Modeling dynamical and multi-modal computer vision data via non-linear probabilistic dimensionality reduction

Andreas Damianou[1]
joint work with Carl Henrik Ek[2], Michalis Titsias[3] and Neil Lawrence[1]

[1] Department of Neuro- and Computer Science, University of Sheffield, UK
[2] Computer Vision and Active Perception Lab , KTH
[3] Wellcome Trust Centre for Human Genetics, University of Oxford

University of Surrey, 14/06/2012

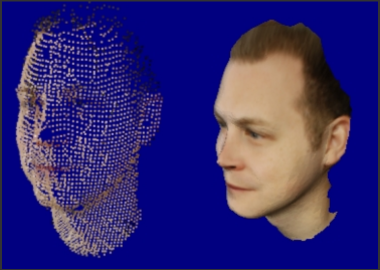
Outline

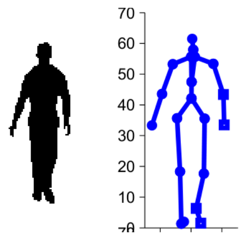
Dimensionality reduction techniques
- From Dual PPCA to GP-LVM

Bayesian GP-LVM

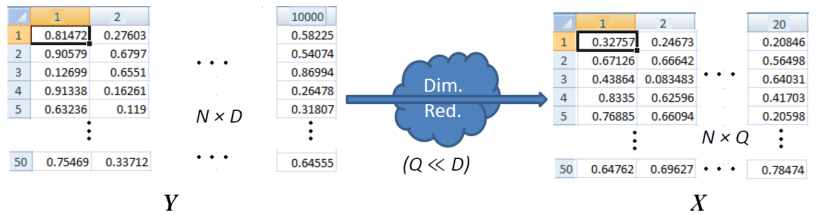
Structure in the latent space
- Modelling dynamics
- Multi-modal modelling

Real-world datasets in computer vision are usually high-dimensional, complex and noisy





Dimensionality reduction



Dimensionality reduction techniques 1/2

Probabilistic vs non-probabilistic

A probabilistic interpretation allows us to:

- Have a model of the data
- Handle incomplete data
- Generate/sample novel data
- Extend the model with prior information or integrate it with other models (e.g. mixtures)

Probabilistic, generative methods

- Observed (high-dimensional) data: $Y \in \mathbb{R}^{N \times D}$
  These contain redundant information
- Actual (low-dimensional) data: $X \in \mathbb{R}^{N \times Q}$, $Q \ll D$
  These are <u>unobserved</u> and (ideally) contain only the minimum amount of information needed to correctly describe the phenomenon
- Work "backwards": learn $f : X \mapsto Y$

Probabilistic, generative methods

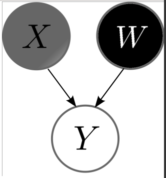
- Model:

$$y_{nd} = f_d(\mathbf{x}_n, W) + \epsilon_n \ , \quad \epsilon_n \sim \mathcal{N}(0, \beta^{-1})$$

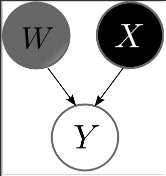
- $p(Y|W, X, \beta) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n | W\mathbf{x}_n, \beta^{-1}\mathbf{I})$ (*linear case*)
- $W, X \in \mathbb{R}^{N \times Q}$, $Q \ll D$
- X is unobserved (latent space)

From dual PPCA to GP-LVM

- PPCA places a prior on and marginalises the latent space X and optimises the *linear* mapping's parameters W
- Dual PPCA does the opposite: the prior is placed on the mapping parameters.



ppca



Dual PPCA

$$p(Y|W,\beta) = \int p(Y|X,W,\beta)p(X)\mathrm{d}X$$

$$p(Y|X,\beta) = \int p(Y|X,W,\beta)p(W)\mathrm{d}W$$

Gaussian process latent variable model (GP-LVM)

- PPCA and Dual PPCA are equivalent (equivalent eigenvalue problems for ML solution)

Gaussian process latent variable model (GP-LVM)

- PPCA and Dual PPCA are equivalent (equivalent eigenvalue problems for ML solution)

- GP-LVM: Instead of placing a prior $p(W)$ on the parametric mapping's parameters, we can place a prior directly on the mapping function $\Rightarrow$ GP prior

Gaussian process latent variable model (GP-LVM)

- PPCA and Dual PPCA are equivalent (equivalent eigenvalue problems for ML solution)

- GP-LVM: Instead of placing a prior $p(W)$ on the parametric mapping's parameters, we can place a prior directly on the mapping function $\Rightarrow$ GP prior

- A GP prior $f \sim \mathcal{GP}(\mathbf{0}, k(x, x'))$ allows for *non-linear mappings* if the kernel k is non-linear. For example:

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

[Lawrence 2005]

Dimensionality reduction: Linear vs non-linear

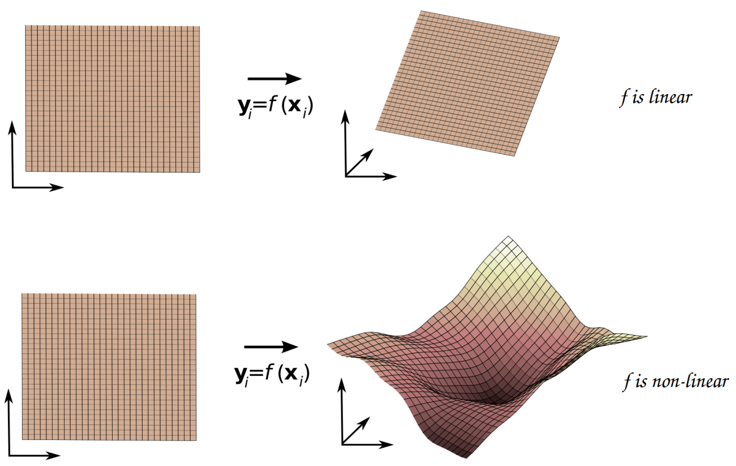
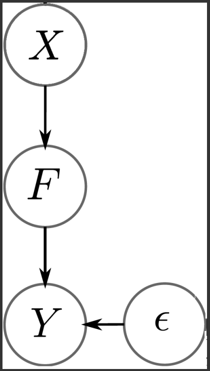


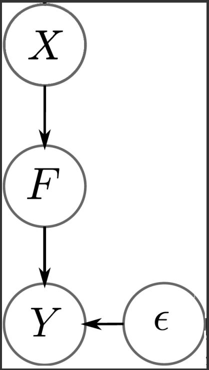
Image from: "Dimensionality Reduction the Probabilistic Way", N. Lawrence, ICML tutorial 2008

Optimising the GP-LVM



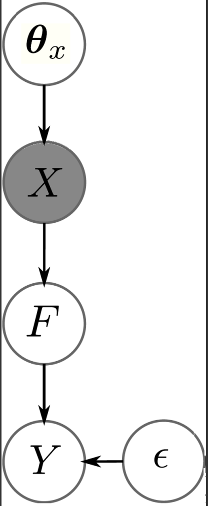
- Objective function for optimisation is $p(Y|X)$ (found analytically, as F is finite)

Optimising the GP-LVM



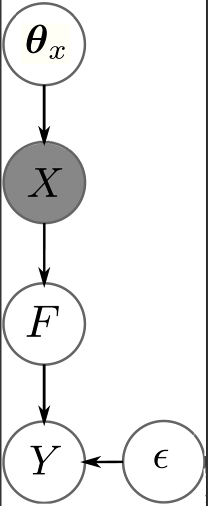
- Objective function for optimisation is $p(Y|X)$ (found analytically, as F is finite)
- Problem: this finds a single point (MAP) estimate for X

Optimising the GP-LVM



- Objective function for optimisation is $p(Y|X)$ (found analytically, as F is finite)
- Problem: this finds a single point (MAP) estimate for X
- We would prefer to instead find a *distribution* over $X \Rightarrow$ *Bayesian GP-LVM*

Optimising the GP-LVM



- Objective function for optimisation is $p(Y|X)$ (found analytically, as F is finite)
- Problem: this finds a single point (MAP) estimate for X
- We would prefer to instead find a *distribution* over $X \Rightarrow$ *Bayesian GP-LVM*
- This allows for:
  - training robust to overfitting
  - automatic detection for the dimensionality of X
  - incorporating known structure on the latent space

Bayesian GPLVM

- GPLVM objective function:

$$p(Y|X) = \int p\left(Y|\mathbf{f}\right) p(\mathbf{f}|X) d\mathbf{f} = \mathcal{N}(Y|\mathbf{0}, K_{NN} + \beta^{-1} I_N)$$

  The GPLVM is trained by maximizing $p(Y|X)$ w.r.t the mapping's parameters and X (jointly) $\Rightarrow$ *MAP* estimate,

- Bayesian GPLVM: Also integrate out X's:

$$p(Y) = \int p\left(Y|X\right) p(X) dX$$

$$p(X) = \prod_{n=1}^{N} N(\mathbf{x}_n|\mathbf{0}, I_Q)$$

Bayesian GPLVM

- GPLVM objective function:

$$p(Y|X) = \int p\left(Y|\mathbf{f}\right) p(\mathbf{f}|X) d\mathbf{f} = \mathcal{N}(Y|\mathbf{0}, K_{NN} + \beta^{-1} I_N)$$

  The GPLVM is trained by maximizing $p(Y|X)$ w.r.t the mapping's parameters and X (jointly) $\Rightarrow$ *MAP* estimate,

- Bayesian GPLVM: Also integrate out X's:

$$p(Y) = \int p\left(Y|X\right) p(X) dX$$

$$p(X) = \prod_{n=1}^{N} N(\mathbf{x}_n|\mathbf{0}, I_Q)$$

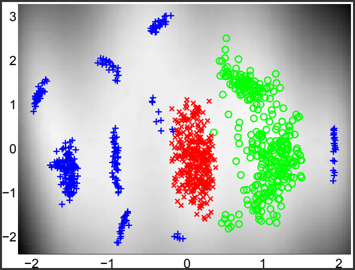
- Tractability: The marginal likelihood as well as the posterior $p(X|Y)$ are intractable $\Rightarrow$ the variational framework of [Titsias and Lawrence, 2010] resolves this

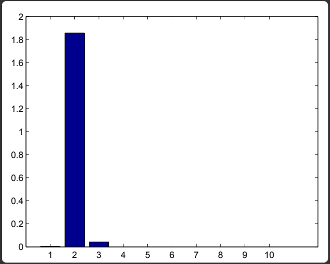
Automatic dimensionality detection

- Achieved by employing *automatic relevance determination (ARD)* priors for the mapping f.
- $f \sim \mathcal{GP}(\mathbf{0}, k_f)$ with:

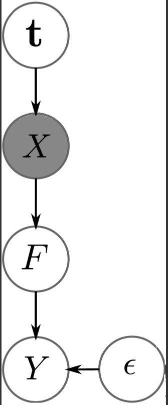
$$k_f\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma^2 e^{-\frac{1}{2}\sum_{q=1}^{Q} w_q (x_{i,q} - x_{j,q})^2}$$

- Example:





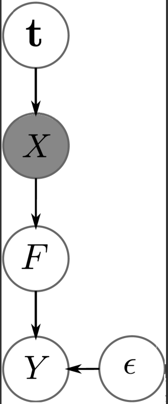
Modelling dynamics



- If Y form is a multivariate time-series, then X also has to be one

[Damianou et al., 2011]

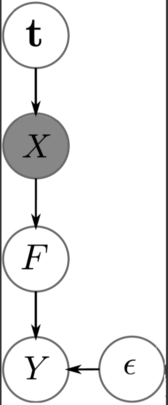
Modelling dynamics



- If Y form is a multivariate time-series, then X also has to be one
- Place a temporal GP prior on the latent space: $\mathbf{x} = x(t) = \mathcal{GP}(\mathbf{0}, k_x)$

[Damianou et al., 2011]

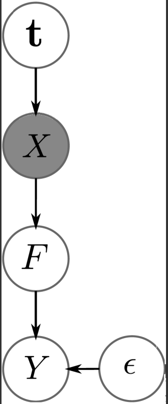
Modelling dynamics



- If Y form is a multivariate time-series, then X also has to be one
- Place a temporal GP prior on the latent space: $\mathbf{x} = x(t) = \mathcal{GP}(\mathbf{0}, k_x)$
- Dynamics are encoded in the covariance matrix $K_x = k_x(\mathbf{t}, \mathbf{t})$, e.g. forcing K_x to be block-diagonal allows to jointly model individual sequences

[Damianou et al., 2011]

Modelling dynamics

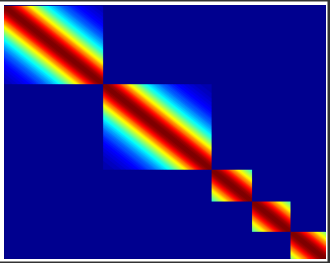


- If Y form is a multivariate time-series, then X also has to be one
- Place a temporal GP prior on the latent space: $\mathbf{x} = x(t) = \mathcal{GP}(\mathbf{0}, k_x)$
- Dynamics are encoded in the covariance matrix $K_x = k_x(\mathbf{t}, \mathbf{t})$, e.g. forcing K_x to be block-diagonal allows to jointly model individual sequences
- *Video examples...*

[Damianou et al., 2011]

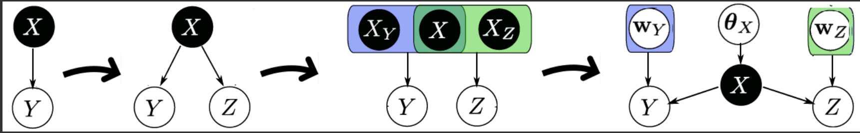
Modelling sequences

- Dynamics are encoded in the covariance matrix $K_x = k_x(\mathbf{t}, \mathbf{t})$, e.g. forcing K_x to be block-diagonal allows to jointly model individual sequences



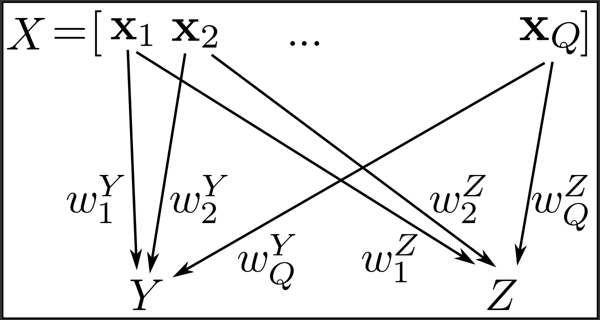
Multi-modal modelling

- Several observation modalities for the same underlying phenomenon
- Challenge: factorise the latent space into parts that are either private or shared for all modalities
- Bayesian solution: use a separate set of *ARD* parameters for each modality
- The ARD weights are optimised to learn the responsibility of each latent dimension for generating each of the observation spaces



Manifold Relevance Determination

- The high-level description of the model:



- Bayesian optimisation ensures that irrelevant dimensions will be assigned a zero weight

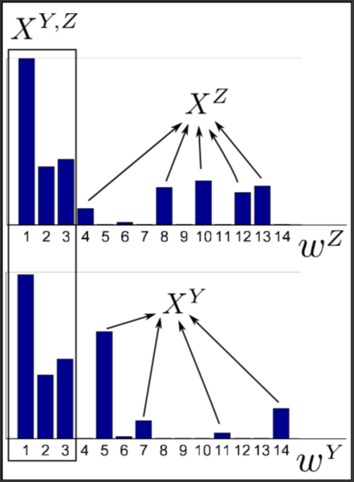
[Damianou et al., 2012]

Example: Yale faces

- Dataset Y: 3 persons under all illumination conditions
- Dataset Z: As above for 3 different persons
- Align datapoints $\mathbf{y}_n$ 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...

Summary

- GP-LVM: probabilistic non-linear dimensionality reduction
- Bayesian GP-LVM: placing a prior over and marginalising the latent space
- Dynamical framework: constraining the latent space to be a timeseries
- Multi-modal framework: automatically segment the latent space to shared and private subspaces

Thanks

KTH

Carl Henrik Ek

Univ. of Oxford

Michalis Titsias

Univ. of Sheffield

Neil Lawrence

Funding

- University of Sheffield Moody endowment fund
- Greek State Scholarships Foundation (IKY)