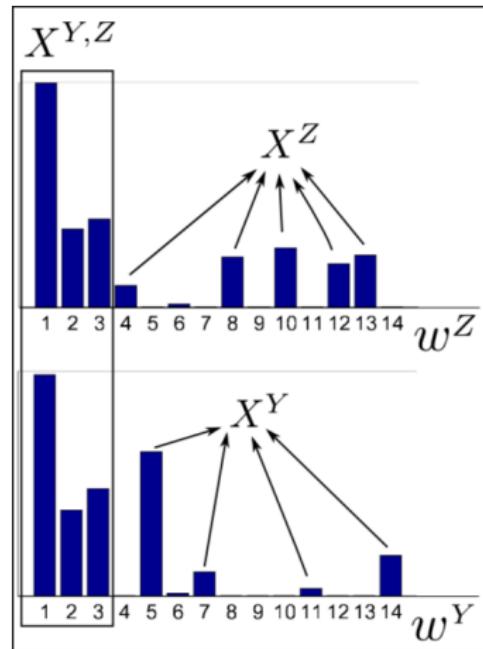
Example: Yale faces



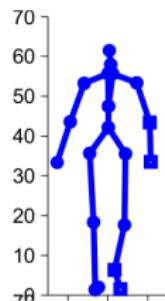
- Dataset Y: 3 persons under all illumination conditions
- Dataset Z: As above for 3 different persons
- Align datapoints \mathbf{x}_n and \mathbf{z}_n only based on the lighting direction

Results

- Latent space X initialised with 14 dimensions
- Weights define a segmentation of X
- Video / demo...

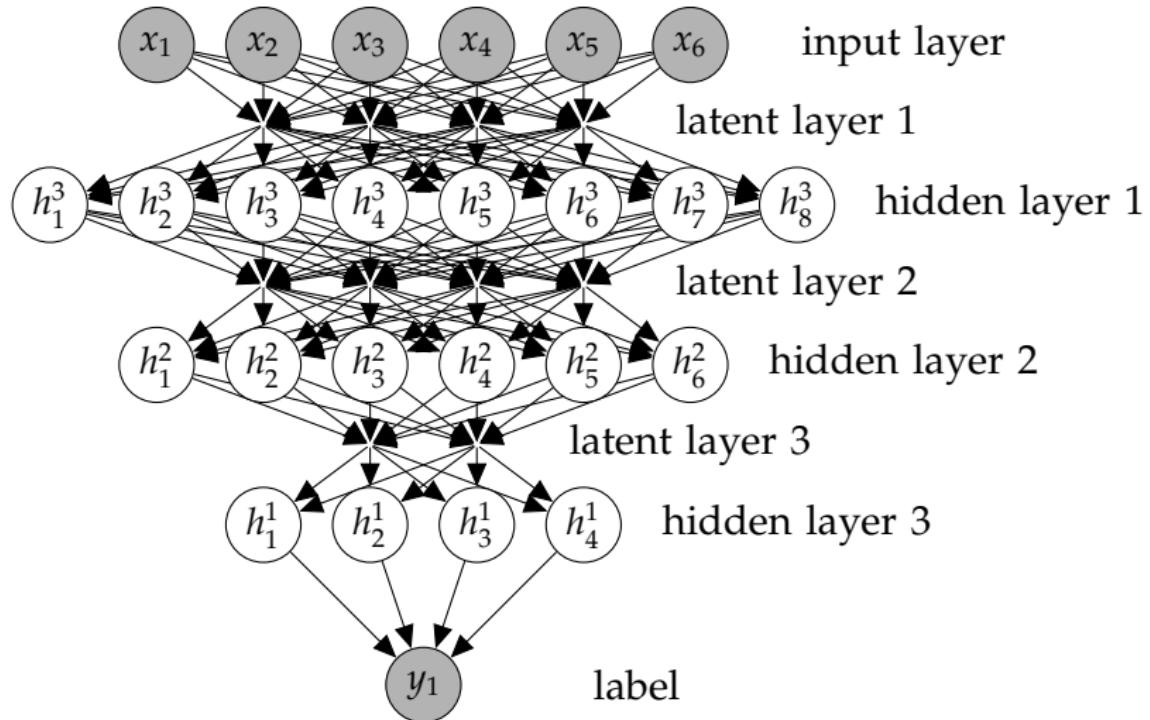


Potential applications..?

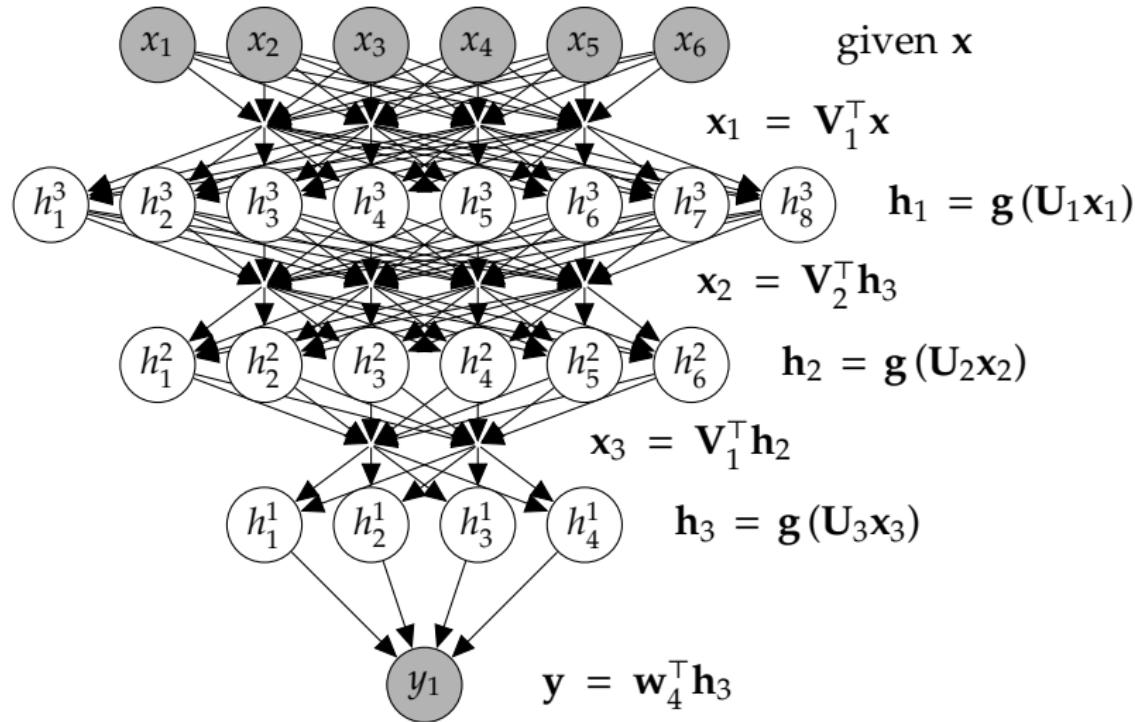


Manifold Relevance Determination

Deep Neural Network



Deep Neural Network



Outline

Probabilistic Linear Dimensionality Reduction

Non Linear Probabilistic Dimensionality Reduction

Conclusions

Conclusions

Summary

- ▶ We've advocated Dimensionality Reduction as a good way of modeling in high dimensions.
- ▶ Spectral techniques lead to convex algorithms.
- ▶ Probabilistic techniques map the “correct way” around.
 - ▶ This leads to problems with local minima.
- ▶ Have shown ability of probabilistic techniques to deal with high dimensional data.

Summary

- ▶ We've advocated Dimensionality Reduction as a good way of *probabilistic* modelling in high dimensions.
- ▶ Probabilistic techniques map the “correct way” around.
 - ▶ This leads to problems with local minima.
- ▶ Probabilistic dimensionality reduction is useful in practice.
- ▶ There are still many open problems to be overcome.

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 Kauffman. [[PDF](#)].
- A. Damianou, M. K. Titsias, and N. D. Lawrence. Variational Gaussian process dynamical systems. In P. Bartlett, F. Pereira, 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](#)].
- A. Elgammal and C. S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.
- Z. Ghahramani, editor. *Proceedings of the International Conference in Machine Learning*, volume 24, 2007. Omnipress. [[Google Books](#)].
- 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.
- A. Klami and S. Kaski. Local dependent components analysis. In Ghahramani (2007). [[Google Books](#)].
- A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72: 39–46, 2008.
- 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.

References II

- 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 Ghahramani (2007), pages 481–488. [[Google Books](#)] . [[PDF](#)].
- G. Leen and C. Fyfe. A Gaussian process latent variable model formulation of canonical correlation analysis. Bruges (Belgium), 26–28 April 2006 2006.
- S. Levine, J. M. Wang, A. Haraux, Z. Popović, and V. Koltun. Continuous character control with low-dimensional embeddings. *ACM Transactions on Graphics (SIGGRAPH 2012)*, 31(4), 2012.
- C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. Technical report,
- 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).
- J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12: 1247–1283, 2000.
- 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](#)].

References III

- M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. In Y. W. Teh and D. M. Titterington, 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](#)].
- L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958.
- R. Urtasun and T. Darrell. Discriminative Gaussian process latent variable model for classification. In Ghahramani (2007). [[Google Books](#)].
- 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.
- M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *European Conference on Computer Vision*, pages 447–460, 2002.
- S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the International Conference in Machine Learning*, volume 28, 2011.
- J. M. Wang, D. J. Fleet, and A. Hertzmann. Multifactor gaussian process models for style-content separation. In Ghahramani (2007), pages 975–982. [[Google Books](#)].
- 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.