Human Motion Modelling through Dimensional Reduction with Gaussian Processes

Neil D. Lawrence
Carl Henrik Ek
Philip H. S. Torr

29th January 2008

Outline

1. Probabilistic Dimensionality Reduction
2. Model Extensions
3. Conclusions

Outline

1. Probabilistic Dimensionality Reduction
2. Model Extensions
3. Conclusions

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 $\mathbf{A}$

X and 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 Latent Variable Model

- Represent data, $\mathbf{Y}$, with a lower dimensional set of latent variables $\mathbf{X}$.

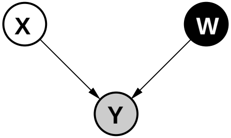
Assume a linear relationship of the form

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

where

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

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



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

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

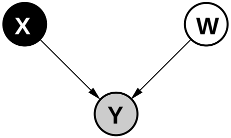
$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^{\top}\mathbf{Y}\right) + \text{const.}$$

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.

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



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

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

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

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

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

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

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

Gaussian Processes

Zero mean Gaussian Process

- A (zero mean) Gaussian process likelihood is of the form

$$p(\mathbf{y}|\mathbf{X}) = N(\mathbf{y}|\mathbf{0}, \mathbf{K}),$$

where **K** is the covariance function or *kernel*.

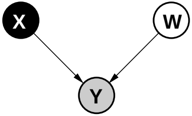
  - The *linear kernel* with noise has the form

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

  - Priors over non-linear functions are also possible.

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*.



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

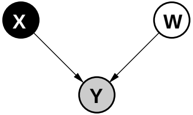
$$p(\mathbf{W}) = \prod_{i=1}^{d} N\left(\mathbf{w}_{i,:}|\mathbf{0},\mathbf{I}\right)$$

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

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).

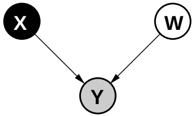


$$p\left(\mathbf{Y}|\mathbf{X}\right)=\prod_{j=1}^{d}N\left(\mathbf{y}_{:,j}|\mathbf{0},\mathbf{X}\mathbf{X}^{\top}+\sigma^{2}\mathbf{I}\right)$$

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).



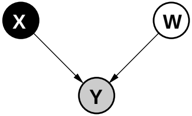
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right)$$

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

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



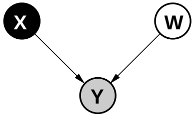
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{d} N\left(\mathbf{y}_{:,j}|\mathbf{0}, \mathbf{K}\right)$$

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

This is a product of Gaussian processes with linear kernels.

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.

RBF Kernel

- The RBF kernel has the form $k_{ij} = k\left(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}\right)$, where

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

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

Extensions

MAP Solutions for Dynamics Models

- Autoregressive Gaussian Processes. Wang et al. [2006]

Force the Model to Respect Local Distances

- Back constrained GP-LVM.

Developments Made Under Pump Priming Grant

- Sparse Approximations for Large Data Sets
- Hierarchical Models for Subject Decomposition
- Three Dimensional Pose Reconstruction from Images

Outline

1. Probabilistic Dimensionality Reduction
2. Model Extensions
3. Conclusions

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.
- Ideally we should marginalise latent spaces
  - In practice we seek MAP solutions.

Two Correlated Subjects

`demHighFive1`

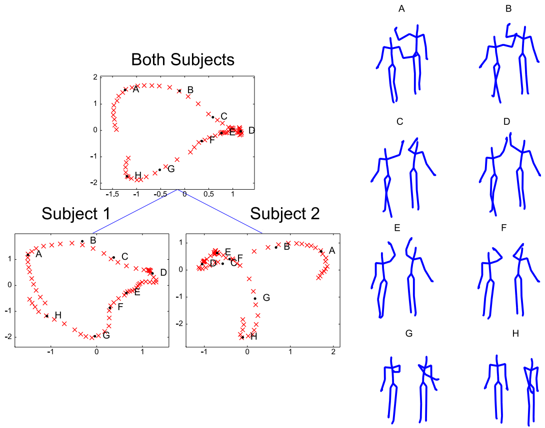


Figure: Hierarchical model of a 'high five'.

Decomposition of Body

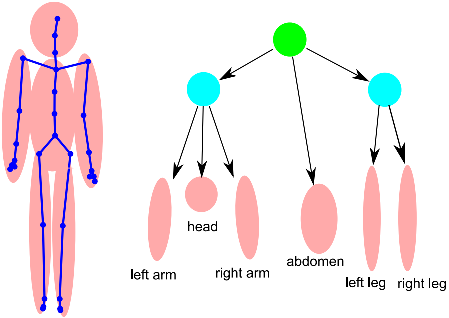


Figure: Decomposition of a subject.

demRunWalk1

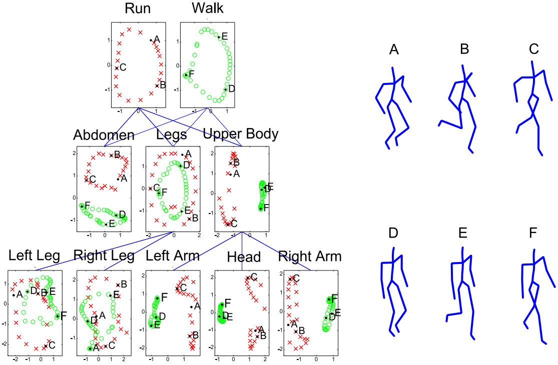


Figure: Hierarchical model of a walk and a run.

Complexity Issues

- Gaussian processes inherently
  - $O\left(N^3\right)$ complexity,
  - $O\left(N^2\right)$ storage.
- Sparse Gaussian processes normally give
  - $O\left(k^2N\right)$ complexity,
  - $O\left(kN\right)$ storage
- FITC Approximation [Snelson and Ghahramani, 2006, Quiñonero Candela and Rasmussen, 2005, Presented/Developed at PASCAL workshop!].

Sparse GP-LVM

- Recreate results of Taylor et al. [2007] on human motion capture data set.
- Data was walking and running motions from subject 35 in the CMU Mocap data base.
- Used dynamical refinement of the GP-LVM proposed by Wang et al. [2006]
- Taylor et al. [2007] applied their binary latent variable model to two missing data problems
    - right leg was removed from the test seqence
    - upper body was removed.
- Reconstruction obtained compared with nearest neighbour.

FITC Approximation

- Used the FITC approximation with 100 inducing points.
- The models were back constrained [Lawrence and Quiñonero Candela, 2006] .
- The data set size was 2613 frames.

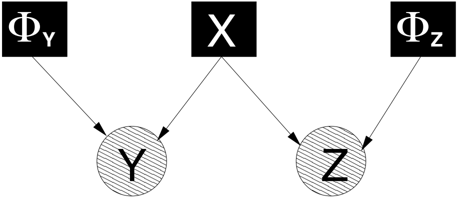
Results

Root mean squared angle error results on test data.

Data	Leg	Body
GP-LVM ($q = 3$)	3.40	**2.49**
GP-LVM ($q = 4$)	**3.38**	2.72
GP-LVM ($q = 5$)	4.25	2.78
NN (s)	4.44	2.62
NN	4.11	3.20

Table: NN: nearest neighbour, NN (s): nearest neighbour in scaled space, GP-LVM (latent dimension): the GP-LVM with different latent dimensions, q.

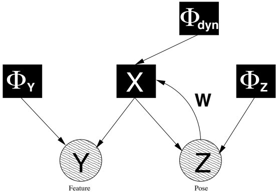
Shared GP-LVM



- Learn two separate kernels from a single shared latent representation **X** [Shon et al., 2006]
- Objective

$$p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \Phi_Y, \Phi_Z) = p(\mathbf{Y}|\mathbf{X}, \Phi_Y)p(\mathbf{Z}|\mathbf{X}, \Phi_Z)$$

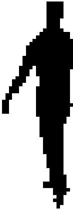
Shared GP-LVM Experiments[1]

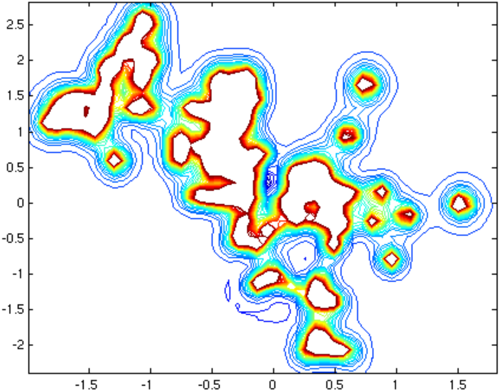


- Silhouette Features: $\mathbf{y}_i \in \Re^{100}$, Pose Parameters: $\mathbf{z}_i \in \Re^{54}$
- **Back constraints:** force bijective mapping between latent space and pose [Lawrence and Quiñonero Candela, 2006].
- **Dynamics:** add GP auto regressive dynamics to latent space [Wang et al., 2006].
- **Artificially generated training data:** from Agarwal and Triggs [2006]**.**

[1] Ek et al. [2007]

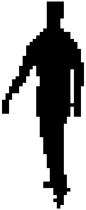
Shared GP-LVM Experiments

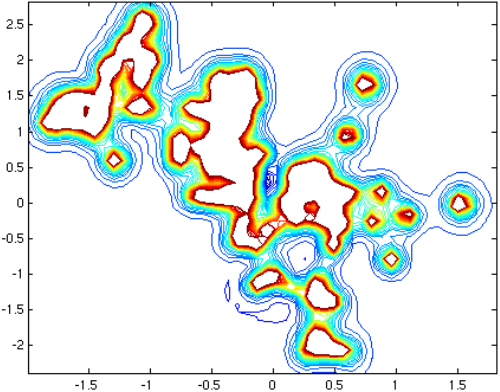




- Highly multimodal latent space given silhouette.

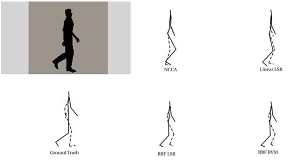
Shared GP-LVM Experiments



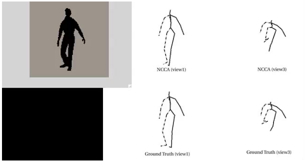


- Highly multimodal latent space given silhouette.

Video

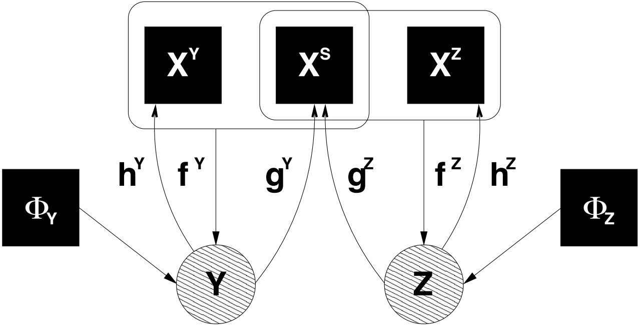


`runcca_all.sh`



`runcca_only.sh`

Modified Model



Initialisation

Shared Latent space by kernel CCA:

- Find directions $\{\mathbf{W_Y}, \mathbf{W_Z}\}$ in each feature space maximizing the correlation
- Canonical variate $\begin{cases} \mathbf{a_Y} & = & \mathbf{YW_Y} \\ \mathbf{a_Z} & = & \mathbf{ZW_Z} \end{cases}$
- Solution through Eigenvalue problem.

Non Shared Latent Space

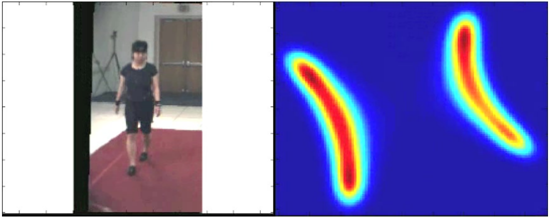
- Find further directions *orthogonal* to CCA directions of maximum variance.
- We named these non-consolidating components.
- Solution through eigenvalue problem.

Kernels

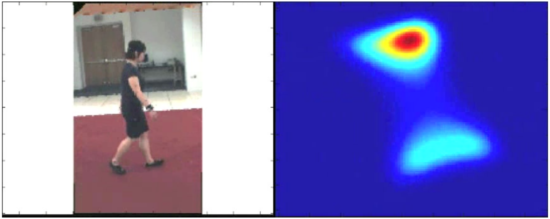
Feature Spaces:

- Many possible choices of feature space
  1. Linear Kernel
  2. RBF
  3. Maximum Variance Unfolding, Isomap
- Choose between them using GP-LVM likelihood [Harmeling, 2007].

Results of Initialisation



`runspectral_test.sh`



`runspectral_test.sh`

Outline

1 Probabilistic Dimensionality Reduction

2 Model Extensions

3 Conclusions

Summary

- GP-LVM is a Probabilistic Non-Linear Generalisation of PCA.
- Pump priming extensions:
  - Hierarchical representations.
  - Larger data sets.
  - Shared latent space models.
- Follow ups:
  - Carl to visit Trevor Darrell in Berkeley later in year.
  - Imminent EPSRC application building on the work.
  - Other on going work on GPs, differential equations and human motion.

References

A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1), 2006. doi: 10.1109/TPAMI.2006.21.

C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In *MLMI 2007*, 2007. To appear.

S. Harmeling. Exploring model selection techniques for nonlinear dimensionality reduction. Technical Report EDI-INF-RR-0960, University of Edinburgh, 2007.

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 and J. Quiñonero Candela. Local distance preservation in the GP-LVM through back constraints. In W. Cohen and A. Moore, editors, *Proceedings of the International Conference in Machine Learning*, volume 23, pages 513–520. Omnipress, 2006. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143909.

J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao. Learning shared latent structure for image synthesis and robotic imitation. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

G. W. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19, Cambridge, MA, 2007. MIT Press.

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.