Tutorial on Gaussian Processes and the Gaussian Process Latent Variable Model

(& discussion on the GPLVM tech. report by Prof. N. Lawrence, '06)

Andreas Damianou

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

University of Surrey, 13/06/2012

Outline

Part 1: Gaussian processes
- Parametric models: ML and Bayesian regression
- Nonparametric models: Gaussian process regression
- Covariance functions

Part 2: Gaussian Process Latent Variable Model
- Dimensionality Reduction: Motivation
- From probabilistic PCA to Dual PPCA
- From Dual PPCA to GP-LVM

Part3: Applications of GP-LVM in vision

Introducing Gaussian Processes: Outline

			Bayesian	Non-parametric
From:	ML / MAP	Regression, to	✗	✗
	Bayesian	Regression, to	✓	✗
	GP	Regression	✓	✓

Maximum Likelihood Regression

$$\begin{array}{cccc} \text{Data:} & X & \rightarrow & Y \\ & \downarrow & & \downarrow \\ & \text{inputs} & & \text{targets} \end{array}$$

- Regression: Assume a *parametric* model with parameters θ
- Likelihood $\mathcal{L}(\theta) = p(Y|X, \theta)$ is obtained from the PDF of the assumed distribution

Maximum Likelihood Regression

$$\begin{array}{cccc} \text{Data:} & X & \rightarrow & Y \\ & \downarrow & & \downarrow \\ & \text{inputs} & & \text{targets} \end{array}$$

- Regression: Assume a *parametric* model with parameters θ
- Likelihood $\mathcal{L}(\theta) = p(Y|X, \theta)$ is obtained from the PDF of the assumed distribution
- Example: Linear Regression
    - $Y = f(X, W) + \epsilon, \;\; f(X, W) = WX, \;\; \epsilon \sim \mathcal{N}(0, \beta^{-1})$

Maximum Likelihood Regression

$$\begin{array}{cccc} \text{Data:} & X & \rightarrow & Y \\ & \downarrow & & \downarrow \\ & \text{inputs} & & \text{targets} \end{array}$$

- Regression: Assume a *parametric* model with parameters θ
- Likelihood $\mathcal{L}(\theta) = p(Y|X,\theta)$ is obtained from the PDF of the assumed distribution
- Example: Linear Regression
  - $Y = f(X,W) + \epsilon, \;\; f(X,W) = WX, \;\; \epsilon \sim \mathcal{N}(0, \beta^{-1})$
  - $\mathcal{L}(\theta) = p(Y|X,\theta) = \mathcal{N}(Y|WX, \beta^{-1}), \;\; \theta = \{W, \beta^{-1}\}$

Maximum Likelihood Regression

$$\begin{array}{cccc} \text{Data:} & X & \rightarrow & Y \\ & \downarrow & & \downarrow \\ & \text{inputs} & & \text{targets} \end{array}$$

- Regression: Assume a *parametric* model with parameters θ
- Likelihood $\mathcal{L}(\theta) = p(Y|X, \theta)$ is obtained from the PDF of the assumed distribution
- Example: Linear Regression
  - $Y = f(X, W) + \epsilon, \quad f(X, W) = WX, \quad \epsilon \sim \mathcal{N}(0, \beta^{-1})$
  - $\mathcal{L}(\theta) = p(Y|X, \theta) = \mathcal{N}(Y|WX, \beta^{-1}), \quad \theta = \{W, \beta^{-1}\}$
  - *Optimise*: $\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta)$. *Predictions* based on $p(y^*|x^*, \hat{\theta})$

Bayesian parametric model

- Bayes rule: $\overbrace{p(\theta|X,Y)}^{\text{posterior}} = \dfrac{\overbrace{p(Y|X,\theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(Y|X)}_{\text{evidence}}}$

Bayesian parametric model

- Bayes rule: $\overbrace{p(\theta|X,Y)}^{\text{posterior}} = \dfrac{\overbrace{p(Y|X,\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(Y|X)}_{\text{evidence}}}$

- Predictions via marginalisation:

$$p(y^*|x^*,X,Y) = \int \underbrace{p(y^*|x^*,X,Y,\theta)}_{\text{likelihood}}\,\underbrace{p(\theta|X,Y)}_{\text{posterior}}\,\mathrm{d}\theta$$

Bayesian parametric model

- Bayes rule: $\overbrace{p(\theta|X,Y)}^{\text{posterior}} = \dfrac{\overbrace{p(Y|X,\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(Y|X)}_{\text{evidence}}}$

- Predictions via marginalisation:
  $p(y^*|x^*,X,Y) = \int \underbrace{p(y^*|x^*,X,Y,\theta)}_{\text{likelihood}}\underbrace{p(\theta|X,Y)}_{\text{posterior}}\,\mathrm{d}\theta$

- θ is integrated out, but we still assume a *parametric model* (i.e. $f(X,\theta = WX, \ \ \theta = \{W,\beta\})$

- The integral (and sometimes $p(Y|X)$) are often intractable

Gaussian process nonparametric models

	Mapping	Prior
Parametric	$f(X, \theta) = WX$	on the function parameters ($p(\theta) = p(W)$)
Nonparametric (GP)	$f \sim \mathcal{GP}$	on the function itself

Gaussian process nonparametric models

	Mapping	Prior
Parametric	$f(X,\theta) = WX$	on the function parameters $(p(\theta) = p(W))$
Nonparametric (GP)	$f \sim \mathcal{GP}$	on the function itself

- A GP is a prior over functions. It depends on a mean and a covariance *function* (NOT matrix!)
-

$$\text{Prior: } f_n = f(x_n) \sim \mathcal{GP}\left(m(x_n), k(x_n, x'_n)\right) \rightarrow \textit{infinite}$$

$$\text{Joint: } f^*, F \sim \mathcal{N}\left(\boldsymbol{\mu}^*, K^*\right) \rightarrow \textit{finite} \quad (F = \{f_n\}_{n=1}^N)$$

- Posterior/predictive process/distribution $f^*|\mathbf{f}$ is also Gaussian!

Gaussian process nonparametric models

(modified from C. E. Rasmussen's tutorial, "Learning with Gaussian Processes")

- Gaussian Likelihood: $Y|X, f(x) = \mathcal{N}(Y|F, \beta^{-1}I)$
- (Zero mean) GP prior: $f(x) \sim \mathcal{GP}(\mathbf{0}, k(x, x'))$

Gaussian process nonparametric models

(modified from C. E. Rasmussen's tutorial, "Learning with Gaussian Processes")

- Gaussian Likelihood: $Y|X, f(x) = \mathcal{N}(Y|F, \beta^{-1}I)$
- (Zero mean) GP prior: $f(x) \sim \mathcal{GP}(\mathbf{0}, k(x, x'))$
- Leads to a GP posterior:

$$f(x)|X,Y \sim \mathcal{GP}\Big(m_{post} = k(x, X)K^{-1}(X, X)F,$$
$$k_{post}(x, x') = k(x, x') - k(x, X)K^{-1}(X, X)k(X, x)\Big)$$

Gaussian process nonparametric models

(modified from C. E. Rasmussen's tutorial, "Learning with Gaussian Processes")

- Gaussian Likelihood: $Y|X, f(x) = \mathcal{N}(Y|F, \beta^{-1}I)$
- (Zero mean) GP prior: $f(x) \sim \mathcal{GP}(\mathbf{0}, k(x, x'))$
- Leads to a GP posterior:

$$f(x)|X,Y \sim \mathcal{GP}\Big(m_{post} = k(x, X)K^{-1}(X, X)F,$$
$$k_{post}(x, x') = k(x, x') - k(x, X)K^{-1}(X, X)k(X, x)\Big)$$

- ... and a Gaussian predictive distribution:

$$y^*|x^*, X, Y \sim \mathcal{N}\Big(k(x^*, X)\left[K(X, X) + \beta^{-1}I\right]^{-1} Y,$$
$$k(x^*, x^*) + \beta^{-1} - k(x^*, X)\left[K(X, X) + \beta^{-1}I\right]^{-1} k(X, x^*)\Big)$$

Covariance functions

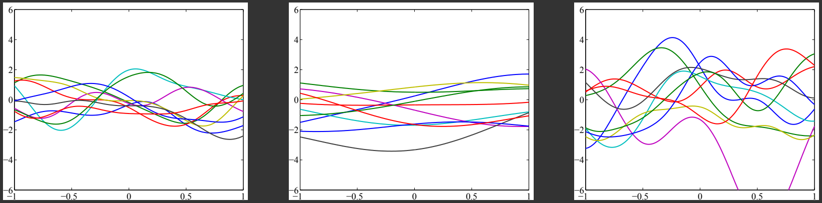
- But where did $k(x, x')$ (and $K(\mathbf{x}, \mathbf{x})$ etc.) come from?
- Assumptions about *properties* of $f \Rightarrow$ define a parametric form for k, e.g:

$$k(x, x') = \alpha \exp\left(-\frac{\gamma}{2}(x - x')^{\mathrm{T}}(x - x')\right)$$

- However, a prior with this cov. function defines a whole *family* of functions
- The parameters $\{\alpha, \gamma\}$ are *hyperparameters*.

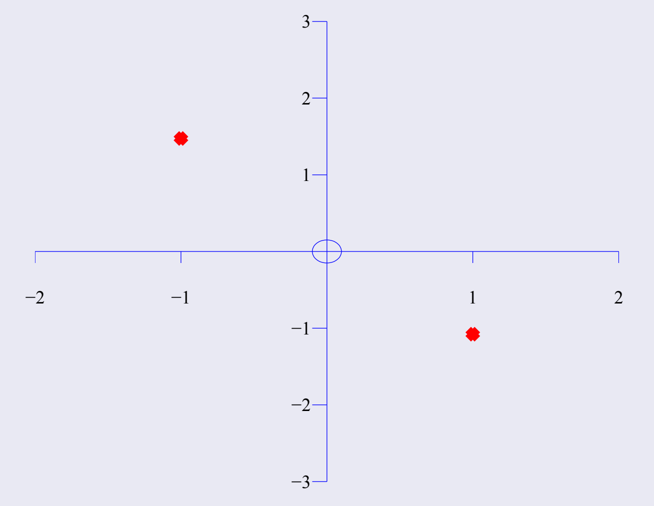
Covariance samples and hyperparameters

- The hyperparameters of the cov. function define the properties (and NOT an explicit form) of the sampled functions



Gaussian Process Regression - demo

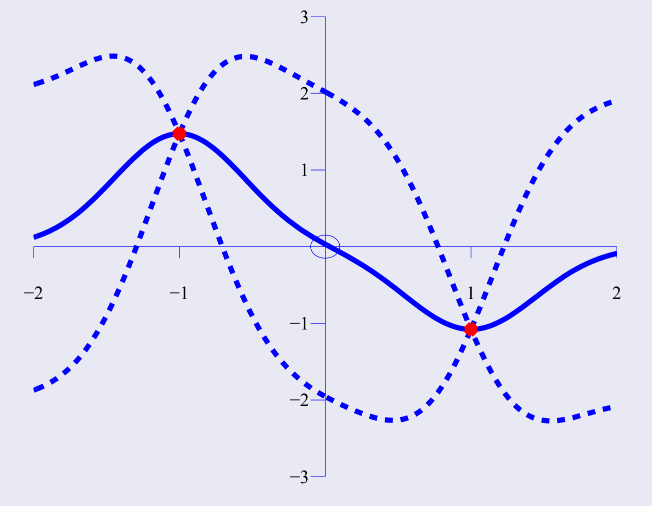
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

Gaussian Process Regression - demo

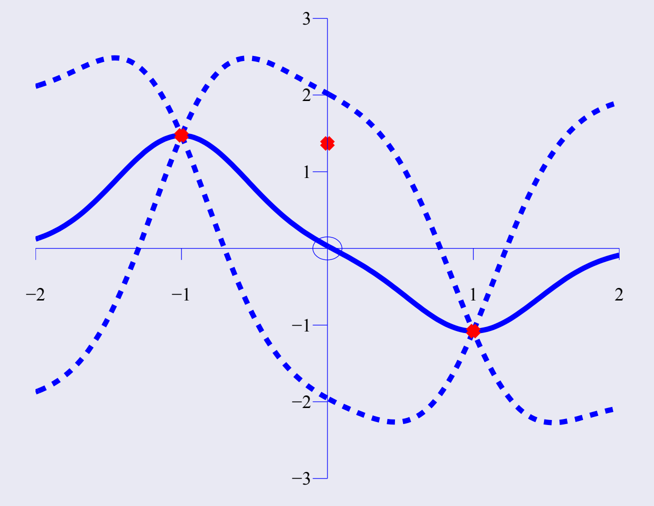
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

Gaussian Process Regression - demo

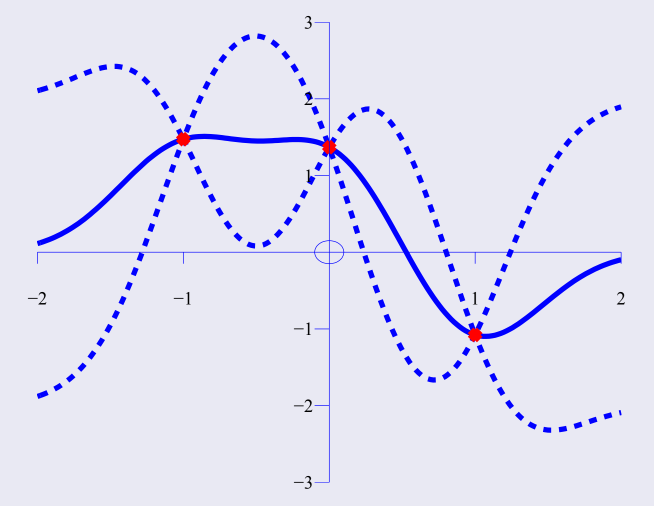
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

Gaussian Process Regression - demo

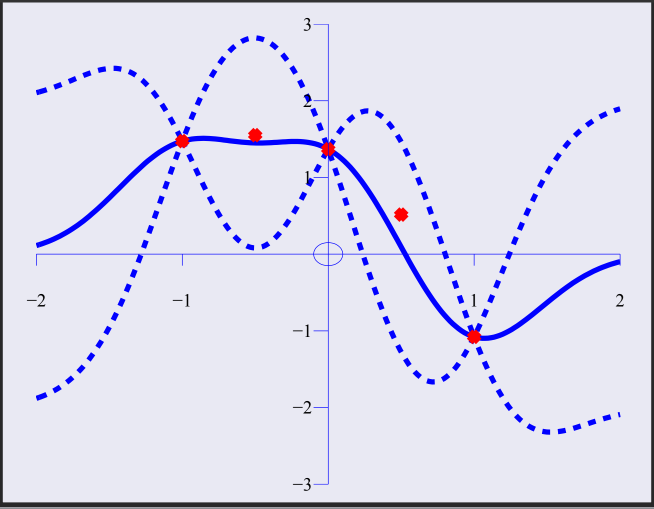
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

Gaussian Process Regression - demo

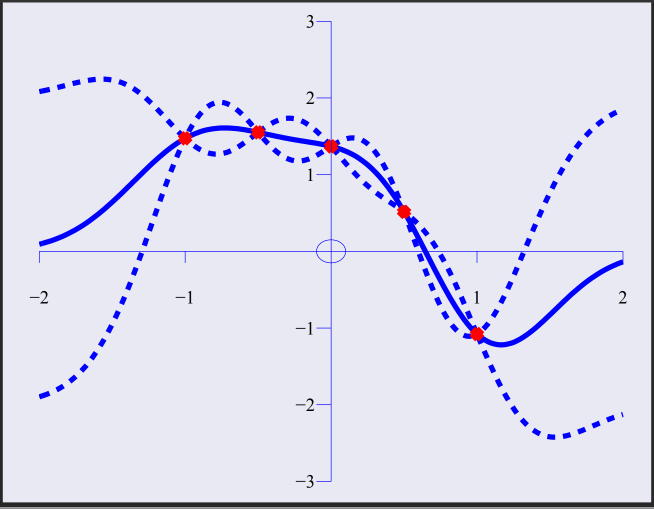
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

Gaussian Process Regression - demo

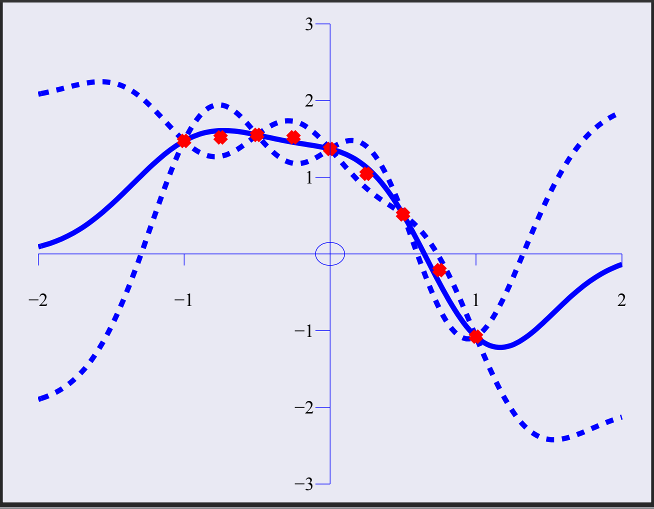
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

Gaussian Process Regression - demo

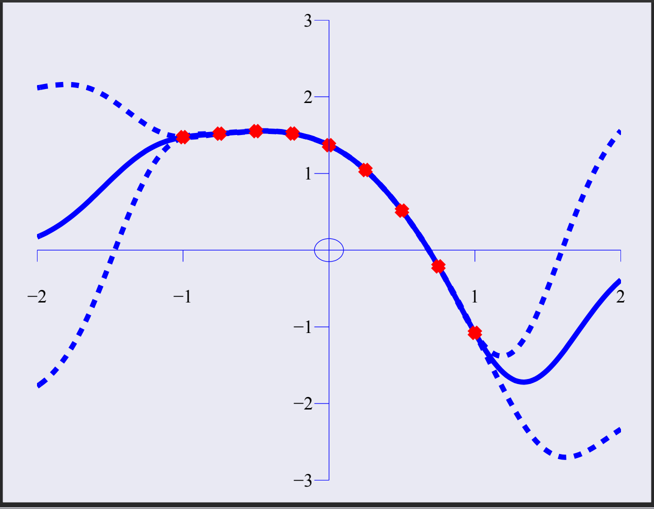
(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

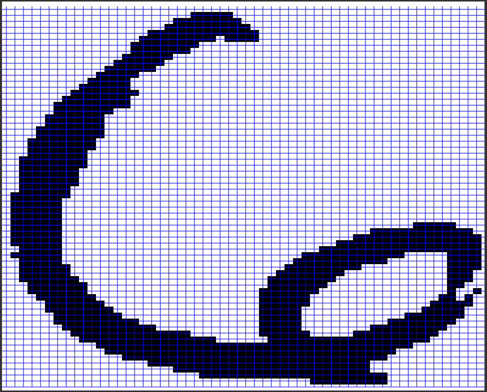
Gaussian Process Regression - demo

(source: N. Lawrence's talk, "Learning and Inference with Gaussian Processes" (2005))



Observing more and more data: Prior + data likelihood get combined

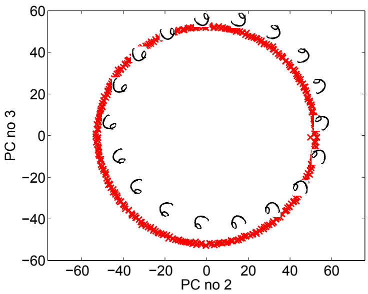
Dimensionality Reduction: Motivation

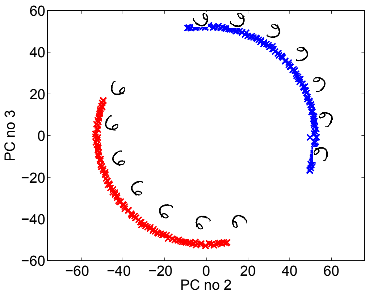


- Each data-point is $64 \times 57 = 3,648$-dimensional (pixel space)
- However, intrinsic dimensionality is lower

Dimensionality Reduction: Motivation

- Consider digit rotations
- Create a new dataset, where a prototype is repeated under one of 360 different angles
- Project into principal components 2 and 3
- Low-dimensional embedding $(3,648 \to 2$ dimensions) captures all necessary information





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 *(compare with regression)*:

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

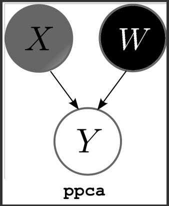
- $p(Y|W,X,\beta) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n | W\mathbf{x}_n, \beta^{-1}\mathbf{I})$

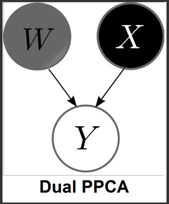
- $W, X \in \mathbb{R}^{N \times Q}$, $Q \ll D$

- X are unobserved

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.





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

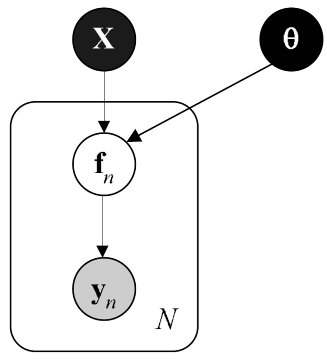
$$p(Y|X,\beta) = \int p(Y|X,W,\beta)p(W)\mathrm{d}W$$

Gaussian process latent variable model (GP-LVM)

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

$$k(x, x') = \alpha \exp\left(-\frac{\gamma}{2}(x - x')^{\mathrm{T}}(x - x')\right)$$

Gaussian process latent variable model (GP-LVM)



Dimensionality reduction: Linear vs non-linear

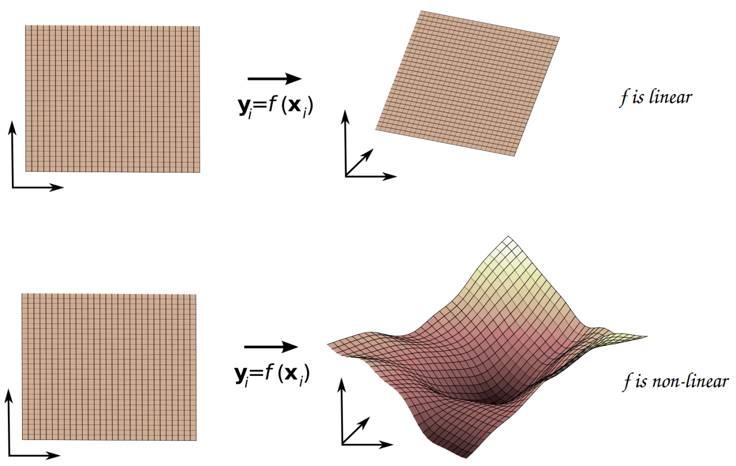


Image from: "Dimensionality Reduction the Probabilistic Way", N. Lawrence, ICML tutorial 2008

Applications of the GP-LVM in vision

- Modelling human motion (inverse kinematics [1] , body parts decomposition [4] , ...) *(Show video...)*
- Animation [2] *(Show video...)*
- Tracking [3]
- Reconstruction & probabilistic generation of HD video/high res. images [5,6]
- ...

[1] Grochow et al. (2004), Style-based Inverse Kinematics (SIGGRAPH)

[2] Baxter and Anjyo (2006), Latent Doodle Space (Eurographics)

[3] Urtasun et al. (2005), Priors for People Tracking from Small Training Sets

[4] Lawrence and Moore. (2007), Hierarchical Gaussian process latent variable models (ICML)

[5] Damianou et al. (2011), Variational Gaussian process dynamical systems (NIPS)

[6] Damianou et al. (2012), Manifold Relevance Determination (ICML)

Main sources:

- 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) "Learning and inference with Gaussian processes: an overview of Gaussian processes and the GP-LVM". Presented at University of Manchester, Machine Learning Course Guest Lecture on 3/11/2006
- 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.
- N. D. Lawrence, lecture notes for "Machine Learning and Adaptive Intelligence" (2012)