Hierarchical Gaussian Process Latent Variable Models

Neil D. Lawrence and Andrew J. Moore
Machine Learning Group
School of Computer Science
University of Manchester, U.K.

22nd June 2007

Outline

Online Resources

All source code and slides are available online

- This talk available from my home page (see talks link on left hand side).
- Examples shown are in the 'oxford' toolbox (vrs 0.131).
  - `http://www.cs.man.ac.uk/~neill/oxford/`.
- And the 'hgplvm' toolbox (vrs 0.1).
  - `http://www.cs.man.ac.uk/~neill/hgplvm/`.
- MATLAB commands used for examples given in `typewriter font`.

Curse of Dimensionality

Incorporating assumptions about data structure

- How do we model high dimensional data probabilistically?
  1. Probabilistic models with sparse connectivity: tree structures, junction trees, Markov random fields.
     - Dictactes conditional independecies in the data.
  2. Assume data inherently lives on a low dimensional manifold.
     - Perhaps all data points are fully interdependent, but they live in a low dimension space.
- Can we combine these two approaches in one model?

Modelling in High Dimensions

Avoiding the Curse of Dimensionality

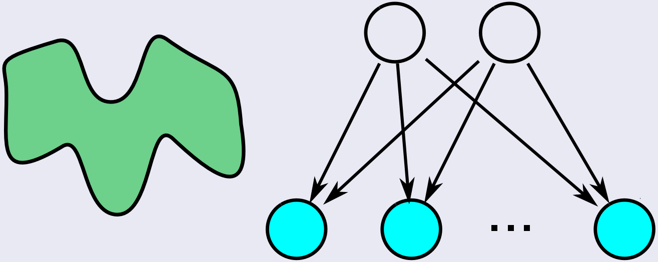


Figure: Probabilistic non-linear dimensional reduction.

Modelling in High Dimensions

Avoiding the Curse of Dimensionality

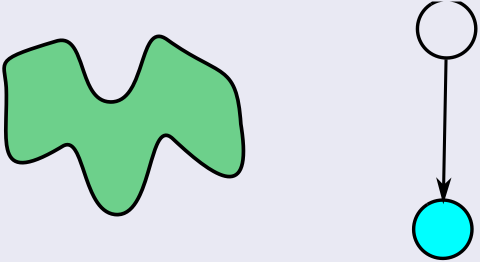


Figure: Probabilistic non-linear dimensional reduction.

Modelling in High Dimensions

Avoiding the Curse of Dimensionality

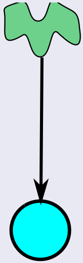


Figure: Probabilistic non-linear dimensional reduction.

Modelling in High Dimensions

Avoiding the Curse of Dimensionality

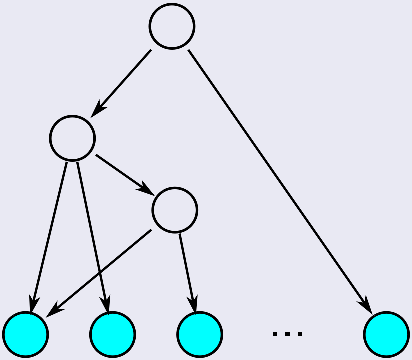


Figure: Hierarchical model (sparse connectivity).

Modelling in High Dimensions

Avoiding the Curse of Dimensionality

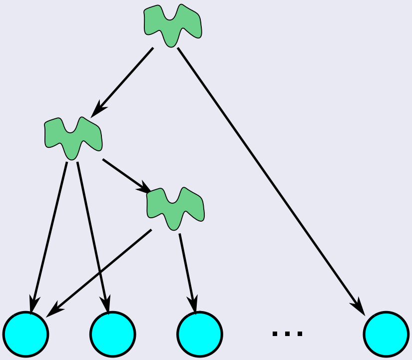


Figure: Hierarchy of non-linear dimensional reductions (**this talk**).

Notation

q— dimension of latent/embedded space

d— dimension of data space

n— number of data points

centred data, $\mathbf{Y} = [\mathbf{y}_{1,:}, \ldots, \mathbf{y}_{n,:}]^{\mathrm{T}} = [\mathbf{y}_{:,1}, \ldots, \mathbf{y}_{:,d}] \in \Re^{n \times d}$

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

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

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

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

Reading Notation

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

- Covariance given by $n^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$.
- Inner product matrix given by $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$.

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

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

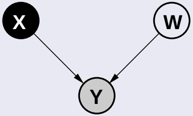
where

$$\eta_{i,:} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right).$$

Linear Latent Variable Model I

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*.

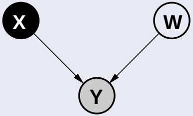


$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model I

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*.

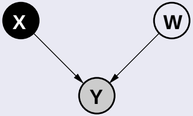


$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model I

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*.



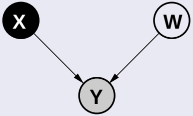
$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:}, \sigma^2\mathbf{I}\right)$$

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

Linear Latent Variable Model I

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*.



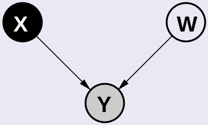
$$p\left(\mathbf{Y}|\mathbf{X},\mathbf{W}\right)=\prod_{i=1}^{n}N\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:},\sigma^2\mathbf{I}\right)$$

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

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

Linear Latent Variable Model II

Dual Probabilistic PCA Max. Likelihood Soln [Lawrence, 2004, 2005]



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

Linear Latent Variable Model II

Dual Probabilistic PCA Max. Likelihood Soln [Lawrence, 2004, 2005]

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

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

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

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

where $\mathbf{V}$ is an arbitrary rotation matrix.

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln [Tipping and Bishop, 1999]

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

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

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

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

where $\mathbf{V}$ is an arbitrary rotation matrix.

Equivalence of Formulations

The Eigenvalue Problems are equivalent

- Solution for Probabilistic PCA (solves for the mapping)

$$\mathbf{Y}^{\top}\mathbf{Y}\mathbf{U}_q = \mathbf{U}_q\Lambda_q \qquad \mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{V}^{\top}$$

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

$$\mathbf{Y}\mathbf{Y}^{\top}\mathbf{U}'_q = \mathbf{U}'_q\Lambda_q \qquad \mathbf{X} = \mathbf{U}'_q\mathbf{L}\mathbf{V}^{\top}$$

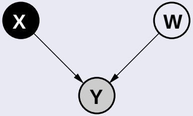
- Equivalence is from

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

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 *parameteters*, **W**.
    - Integrate out *parameters*.



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

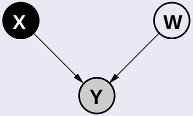
$$p\left(\mathbf{W}\right)=\prod_{i=1}^{d}N\left(\mathbf{w}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

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

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).

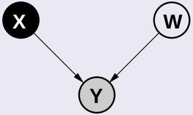


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}\right)$$

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).



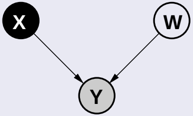
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right)$$

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

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’.
  - We call this the Gaussian Process Latent Variable model (GP-LVM).



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

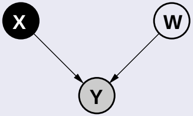
$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I}$$

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).



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

$$\mathbf{K} = ?$$

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

Stick Man

Generalization with less Data than Dimensions

- Powerful uncertainty handling of GPs leads to suprising 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

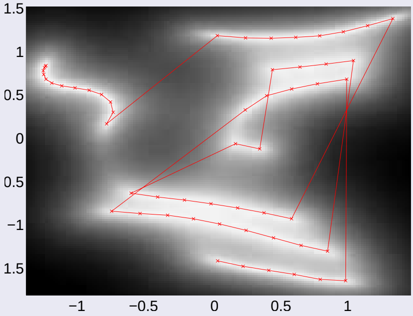


Figure: The latent space for the stick man motion capture data.

Adding Dynamics

MAP Solutions for Dynamics Models

- Introduce dynamical model in latent space.
  - Marginalising such dynamics is intractable.
  - But: **MAP solutions** are trivial to implement.
- Wang et al. [2006] suggest using a auto regressive Gaussian Process.
- Here we use a regressive Gaussian process.
  $p(\mathbf{Y}|\mathbf{t}) = \int p(\mathbf{Y}|\mathbf{X})\, p(\mathbf{X}|\mathbf{t})\, d\mathbf{X}$

Regressive Dynamics

Direct use of Time Variable

- Take $\mathbf{t}$ as an input, use a prior $p(\mathbf{X}|\mathbf{t})$.
- User a Gaussian process prior for $p(\mathbf{X}|\mathbf{t})$.
- Also allows us to consider variable sample rate data.

Motion Capture Results

`demStick1` and `demStick5`

Figure: The latent space for the motion capture data without dynamics (*left*) and with regressive dynamics (*right*) based on an RBF kernel.

Motion Capture Results

demStick1 and demStick5

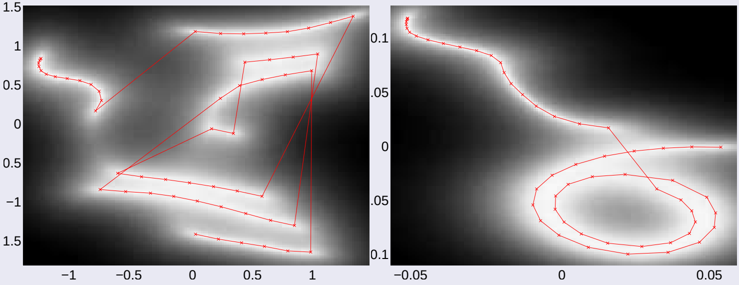


Figure: The latent space for the motion capture data without dynamics (*left*) and with regressive dynamics (*right*) based on an RBF kernel.

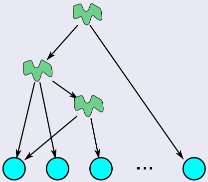
Hierarchical GP-LVM

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.

Hierarchical GP-LVM

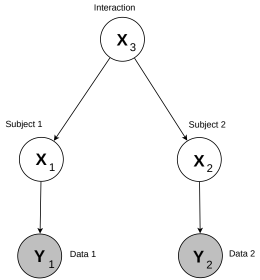
Stacking GP-LVMs



- This provides a route to incoporate conditional independencies.
- Ideally we should marginalise latent spaces
    - In practice we seek **MAP solutions**.

Two Correlated Subjects

- Simple hiearchy:
  - Motion capture data with two subjects.
- Subjects interact: approach each other and 'high five'.
- Model as a very simple tree.



$$p(\mathbf{Y}_1, \mathbf{Y}_2) = \int p(\mathbf{Y}_1|\mathbf{X}_1) \int p(\mathbf{Y}_2|\mathbf{X}_2) \int p(\mathbf{X}_1|\mathbf{X}_3)\, p(\mathbf{X}_2|\mathbf{X}_3)\, d\mathbf{X}_1 d\mathbf{X}_2 d\mathbf{X}_3$$

Two Correlated Subjects

demHighFive1

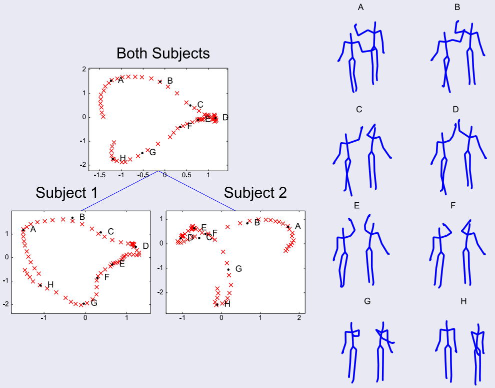


Figure: Hierarchical model of a 'high five'.

Within Subject Hierarchy

Decomposition of Body

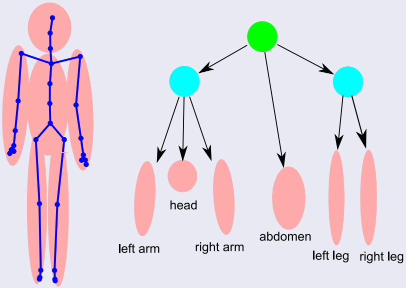


Figure: Decomposition of a subject.

Single Subject Run/Walk

demRunWalk1

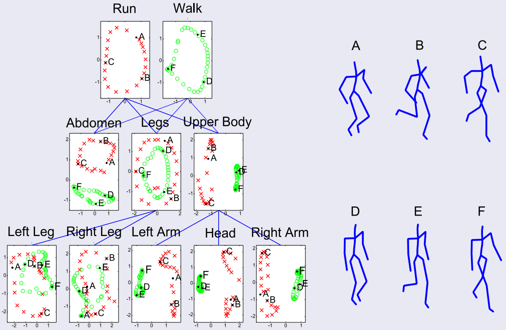


Figure: Hierarchical model of a walk and a run.

Overfitting

More parameters than data

- Large number of parameters: why doesn't it overfit?
- Standard GP-LVM: parameters increase linearly $\frac{q}{d} \times N$, $q < d$.
- HGP-LVM: adding more latent variables (parameters), will we overfit?
  - Upper levels only regularise the leaf nodes: if the leaf nodes don't overfit model won't.
  - Best likelihood obtained by *removing regularisation*.
  - Counter this potential problem in two ways.
    1. Provide a fixed dynamical prior at the top level.
    2. Constraine the noise variance of each non-leaf Gaussian process to 1×10^{-6}.

Summary

Conclusions

- GP-LVM is a Probabilistic Non-Linear Generalisation of PCA.
- We can stack GP-LVMs to provide:
  - A dynamical model.
  - A hierarchical decomposition of our data.
- MAP Solutions still provide interesting decompositions.

References

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, Nov 2005.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999.

J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.