Deep Gaussian processes

Andreas Damianou

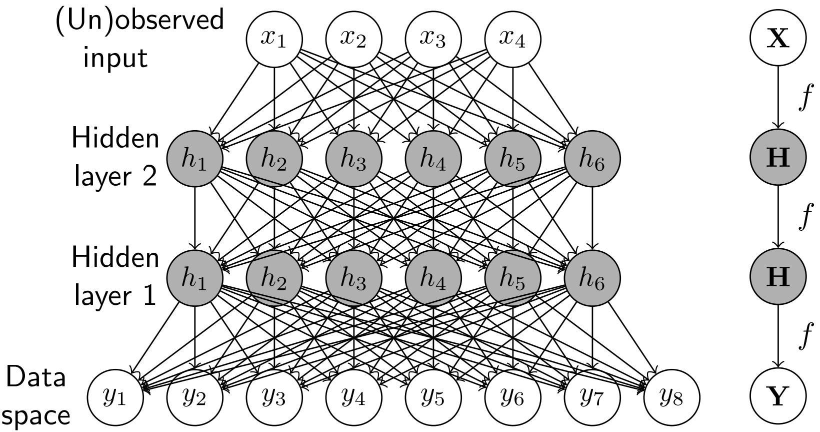
Department of Neuro- and Computer Science, University of Sheffield, UK

Deep Probabilistic Models Workshop, Sheffield, 02/10/2014

Outline

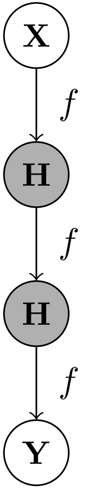
Outline

Deep learning (directed graph)



$$\mathbf{Y} = f(f(\cdots f(\mathbf{X}))), \qquad \mathbf{H}_i = f_i(\mathbf{H}_{i-1})$$

Deep Gaussian processes - Big Picture



Deep GP:

- Directed graphical model
- Non-parametric, non-linear mappings f
- Mappings f marginalised out analytically
- Likelihood is a non-linear function of the inputs
- Continuous variables
- NOT a GP!

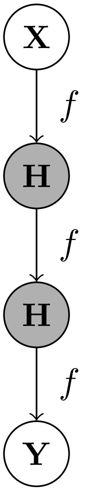
Challenges:

- Marginalise out $\mathbf{H}$
- No sampling: analytic approximation of objective

Solution:

- Variational approximation
- This also gives access to the *model evidence*

Deep Gaussian processes - Big Picture



Deep GP:

- Directed graphical model
- Non-parametric, non-linear mappings f
- Mappings f marginalised out analytically
- Likelihood is a non-linear function of the inputs
- Continuous variables
- NOT a GP!

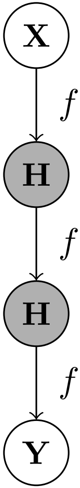
Challenges:

- Marginalise out $\mathbf{H}$
- No sampling: analytic approximation of objective

Solution:

- Variational approximation
- This also gives access to the *model evidence*

Deep Gaussian processes - Big Picture



Deep GP:

- Directed graphical model
- Non-parametric, non-linear mappings f
- Mappings f marginalised out analytically
- Likelihood is a non-linear function of the inputs
- Continuous variables
- NOT a GP!

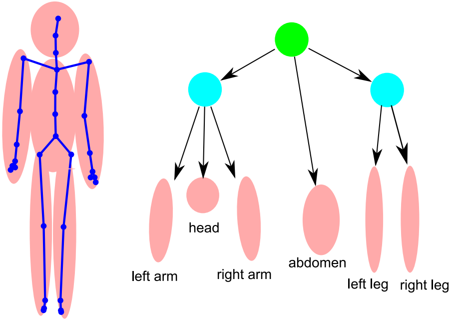
Challenges:

- Marginalise out $\mathbf{H}$
- No sampling: analytic approximation of objective

Solution:

- Variational approximation
- This also gives access to the *model evidence*

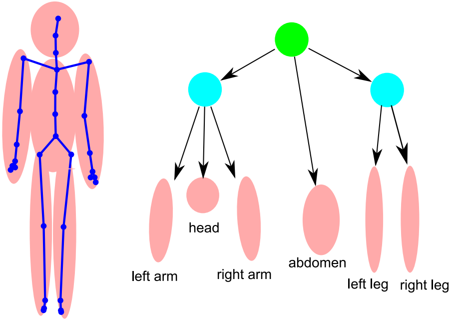
Hierarchical GP-LVM



- Hidden layers are not marginalised out.
- This leads to some difficulties.

[Lawrence and Moore, 2004]

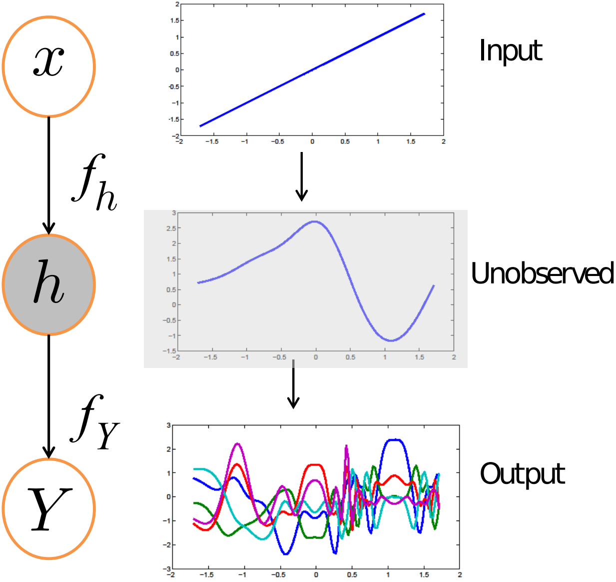
Hierarchical GP-LVM



- Hidden layers are not marginalised out.
- This leads to some difficulties.

[Lawrence and Moore, 2004]

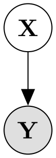
Sampling from a deep GP



Outline

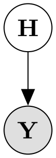
Dynamics

GP-LVM



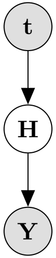
Dynamics

GP-LVM



Dynamics

GP-LVM



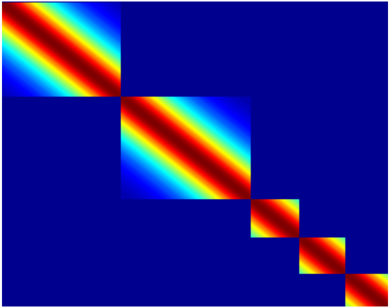
- If $\mathbf{Y}$ form is a multivariate time-series, then $\mathbf{H}$ also has to be one
- Place a temporal GP prior on the latent space:

$$\mathbf{h} = h(t) = \mathcal{GP}(\mathbf{0}, k_h(t,t))$$
$$\mathbf{f} = f(h) = \mathcal{GP}(\mathbf{0}, k_f(h,h))$$
$$\mathbf{y} = f(h) + \epsilon$$

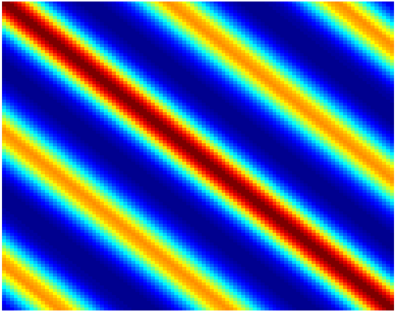
- Still, we didn't introduce uncertainty for the inputs to the second GP.

Dynamics

- Dynamics are encoded in the covariance matrix $\mathbf{K} = k(\mathbf{t}, \mathbf{t})$.
- We can consider special forms for $\mathbf{K}$.



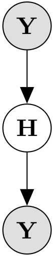
Model individual sequences



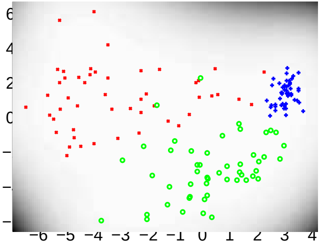
Model periodic data

- https://www.youtube.com/watch?v=i9TEoYxaBxQ (missa)
- https://www.youtube.com/watch?v=mUY1XHPnoCU (dog)
- https://www.youtube.com/watch?v=fHDWloJtgk8 (mocap)

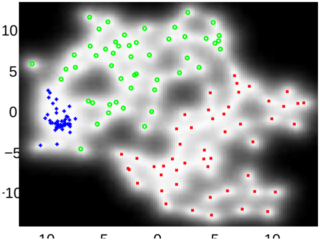
Autoencoder



GP-LVM:



Non-parametric auto-encoder:



Outline

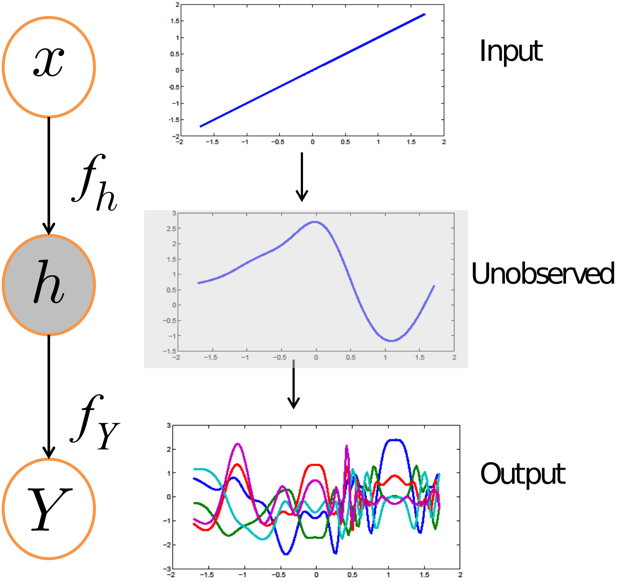
Part 1: A general view

Part 2: Structure in the latent space
- Dynamics
- Autoencoders

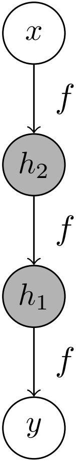
Part 3: Deep Gaussian processes
- Bayesian regularization
- Inducing Points
- Structure: ARD and MRD (multi-view)
- Examples

Summary

Sampling from a deep GP



MAP optimisation?



- Joint $= p(y|h_1)p(h_1|h_2)p(h_2|x)$

- MAP optimization is extremely problematic because:
  - Dimensionality of hs has to be decided a priori
  - Prone to overfitting, if h are treated as parameters
  - Deep structures are not supported by the model's objective but have to be forced [Lawrence & Moore '07]

Regularization solution: approximate Bayesian framework

- Analytic variational bound $\mathcal{F} \leq p(y|x)$
  - Extend the Variational Free Energy sparse GPs (Titsias 09) / Variational Compression tricks.
  - *Approximately* marginalise out h
- Automatic structure discovery (nodes, connections, layers)
  - Use the Automatic / Manifold Relevance Determination trick
- ...

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$
- $p(h_1|x) \quad = \int_{h_2, f_2} \; p(h_1|f_2) p(f_2|h_2) \; p(h_2|x)$
- $p(h_1|x, \tilde{h}_2) = \int_{h_2, f_2, u_2} p(h_1|f_2) p(f_2|u_2, h_2) p(u_2|\tilde{h}_2) p(h_2|x)$
- $\log p(h_1|x, \tilde{h}_2) \geq \int_{h_2, f_2, u_2} \mathcal{Q} \log \frac{p(h_1|f_2) \cancel{p(f_2|u_2, h_2)} p(u_2|\tilde{h}_2) p(h_2|x)}{\mathcal{Q} = \cancel{p(f_2|u_2, h_2)} q(u_2) q(h_2)}$
- $\log p(h_1|x, \tilde{h}_2) \geq \int_{h_2, f_2, u_2} \mathcal{Q} \log \frac{p(h_1|f_2) p(u_2|\tilde{h}_2) p(h_2|x)}{\mathcal{Q} = q(u_2) q(h_2)}$

$p(u_2|\tilde{h}_2)$ contains $k(\tilde{h}_2, \tilde{h}_2)^{-1}$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$
- $p(h_1|x) \quad = \int_{h_2, f_2} \quad p(h_1|f_2) p(f_2|h_2) \quad p(h_2|x)$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$
- $p(h_1|x) \quad = \int_{h_2, f_2} \quad p(h_1|f_2) p(f_2|h_2) \quad p(h_2|x)$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$

- $p(h_1|x) \quad = \int_{h_2, f_2} p(h_1|f_2) \underbrace{p(f_2|h_2)}_{\substack{\text{contains} \\ (k(h_2,h_2))^{-1}}} p(h_2|x)$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \Big(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \Big)$
- $p(h_1|x) \quad = \int_{h_2, f_2} \quad p(h_1|f_2) p(f_2|h_2) \quad p(h_2|x)$
- $p(h_1|x, \tilde{h}_2) = \int_{h_2, f_2, u_2} p(h_1|f_2) p(f_2|u_2, h_2) p(u_2|\tilde{h}_2) p(h_2|x)$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$
- $p(h_1|x) \quad = \int_{h_2,f_2} \quad p(h_1|f_2) p(f_2|h_2) \quad p(h_2|x)$
- $p(h_1|x, \tilde{h}_2) = \int_{h_2,f_2,u_2} p(h_1|f_2) p(f_2|u_2, h_2) p(u_2|\tilde{h}_2) p(h_2|x)$
- $\log p(h_1|x, \tilde{h}_2) \geq \int_{h_2,f_2,u_2} \mathcal{Q} \log \frac{p(h_1|f_2) p(f_2|u_2,h_2) p(u_2|\tilde{h}_2) p(h_2|x)}{\mathcal{Q}}$

Direct marginalisation of h is intractable (O_o)

- New objective: $p(y|x) = \int_{h_1} \left(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \right)$
- $p(h_1|x) = \int_{h_2,f_2} p(h_1|f_2) p(f_2|h_2) p(h_2|x)$
- $p(h_1|x,\tilde{h}_2) = \int_{h_2,f_2,u_2} p(h_1|f_2) p(f_2|u_2,h_2) p(u_2|\tilde{h}_2) p(h_2|x)$
- $\log p(h_1|x,\tilde{h}_2) \geq \int_{h_2,f_2,u_2} \mathcal{Q} \log \frac{p(h_1|f_2)\cancel{p(f_2|u_2,h_2)}p(u_2|\tilde{h}_2)p(h_2|x)}{\mathcal{Q}=\cancel{p(f_2|u_2,h_2)}q(u_2)q(h_2)}$

Direct marginalisation of h is intractable (O_o)

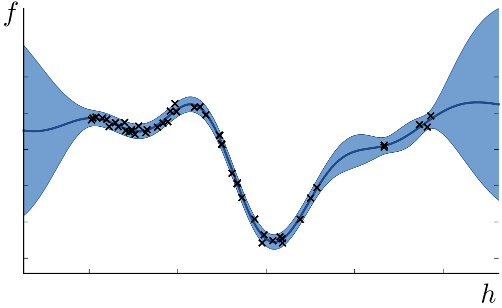
- New objective: $p(y|x) = \int_{h_1} \Big(p(y|h_1) \int_{h_2} p(h_1|h_2) p(h_2|x) \Big)$
- $p(h_1|x) \quad = \int_{h_2, f_2} p(h_1|f_2) p(f_2|h_2) \; p(h_2|x)$
- $p(h_1|x, \tilde{h}_2) = \int_{h_2, f_2, u_2} p(h_1|f_2) p(f_2|u_2, h_2) p(u_2|\tilde{h}_2) p(h_2|x)$
- $\log p(h_1|x, \tilde{h}_2) \geq \int_{h_2, f_2, u_2} \mathcal{Q} \log \frac{p(h_1|f_2)\cancel{p(f_2|u_2,h_2)}p(u_2|\tilde{h}_2)p(h_2|x)}{\mathcal{Q}=\cancel{p(f_2|u_2,h_2)}q(u_2)q(h_2)}$
- $\log p(h_1|x, \tilde{h}_2) \geq \int_{h_2, f_2, u_2} \mathcal{Q} \log \frac{p(h_1|f_2)p(u_2|\tilde{h}_2)p(h_2|x)}{\mathcal{Q}=q(u_2)q(h_2)}$

$p(u_2|\tilde{h}_2)$ contains $k(\tilde{h}_2, \tilde{h}_2)^{-1}$

The above trick is applied to all layers simultaneously.

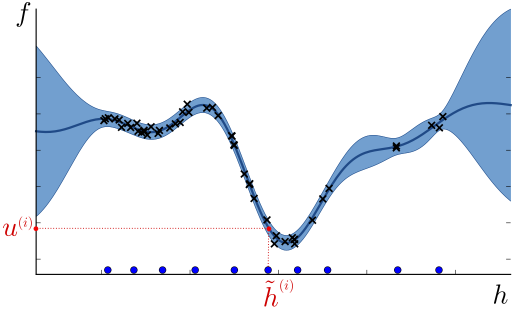
Inducing points: sparseness, tractability and Big Data

h_1	$\mathbf{f}_1$
h_2	$\mathbf{f}_2$
$\cdots$	$\cdots$
h_{30}	$\mathbf{f}_{30}$
h_{31}	$\mathbf{f}_{31}$
$\cdots$	$\cdots$
h_N	$\mathbf{f}_N$



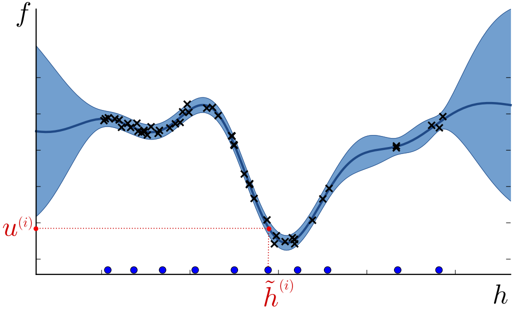
Inducing points: sparseness, tractability and Big Data

h_1	$\mathbf{f}_1$
h_2	$\mathbf{f}_2$
$\ldots$	$\ldots$
h_{30}	$\mathbf{f}_{30}$
$\tilde{h}^{(i)}$	$u^{(i)}$
h_{31}	$\mathbf{f}_{31}$
$\ldots$	$\ldots$
h_N	$\mathbf{f}_N$



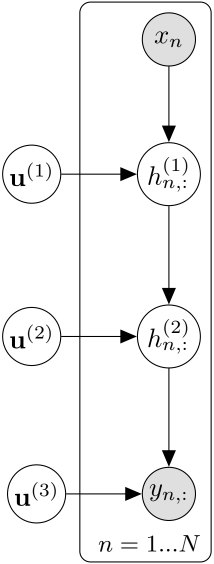
Inducing points: sparseness, tractability and Big Data

h_1	$\mathbf{f}_1$
h_2	$\mathbf{f}_2$
$\dots$	$\dots$
h_{30}	$\mathbf{f}_{30}$
$\tilde{h}^{(i)}$	$u^{(i)}$
h_{31}	$\mathbf{f}_{31}$
$\dots$	$\dots$
h_N	$\mathbf{f}_N$



- Inducing points originally introduced for faster **(sparse) GPs**
- But this also induces **tractability** in our models, due to the conditional independencies assumed
- Viewing them as **global variables**
  $\Rightarrow$ extension to **Big Data** [Hensman et al., UAI 2013]

Factorised vs non-factorised bound



- Preliminary bound:

  $$\mathcal{L} \leq \log p(\mathbf{Y}, \{\mathbf{H}_l\}_{l=1}^{L} | \{\mathbf{U}_l\}_{l=1}^{L+1}, \mathbf{X})$$

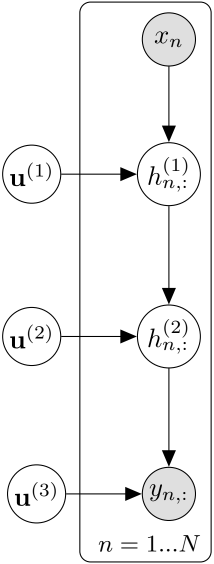
Factorised vs non-factorised bound

- Preliminary bound

$$
\begin{aligned}
\mathcal{L} &\leq \log p(\mathbf{Y}, \{\mathbf{H}_l\}_{l=1}^{L} | \{\mathbf{U}_l\}_{l=1}^{L+1}, \mathbf{X}) \\
\mathcal{L} &= \sum_{n=1}^{N} \left[\sum_{l=1}^{L} \left(\sum_{q=1}^{Q_l} \log \mathcal{N}\left(h_l^{(n,q)} | \mathbf{k}_l^{(n,:)} \mathbf{K}^{-1} \mathbf{u}_l^{(:,d)}, \beta_l^{-1} \mathbf{I} \right) \right.\right. \\
&\qquad \left.\left. - \frac{\beta_l^{-1} \tilde{\mathbf{k}}_l^{(n)}}{2} \right) \right] \\
&= \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{q=1}^{Q_l} \mathcal{L}_l^{n,q}
\end{aligned}
$$

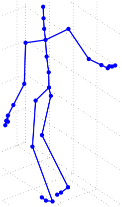
- Fully factorised.

SVI for factorised deep GPs

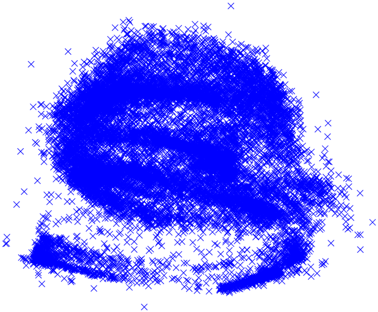


- We can additionally marginalise out $\mathbf{h}$ and maintain factorisation.
- We can consider SVI.
- Unlike $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}$, $\mathbf{h}$ are *not* global variables.
- So, estimate $\mathbf{h}^{(batch)}$ given the current $\boldsymbol{\theta}_t$
- Adjusting the step-length for SVI is tricky.

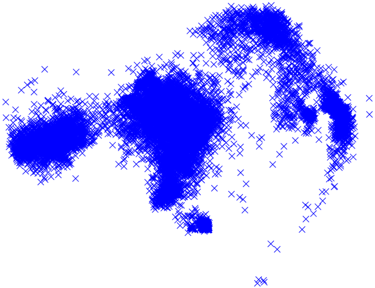
SVI - 18K mocap examples



Hidden space projections:

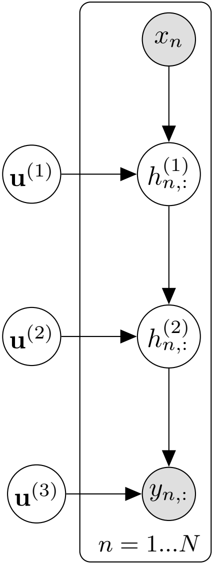


Global motion features

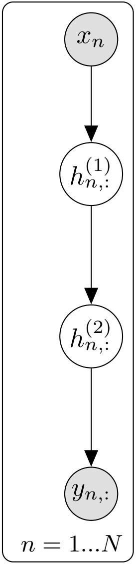


Clustered motion features

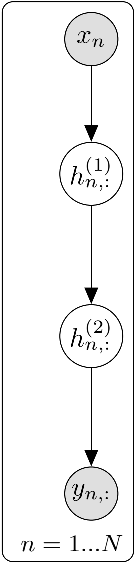
Integrate out inducing outputs



Integrate out inducing outputs



Integrate out inducing outputs



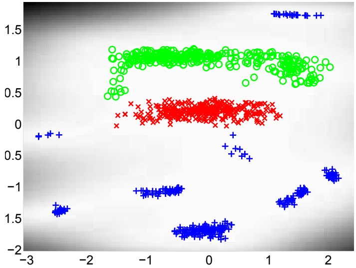
- Integrating $\mathbf{u}$ introduces coupling.
- But we can still distribute the computations efficiently (work by Z. Dai).
- An alternative approach is to collapse the effect of $q(\mathbf{h})$ (next talk by J. Hensman).

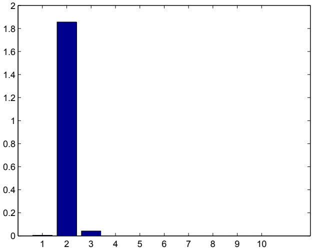
Automatic dimensionality detection

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

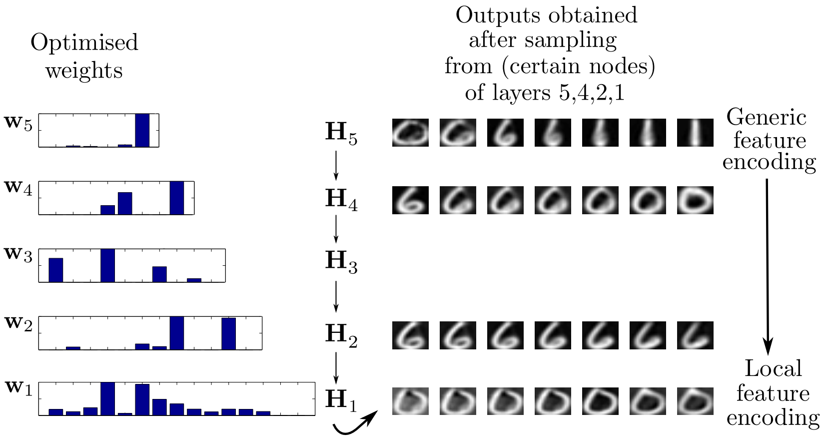
$$k_f\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{q=1}^{Q} w_q \left(x_{i,q} - x_{j,q}\right)^2\right)$$

- Example:





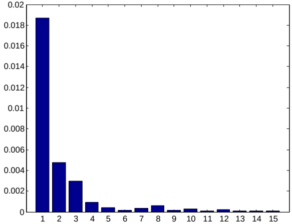
Deep GP: digits example



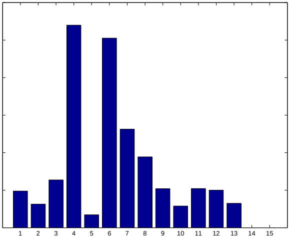
▸ http://staffwww.dcs.sheffield.ac.uk/people/A.Damianou/research/index.html#DeepGPs

MNIST: The first layer

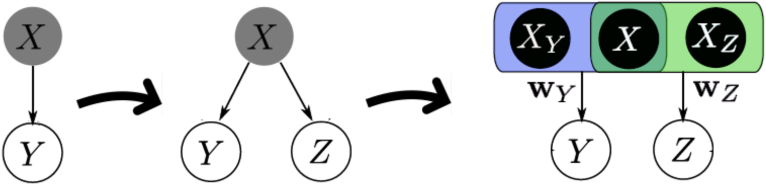
1 layer GP-LVM:



5 layer deep GP (showing 1st layer):

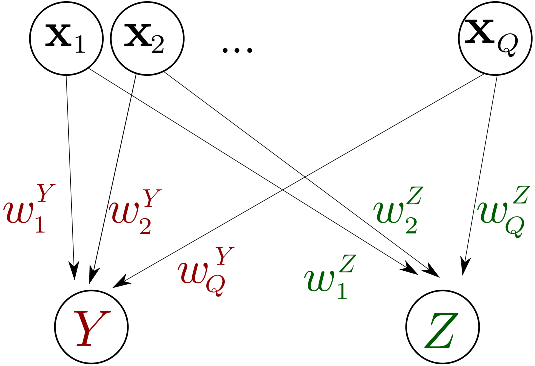


Manifold Relevance Determination

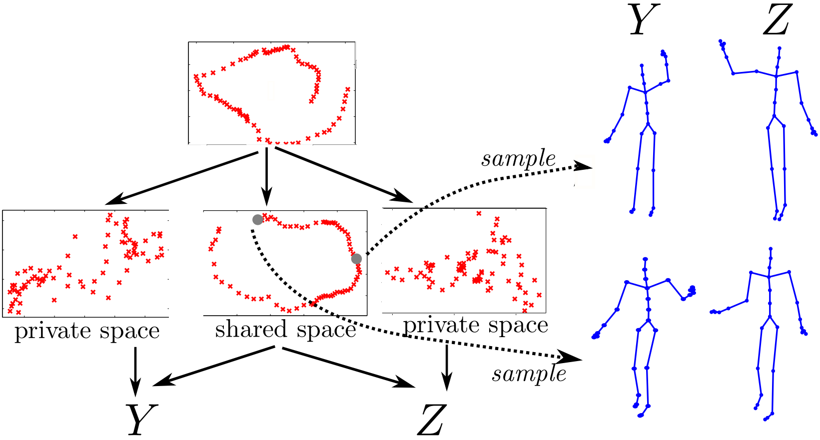


- Observations come into two different *views*: Y and Z.
- The latent space is segmented into parts private to Y, private to Z and shared between Y and Z.
- Used for data consolidation and discovering commonalities.

MRD weights



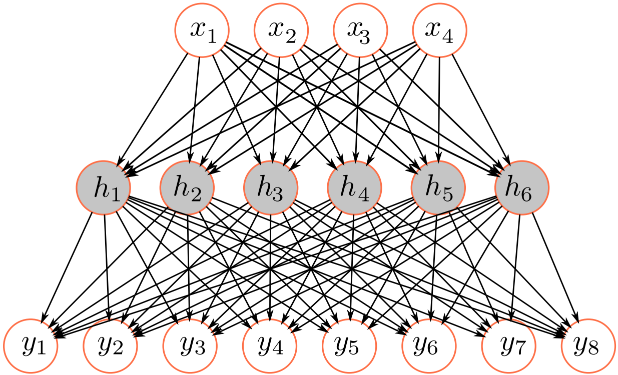
Deep GPs: Another multi-view example



Automatic structure discovery

Tools:

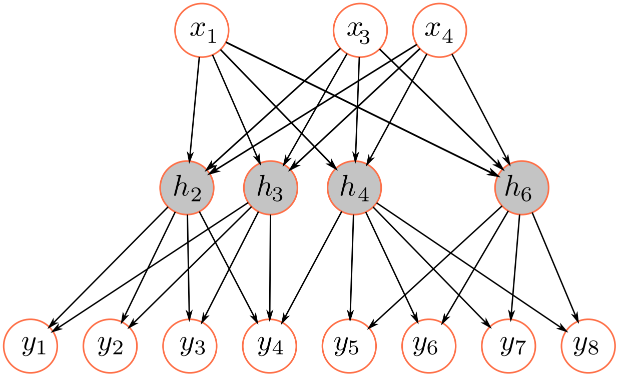
- ARD: Eliminate unnecessary nodes/connections
- MRD: Conditional independencies
- Approximating evidence: Number of layers (?)



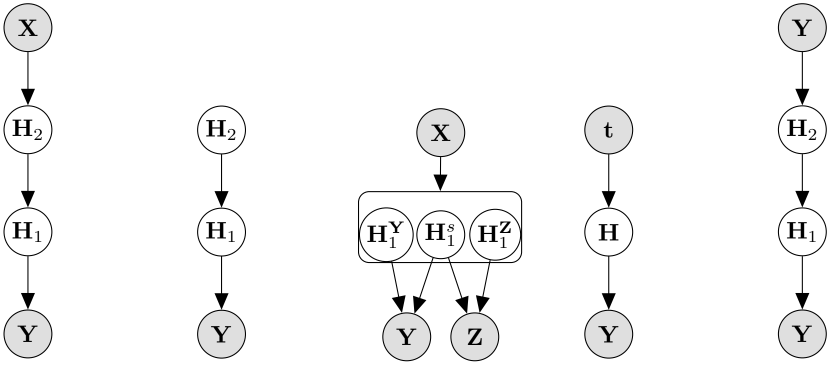
Automatic structure discovery

Tools:

- ARD: Eliminate unnecessary nodes/connections
- MRD: Conditional independencies
- Approximating evidence: Number of layers (?)



Deep GP variants



Deep GP - Supervised

Deep GP - Unsupervised

Multi-view

Temporal

Autoencoder

Summary

- A deep GP is not a GP.
- Sampling is straight-forward. Regularization and training needs to be worked out.
- The solution is a special treatment of auxiliary variables.
- Many variants: multi-view, temporal, autoencoders ...
- Future: make it scalable with distributed computations.
- Future: how does it compare to / complement more traditional deep models?

Thanks

Thanks to Neil Lawrence, Carl Henrik Ek, James Hensman, Michalis Titsias.

References:

- N. D. Lawrence (2006) "The Gaussian process latent variable model" Technical Report no CS-06-03, The University of Sheffield, Department of Computer Science
- N. D. Lawrence (2006) "Probabilistic dimensional reduction with the Gaussian process latent variable model" (talk)
- C. E. Rasmussen (2008), "Learning with Gaussian Processes", Max Planck Institute for Biological Cybernetics, Published: Feb. 5, 2008 (Videolectures.net)
- Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, 2006. ISBN 026218253X.
- M. K. Titsias (2009), "Variational learning of inducing variables in sparse Gaussian processes", AISTATS 2009
- A. C. Damianou, M. K. Titsias and N. D. Lawrence (2011), "Variational Gaussian process dynamical systems", NIPS 2011
- A. C. Damianou, C. H. Ek, M. K. Titsias and N. D. Lawrence (2012), "Manifold Relevance Determination", ICML 2012
- A. C. Damianou and N. D. Lawrence (2013), "Deep Gaussian processes", AISTATS 2013
- J. Hensman (2013), "Gaussian processes for Big Data", UAI 2013