

Variational Gaussian Processes

GP Variational Approximations

Neil D. Lawrence

Departments of Neuro- and Computer Science, University of
Sheffield, U.K.

MPI Tuebingen

11th March 2013

Outline

Notation

q — dimension of latent/embedded space

p — dimension of data space

n — number of data points

centred data, $\mathbf{Y} = [\mathbf{y}_{1,:}, \dots, \mathbf{y}_{n,:}]^\top = [\mathbf{y}_{:,1}, \dots, \mathbf{y}_{:,p}] \in \mathbb{R}^{n \times p}$

latent variables, $\mathbf{X} = [\mathbf{x}_{1,:}, \dots, \mathbf{x}_{n,:}]^\top = [\mathbf{x}_{:,1}, \dots, \mathbf{x}_{:,q}] \in \mathbb{R}^{n \times q}$

mapping matrix, $\mathbf{W} \in \mathbb{R}^{p \times q}$

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

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

Reading Notation

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

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

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

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

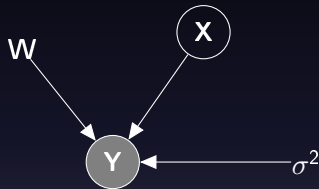
where

$$\boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:
 - Define Gaussian prior over *latent space*, $\mathbf{X}$.
 - Integrate out *latent variables*.

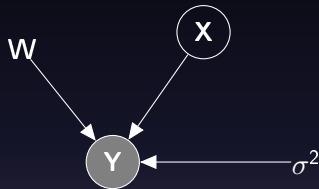


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:
 - Define Gaussian prior over *latent space*, $\mathbf{X}$.
 - Integrate out *latent variables*.

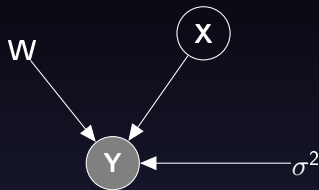


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:
 - Define Gaussian prior over *latent space*, $\mathbf{X}$.
 - Integrate out *latent variables*.



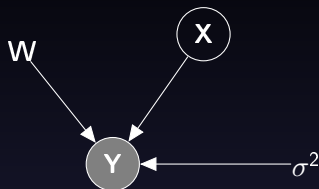
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{X}) = \prod_{i=1}^n \mathcal{N}(\mathbf{x}_{i,:} | \mathbf{0}, \mathbf{I})$$

Linear Latent Variable Model

Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Standard** Latent variable approach:
 - Define Gaussian prior over *latent space*, $\mathbf{X}$.
 - Integrate out *latent variables*.



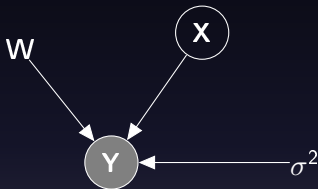
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

$$p(\mathbf{X}) = \prod_{i=1}^n \mathcal{N}(\mathbf{x}_{i,:} | \mathbf{0}, \mathbf{I})$$

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

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)



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

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

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

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

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

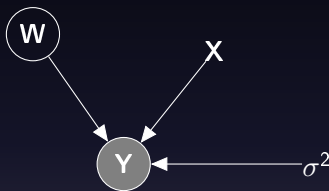
$$\mathbf{W} = \mathbf{U}_q \mathbf{L} \mathbf{R}^\top, \quad \mathbf{L} = (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I})^{\frac{1}{2}}$$

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

Linear Latent Variable Model III

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*.

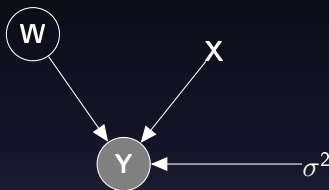


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model III

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.

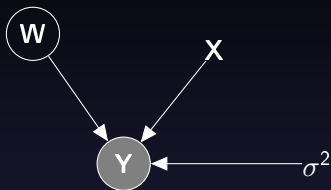


$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W} \mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

Linear Latent Variable Model III

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*.



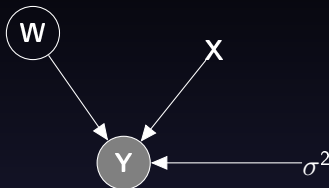
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W} \mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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

Linear Latent Variable Model III

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*.



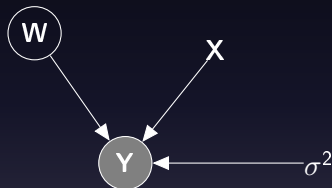
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_{i,:} | \mathbf{W}\mathbf{x}_{i,:}, \sigma^2 \mathbf{I})$$

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

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

Linear Latent Variable Model IV

Dual Probabilistic PCA Max. Likelihood Soln (Lawrence, 2004, 2005)



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

Linear Latent Variable Model IV

Dual Probabilistic PCA Max. Likelihood Soln (Lawrence, 2004, 2005)

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

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

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

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

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

Linear Latent Variable Model IV

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

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

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

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

$$\mathbf{W} = \mathbf{U}_q \mathbf{L} \mathbf{R}^\top, \quad \mathbf{L} = (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I})^{\frac{1}{2}}$$

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

Equivalence of Formulations

The Eigenvalue Problems are equivalent

- Solution for Probabilistic PCA (solves for the mapping)

$$\mathbf{Y}^\top \mathbf{Y} \mathbf{U}_q = \mathbf{U}_q \mathbf{\Lambda}_q \quad \mathbf{W} = \mathbf{U}_q \mathbf{L} \mathbf{R}^\top$$

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

$$\mathbf{Y} \mathbf{Y}^\top \mathbf{U}'_q = \mathbf{U}'_q \mathbf{\Lambda}_q \quad \mathbf{X} = \mathbf{U}'_q \mathbf{L} \mathbf{R}^\top$$

- Equivalence is from

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

Sampling a Function

Multi-variate Gaussians

- We will consider a Gaussian with a particular structure of covariance matrix.
- Generate a single sample from this 25 dimensional Gaussian distribution, $\mathbf{f} = [f_1, f_2 \dots f_{25}]$.
- We will plot these points against their index.

Gaussian Distribution Sample

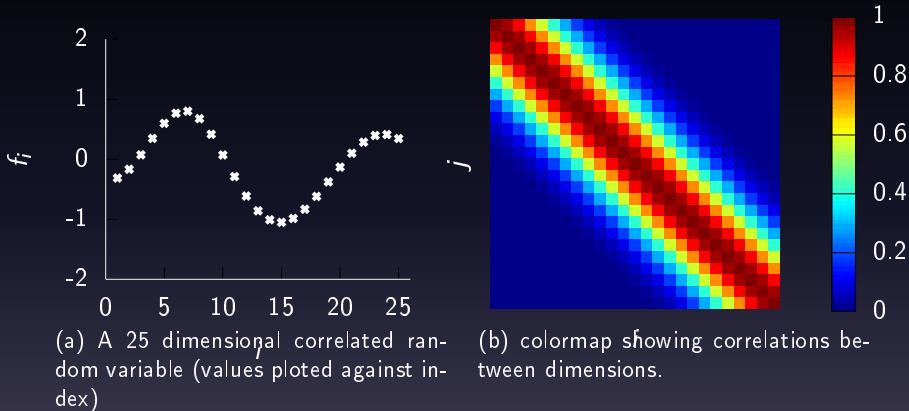


Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample

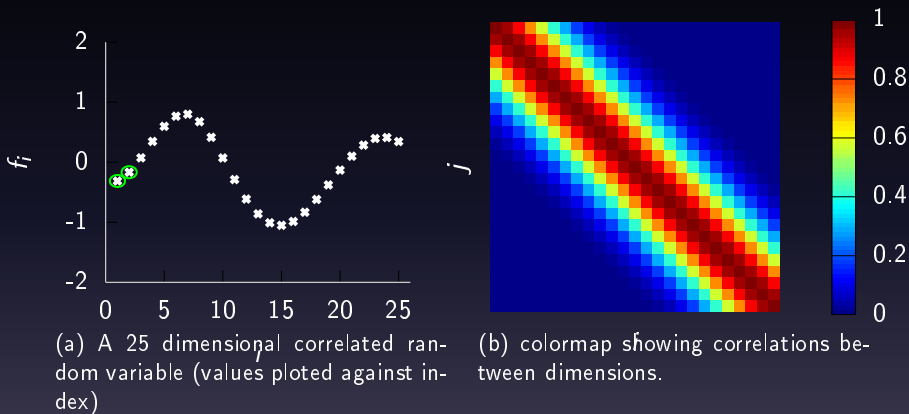
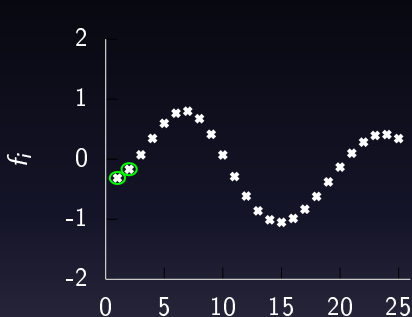
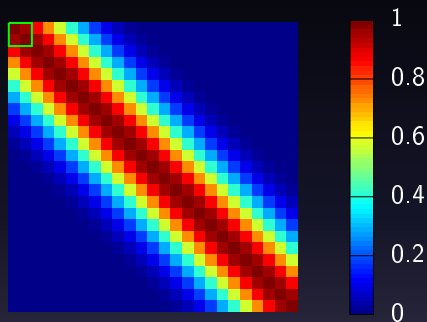


Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



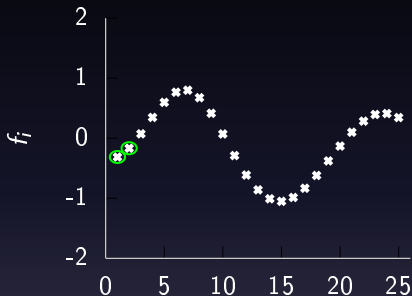
(a) A 25 dimensional correlated random variable (values plotted against index)



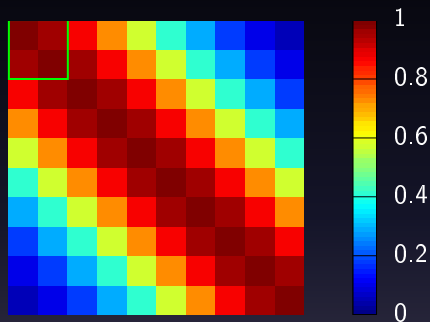
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



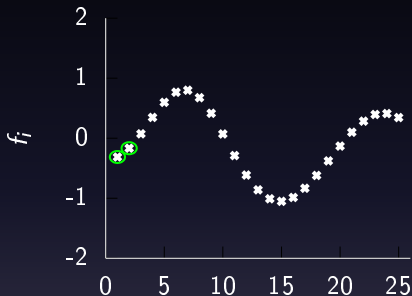
(a) A 25 dimensional correlated random variable (values plotted against index)



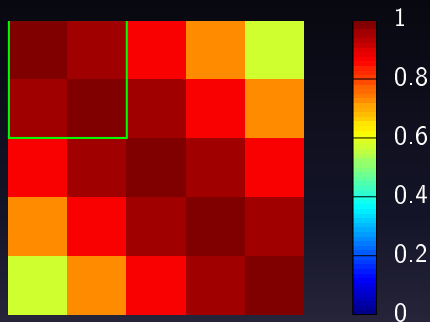
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



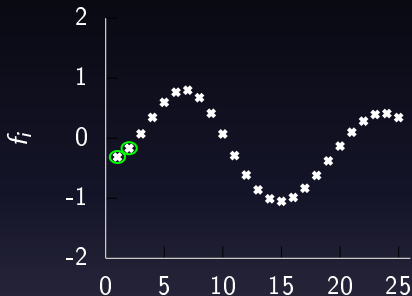
(a) A 25 dimensional correlated random variable (values plotted against index)



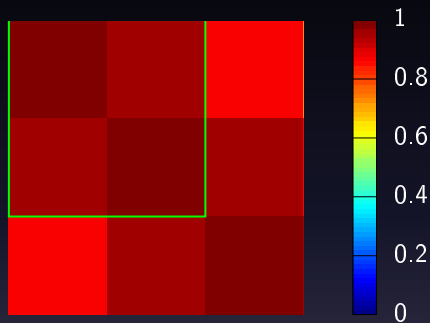
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



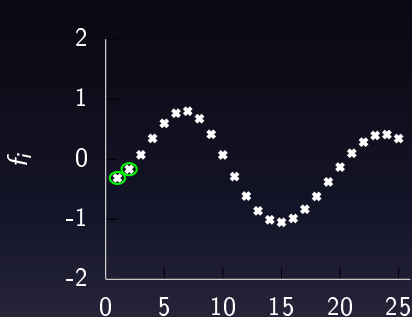
(a) A 25 dimensional correlated random variable (values plotted against index)



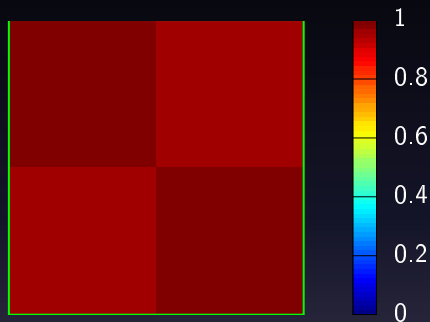
(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



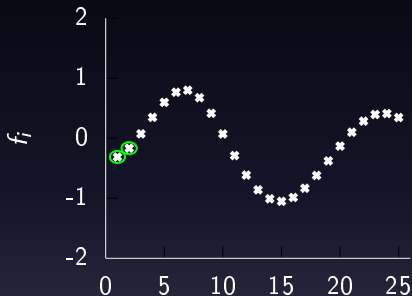
(a) A 25 dimensional correlated random variable (values plotted against index)



(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.

Gaussian Distribution Sample



(a) A 25 dimensional correlated random variable (values plotted against index)

$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

(b) correlation between f_1 and f_2 .

Figure: A sample from a 25 dimensional Gaussian distribution.

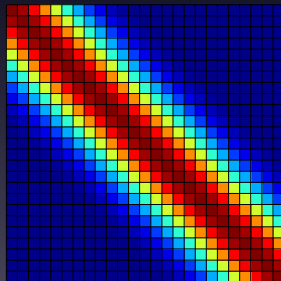
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2} \right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.



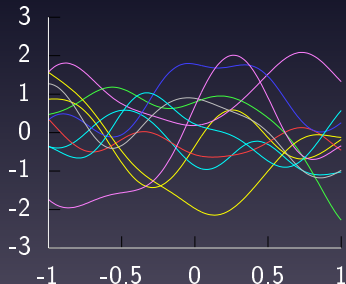
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2} \right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.



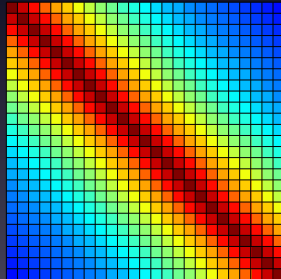
Covariance Functions

Where did this covariance matrix come from?

Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|}{2\ell^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.



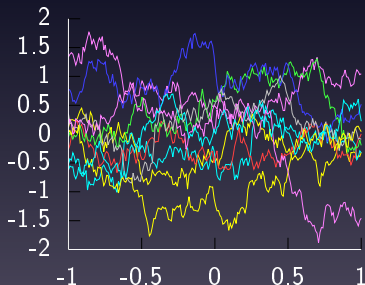
Covariance Functions

Where did this covariance matrix come from?

Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|}{2\ell^2}\right)$$

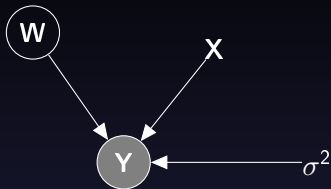
- Covariance matrix is built using the *inputs* to the function $\mathbf{x}$.



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 *parameters*, $\mathbf{W}$.
 - Integrate out *parameters*.



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

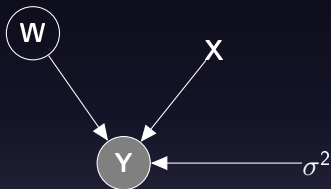
$$p(\mathbf{W}) = \prod_{i=1}^p \mathcal{N}(\mathbf{w}_{i,:} | \mathbf{0}, \mathbf{I})$$

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

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).

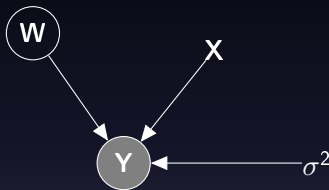


$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})$$

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).



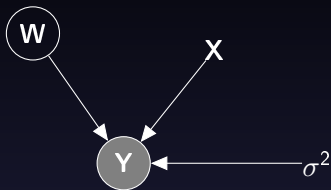
$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^p \mathcal{N}(\mathbf{y}_{:,j} | \mathbf{0}, \mathbf{K})$$

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

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).



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

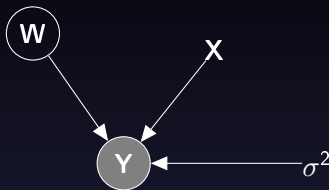
$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I}$$

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).



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

$$\mathbf{K} = ?$$

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

Non-linear Latent Variable Models

Exponentiated Quadratic (EQ) Covariance

- The EQ covariance has the form $k_{ij} = k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:})$, where

$$k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}) = \alpha \exp \left(-\frac{\|\mathbf{x}_{i,:} - \mathbf{x}_{j,:}\|_2^2}{2\ell^2} \right).$$

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

Learning in Larger Datasets

(Lawrence, 2007; Titsias, 2009)

- Complexity of standard GP:
 - $O(n^3)$ in computation.
 - $O(n^2)$ in storage.
- Via low rank representations of covariance:
 - $O(nm^2)$ in computation.
 - $O(nm)$ in storage.
- Where m is user chosen number of *inducing* variables. They give the rank of the resulting covariance.

Inducing Variable Approximations

- Date back to (Williams and Seeger, 2001; Smola and Bartlett, 2001; Csató and Opper, 2002; Seeger et al., 2003; Snelson and Ghahramani, 2006). See Quiñero Candela and Rasmussen (2005) for a review.
- We follow variational perspective of (Titsias, 2009).
- This is an augmented variable method, followed by a collapsed variational approximation (King and Lawrence, 2006; Hensman et al., 2012).

Augmented Variable Model

Augment standard GP model with a set of m new inducing variables, $\mathbf{u}$.

$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{u}) d\mathbf{u}$$

y

Augmented Variable Model

Augment standard GP model with a set of m new inducing variables, $\mathbf{u}$.

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u}$$



Augmented Variable Model

Assume that relationship is through $\mathbf{f}$.

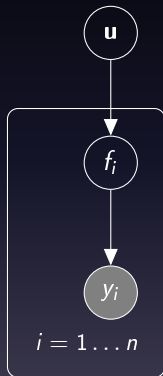
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}$$



Augmented Variable Model

Very often likelihood factorizes.

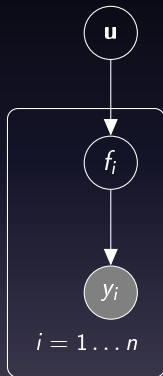
$$p(\mathbf{y}) = \int \prod_{i=1}^n p(y_i | f_i) p(\mathbf{f} | \mathbf{u}) p(\mathbf{u}) d\mathbf{f} d\mathbf{u}$$



Augmented Variable Model

Focus on integral over $\mathbf{f}$.

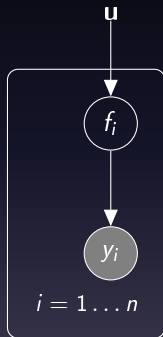
$$p(\mathbf{y}) = \int \int \prod_{i=1}^n p(y_i | f_i) p(\mathbf{f} | \mathbf{u}) d\mathbf{f} p(\mathbf{u}) d\mathbf{u}$$



Augmented Variable Model

Focus on integral over $\mathbf{f}$.

$$p(\mathbf{y}|\mathbf{u}) = \int \prod_{i=1}^n p(y_i|f_i)p(\mathbf{f}|\mathbf{u})d\mathbf{f}$$



Variational Bound on $p(\mathbf{y}|\mathbf{u})$

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{u}) &= \log \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})d\mathbf{f} \\ &\geq \int q(\mathbf{f}) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})}{q(\mathbf{f})}d\mathbf{f}\end{aligned}$$

- For variational approximation of (Titsias, 2009) set $q(\mathbf{f}) = p(\mathbf{f}|\mathbf{u})$,

$$\log p(\mathbf{y}|\mathbf{u}) \geq \log \int p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})d\mathbf{f}.$$

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \int p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})d\mathbf{f}.$$

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \int p(\mathbf{f}|\mathbf{u}) \log \prod_{i=1}^n p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \int p(\mathbf{f}|\mathbf{u}) \log \prod_{i=1}^n p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \int p(\mathbf{f}|\mathbf{u}) \sum_{i=1}^n \log p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \int p(\mathbf{f}|\mathbf{u}) \sum_{i=1}^n \log p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \sum_{i=1}^n \int p(f_i|\mathbf{u}) \log p(y_i|f_i) df_i.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \exp \sum_{i=1}^n \int p(f_i|\mathbf{u}) \log p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \prod_{i=1}^n \exp \int p(f_i|\mathbf{u}) \log p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \prod_{i=1}^n \exp \int p(f_i|\mathbf{u}) \log p(y_i|f_i) d\mathbf{f}.$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \prod_{i=1}^n \exp \langle \log p(y_i|f_i) \rangle_{p(f_i|\mathbf{u})}$$

- Then the bound factorizes.

Factorizing Likelihoods

- If the likelihood, $p(\mathbf{y}|\mathbf{f})$, factorizes

$$p(\mathbf{y}|\mathbf{u}) \geq \prod_{i=1}^n \exp \langle \log p(y_i|f_i) \rangle_{p(f_i|\mathbf{u})}$$

- Then the bound factorizes.

Gaussian $p(y_i|f_i)$

For Gaussian likelihoods:

$$\langle \log p(y_i|f_i) \rangle_{p(f_i|\mathbf{u})} = -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (y_i - \langle f_i \rangle)^2 - \frac{1}{2\sigma^2} \left(\langle f_i^2 \rangle - \langle f_i \rangle^2 \right)$$

Gaussian $p(y_i|f_i)$

For Gaussian likelihoods:

$$\langle \log p(y_i|f_i) \rangle_{p(f_i|\mathbf{u})} = -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (y_i - \langle f_i \rangle)^2 - \frac{1}{2\sigma^2} \left(\langle f_i^2 \rangle - \langle f_i \rangle^2 \right)$$

Implying:

$$p(y_i|\mathbf{u}) \geq \exp \langle \log c_i \rangle \mathcal{N}(y_i | \langle f_i \rangle, \sigma^2)$$

Gaussian Process Over $\mathbf{f}$ and $\mathbf{u}$

Define:

$$q_{i,i} = \text{var}_{p(f_i|\mathbf{u})} (f_i) = \langle f_i^2 \rangle_{p(f_i|\mathbf{u})} - \langle f_i \rangle_{p(f_i|\mathbf{u})}^2$$

We can write:

$$c_i = \exp \left(-\frac{q_{i,i}}{2\sigma^2} \right)$$

If joint distribution of $p(\mathbf{f}, \mathbf{u})$ is Gaussian then:

$$q_{i,i} = k_{i,i} - \mathbf{k}_{i,\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{k}_{i,\mathbf{u}}$$

c_i is not a function of $\mathbf{u}$ but *is* a function of $\mathbf{X}_{\mathbf{u}}$.

Lower Bound on Likelihood

Substitute variational bound into marginal likelihood:

$$p(\mathbf{y}) \geq \prod_{i=1}^n c_i \int \mathcal{N}(\mathbf{y} | \langle \mathbf{f} \rangle, \sigma^2 \mathbf{I}) p(\mathbf{u}) d\mathbf{u}$$

Note that:

$$\langle \mathbf{f} \rangle_{p(\mathbf{f}|\mathbf{u})} = \mathbf{K}_{\mathbf{f},\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}$$

is *linearly* dependent on $\mathbf{u}$.

Deterministic Training Conditional

Making the marginalization of $\mathbf{u}$ straightforward. In the Gaussian case:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$$

$$\int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \geq \prod_{i=1}^n c_i \int \mathcal{N}(\mathbf{y}|\mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \sigma^2) \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}}) d\mathbf{u}$$

Deterministic Training Conditional

Making the marginalization of $\mathbf{u}$ straightforward. In the Gaussian case:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$$

$$\int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \geq \prod_{i=1}^n c_i \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}}\right)$$

Deterministic Training Conditional

Making the marginalization of $\mathbf{u}$ straightforward. In the Gaussian case:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$$

$$\int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \geq \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}})$$

Maximize log of the bound to find covariance function parameters,

$$L \geq \sum_{i=1}^n \log c_i + \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}})$$

Deterministic Training Conditional

Making the marginalization of $\mathbf{u}$ straightforward. In the Gaussian case:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$$

$$\int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \geq \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}})$$

Maximize log of the bound to find covariance function parameters,

$$L \geq \sum_{i=1}^n \log c_i + \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}})$$

Deterministic Training Conditional

Making the marginalization of $\mathbf{u}$ straightforward. In the Gaussian case:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$$

$$\int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \geq \prod_{i=1}^n c_i \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}}\right)$$

Maximize log of the bound to find covariance function parameters,

$$L \approx \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}}\right)$$

- If the bound is normalized, the c_i terms are removed.
- This results in the projected process approximation (Rasmussen and Williams, 2006) or DTC (Quiñonero Candela and Rasmussen, 2005).
Proposed by (Smola and Bartlett, 2001; Seeger et al., 2003; Csató and Opper, 2002; Csató, 2002).

Deterministic Training Conditional

Making the marginalization of $\mathbf{u}$ straightforward. In the Gaussian case:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$$

$$\int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \geq \prod_{i=1}^n c_i \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{K}_{\mathbf{f},\mathbf{u}}^\top \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{K}_{\mathbf{u},\mathbf{f}})$$

Maximize log of the bound to find covariance function parameters,

- If the bound is normalized, the c_i terms are removed.
- This results in the projected process approximation (Rasmussen and Williams, 2006) or DTC (Quiñonero Candela and Rasmussen, 2005).
Proposed by (Smola and Bartlett, 2001; Seeger et al., 2003; Csató and Opper, 2002; Csató, 2002).

Stochastic Variational Inference

- James Hensman has shown this same bound can be used for *stochastic* variational inference (Blei et al).
- By retaining the variational distribution over $\mathbf{u}$ explicitly.
- Currently running GPs over 120,000,000 data points.
- This bound seems extendable to deep GPs.

References I

- L. Csató. *Gaussian Processes — Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.
- L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.
- J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the exponential family. *NIPS 2012*, 2012.
- N. J. King and N. D. Lawrence. Fast variational inference for Gaussian Process models through KL-correction. In *ECML, Berlin, 2006*, Lecture Notes in Computer Science, pages 270–281, Berlin, 2006. Springer-Verlag. [PDF].
- 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, 11 2005.
- N. D. Lawrence. Learning for larger datasets with the Gaussian process latent variable model. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*, pages 243–250, San Juan, Puerto Rico, 21-24 March 2007. Omnipress. [PDF].
- T. K. Leen, T. G. Dietterich, and V. Tresp, editors. *Advances in Neural Information Processing Systems*, volume 13, Cambridge, MA, 2001. MIT Press.
- D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, U.K., 2003. [Google Books] .
- R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996. Lecture Notes in Statistics 118.
- 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.
- C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. [Google Books] .

References II

- M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.
- A. J. Smola and P. L. Bartlett. Sparse greedy Gaussian process regression. In Leen et al. (2001), pages 619–625.
- 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.
- M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999. [[PDF](#)]. [[DOI](#)].
- M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16–18 April 2009. JMLR W&CP 5.
- M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. In Y. W. Teh and D. M. Titterton, editors, *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, volume 9, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13–16 May 2010. JMLR W&CP 9. [[PDF](#)].
- J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008. ISSN 0162-8828. [[DOI](#)].
- C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In Leen et al. (2001), pages 682–688.