DD2434 - Advanced Machine Learning

Representation Learning

Carl Henrik Ek
{chek}@csc.kth.se

Royal Institute of Technology

November 11th, 2015

Last Lecture

- Gaussian Processes
  - Prior over the space of functions
  - Posterior
  - Marginal Likelihood
  - Learning

Regression

Regression model,

$$\mathbf{y}_i = f(\mathbf{x}_i) + \boldsymbol{\epsilon} \tag{1}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{2}$$

Introduce f_i as *instansiation* of function,

$$f_i = f(\mathbf{x}_i), \tag{3}$$

as a new random variable.

Regression

Model,

$$p(\mathbf{Y}, \mathbf{f}, \mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X})p(\boldsymbol{\theta}) \tag{4}$$

Want to “push” $\mathbf{X}$ through a mapping f of which we are uncertain,

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}), \tag{5}$$

prior over instansiations of function.

Gaussian Processes[1]

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\mathbf{X}), k(\mathbf{X}, \mathbf{X})) \tag{6}$$

Defenition

A Gaussian Process is an infinite collection of random variables who **any** subset is jointly gaussian. The process is specified by a mean function $\mu(\cdot)$ and a co-variance function $k(\cdot, \cdot)$

$$f \sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot)) \tag{7}$$

[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]

$$p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\mathbf{X}), k(\mathbf{X},\mathbf{X})) \tag{8}$$

$$\mathbf{y}_i = f_i + \boldsymbol{\epsilon} \tag{9}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{10}$$

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta}) \mathrm{d}f \tag{11}$$

Connection to Distribution

$\mathcal{GP}$ is infinite, but we only observe finite amount of data. This means conditioning on a subset of the data, the $\mathcal{GP}$ is a just a Gaussian distribution, which is self-conjugate.

[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]

The mean function

- Function of only the input location
- What do I expect the function value to be **only** accounting for the input location
- We will assume this to be constant

The co-variance function

- Function of **two** input locations
- How should the information from other locations with **known** function value observations effect my estimate
- Encodes the behavior of the function

[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]

The mean function

- Function of only the input location
- What do I expect the function value to be **only** accounting for the input location
- We will assume this to be constant

The co-variance function

- Function of **two** input locations
- How should the information from other locations with **known** function value observations effect my estimate
- Encodes the behavior of the function

[1]Bishop 2006, p. 6.4.2

Gaussian Processes[1]

The mean function

- Function of only the input location
- What do I expect the function value to be **only** accounting for the input location
- We will assume this to be constant

The co-variance function

- Function of **two** input locations
- How should the information from other locations with **known** function value observations effect my estimate
- Encodes the behavior of the function

[1]Bishop 2006, p. 6.4.2

Gaussian Processes[1]

The Prior

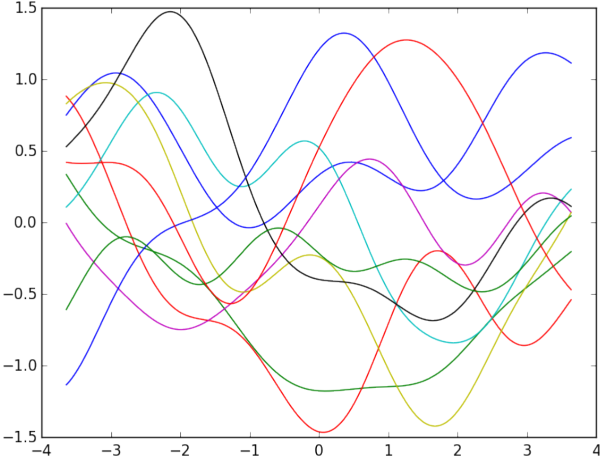
$$p(f|\mathbf{X},\boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}')) \tag{12}$$

$$\mu(\mathbf{x}) = \mathbf{0} \tag{13}$$

$$k(\mathbf{x}_i,\mathbf{x}_j) = \sigma^2 e^{-\frac{1}{2\ell^2}(\mathbf{x}_i-\mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i-\mathbf{x}_j)} \tag{14}$$

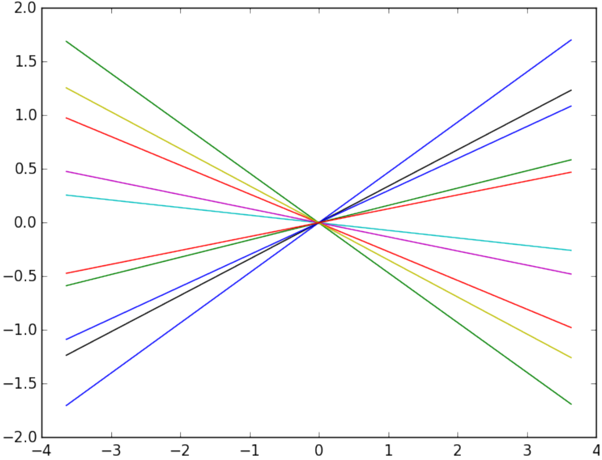
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



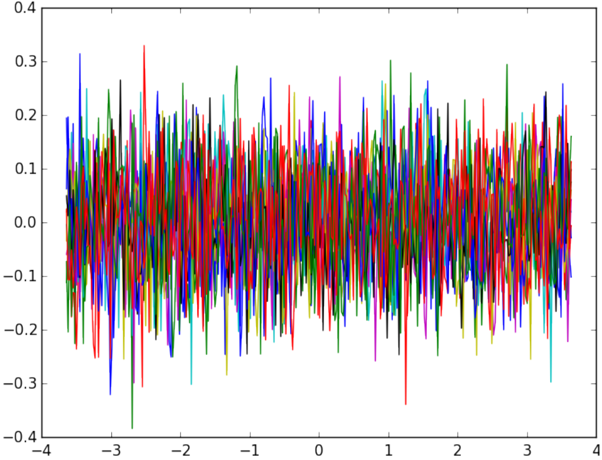
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



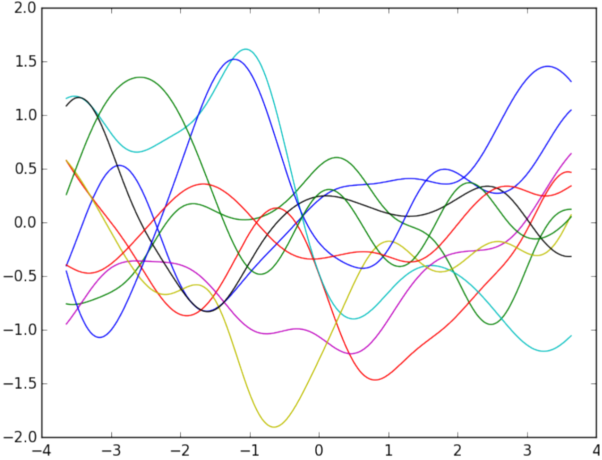
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



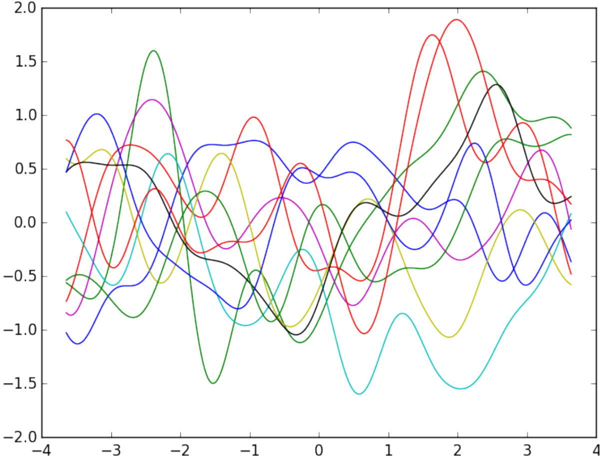
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



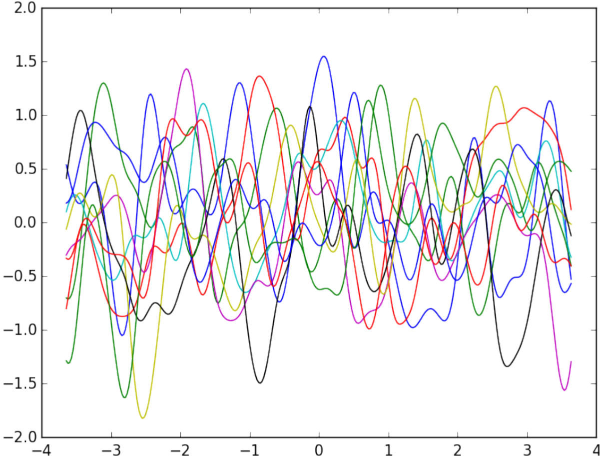
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



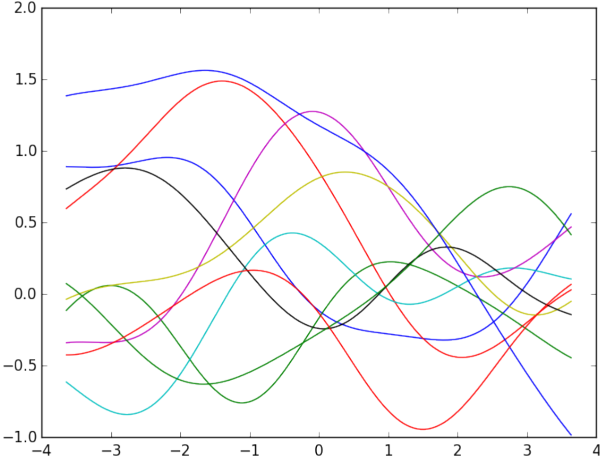
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



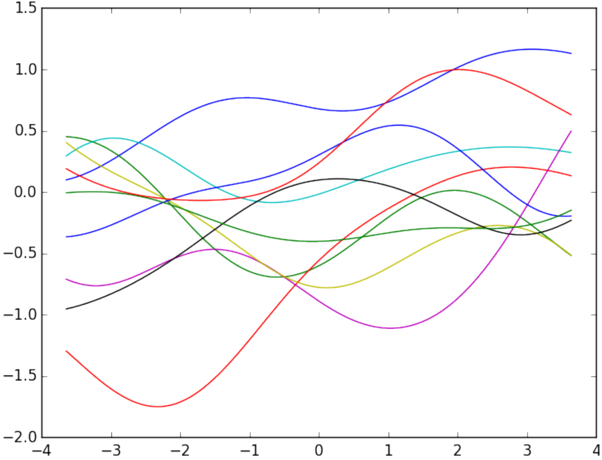
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



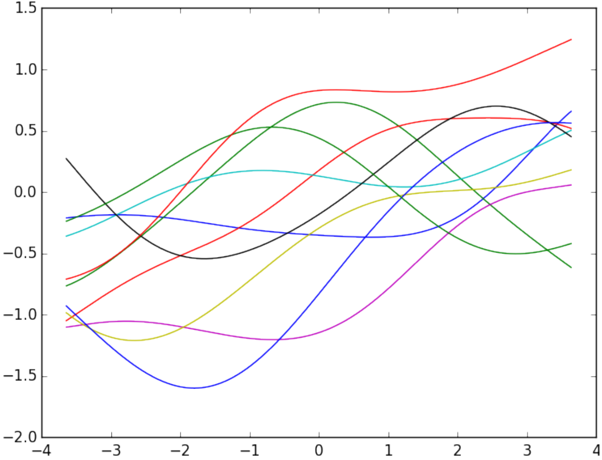
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



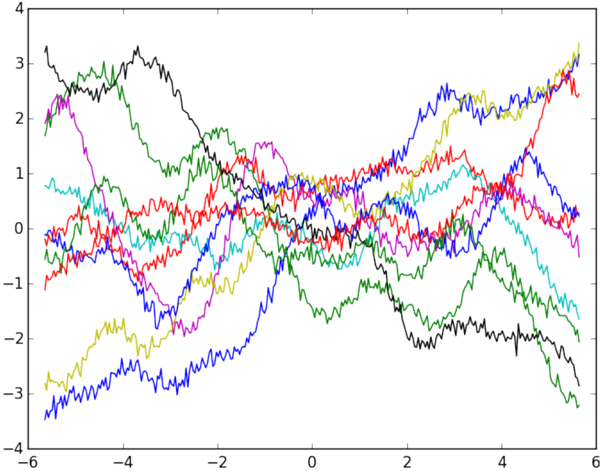
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



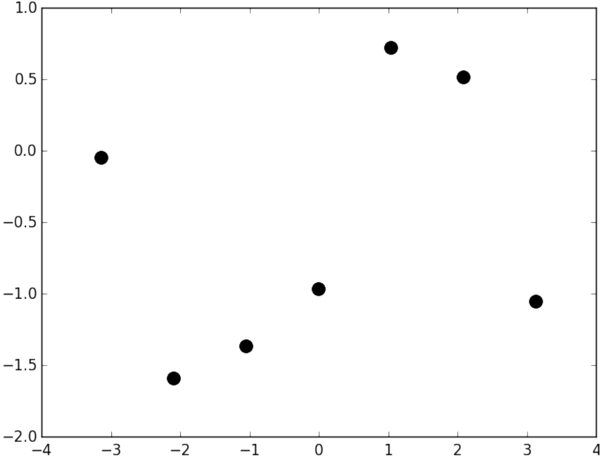
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]

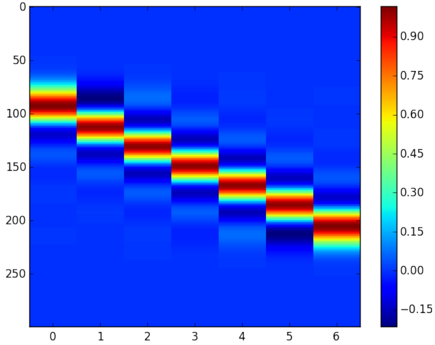
The (predictive) Posterior

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X},\mathbf{X}) & k(\mathbf{X},\mathbf{x}_*) \\ k(\mathbf{x}_*,\mathbf{X}) & k(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix} \right) \tag{15}$$

$$\begin{aligned} p(f_*|\mathbf{x}_*,\mathbf{X},\mathbf{f},\boldsymbol{\theta}) = \mathcal{N}(& k(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}} K(\mathbf{X},\mathbf{X})^{-1}\mathbf{f}, \\ & k(\mathbf{x}_*,\mathbf{x}_*) - k(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}} K(\mathbf{X},\mathbf{X})^{-1} K(\mathbf{X},\mathbf{x}_*)) \end{aligned} \tag{16}$$

[1] Bishop 2006, p. 6.4.2

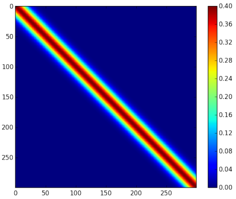
Gaussian Processes[1]



$$k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f} \tag{17}$$

[1] Bishop 2006, p. 6.4.2

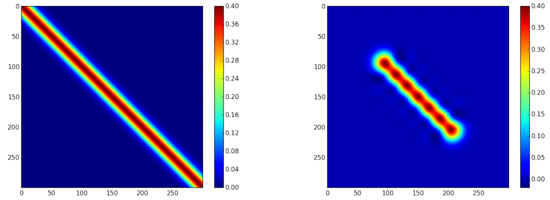
Gaussian Processes[1]



$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}, \mathbf{x}_*) \tag{18}$$

[1] Bishop 2006, p. 6.4.2

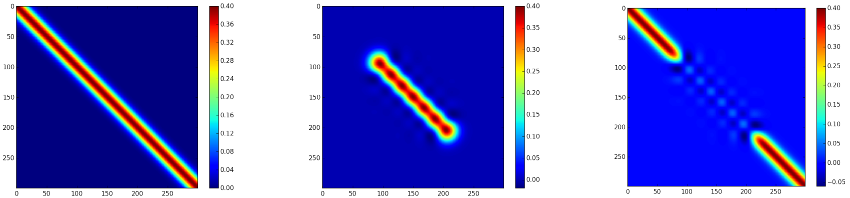
Gaussian Processes[1]



$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}, \mathbf{x}_*) \tag{19}$$

[1] Bishop 2006, p. 6.4.2

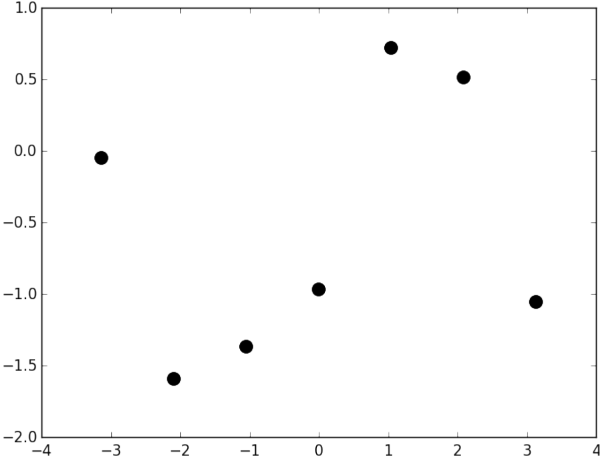
Gaussian Processes[1]



$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}, \mathbf{x}_*) \tag{20}$$

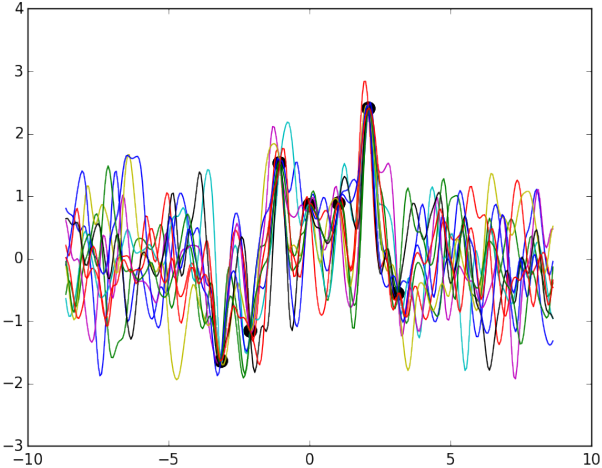
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



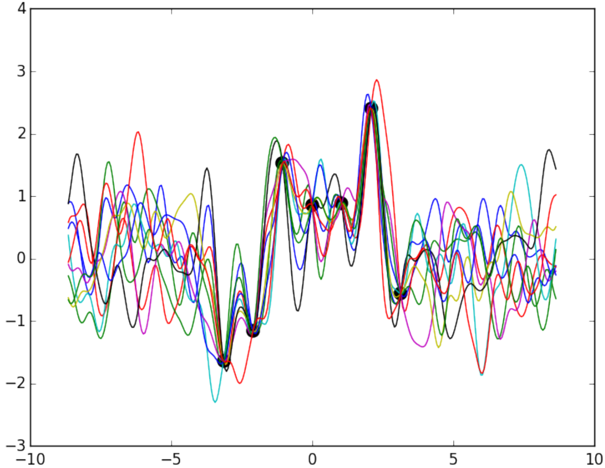
[1]Bishop 2006, p. 6.4.2

Gaussian Processes[1]



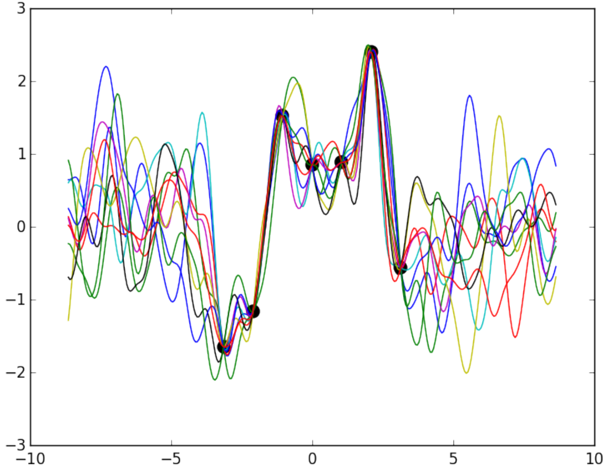
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



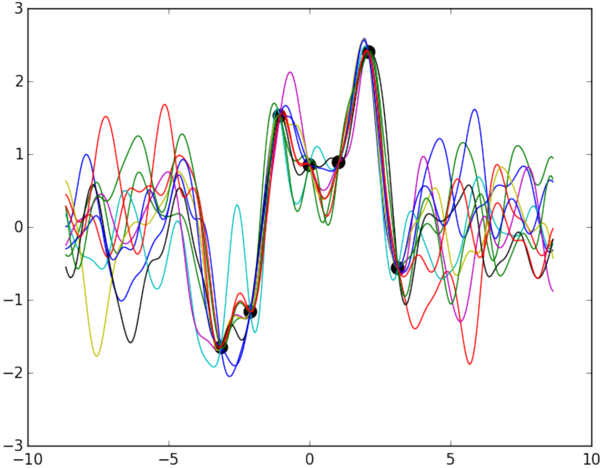
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



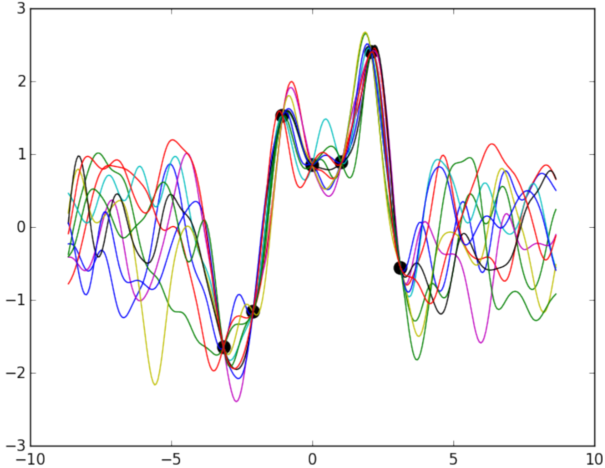
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



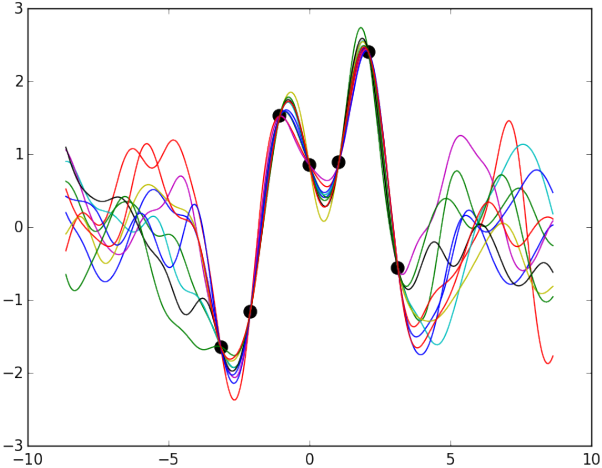
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



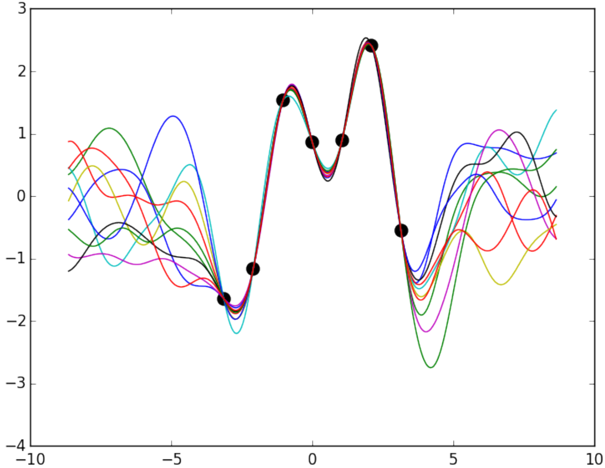
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



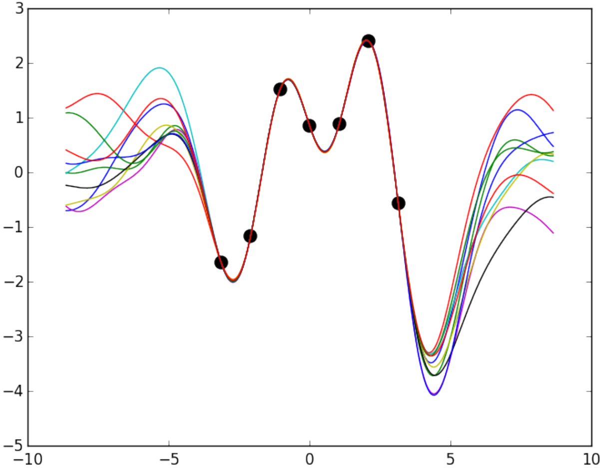
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



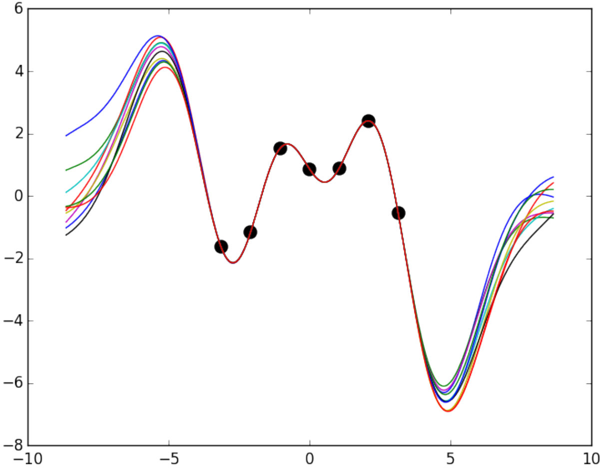
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



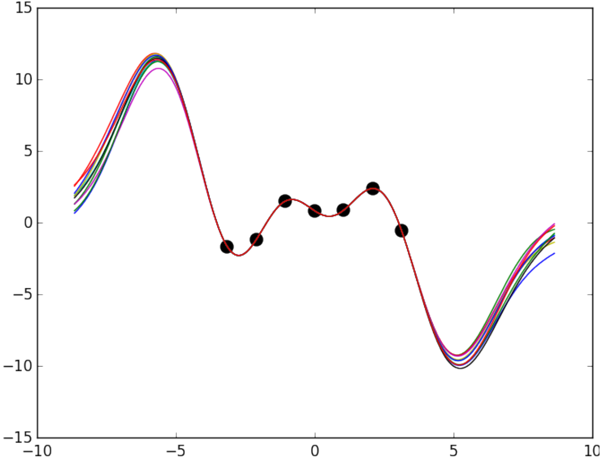
[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



[1] Bishop 2006, p. 6.4.2

Gaussian Processes[1]



[1]Bishop 2006, p. 6.4.2

Learning in Gaussian Processes[2]

Hyper-parameters

- Prior has parameters
  - referred to as *hyper*-parameters
  - SE have lengthscale and variance
- Learning in $\mathcal{GP}$s implies inferring hyper-parameters from the model

[2]Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\mathrm{d}f \tag{21}$$

Marginal Likelihood

- We are not interested in $\mathbf{f}$ directly
- Marginalise out $\mathbf{f}$!

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\mathrm{d}f \tag{22}$$

Marginal Likelihood

- We are not interested in $\mathbf{f}$ directly
- Marginalise out $\mathbf{f}$!

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})\mathrm{d}f \tag{23}$$

Marginal Likelihood

- We are not interested in $\mathbf{f}$ directly
- Marginalise out $\mathbf{f}$!

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

$$\text{argmax}_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \text{argmin}_{\boldsymbol{\theta}} - \log\left(p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})\right) = \text{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{24}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi) \tag{25}$$

Type-II Maximum Likelihood

- Can be minimised using gradient based methods
- Data-fit: $\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$
- Complexity: $\frac{1}{2}\log|\mathbf{K}|$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

$$\text{argmax}_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \text{argmin}_{\boldsymbol{\theta}} - \log\left(p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})\right) = \text{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{26}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi) \tag{27}$$

Type-II Maximum Likelihood

- Can be minimised using gradient based methods
- Data-fit: $\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$
- Complexity: $\frac{1}{2}\log|\mathbf{K}|$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

$$\text{argmax}_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \text{argmin}_{\boldsymbol{\theta}} - \log\left(p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})\right) = \text{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{28}$$

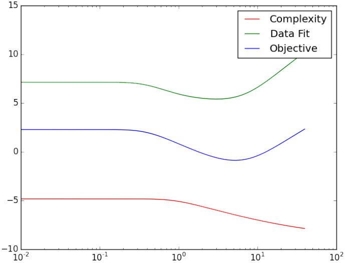
$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi) \tag{29}$$

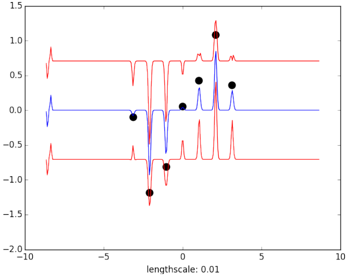
Type-II Maximum Likelihood

- Can be minimised using gradient based methods
- Data-fit: $\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$
- Complexity: $\frac{1}{2}\log|\mathbf{K}|$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

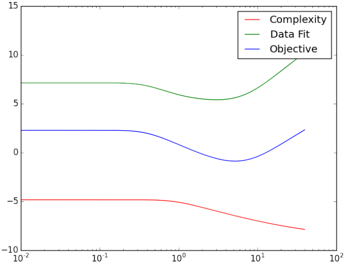


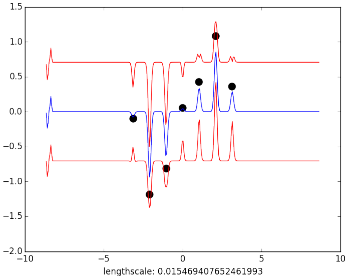


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

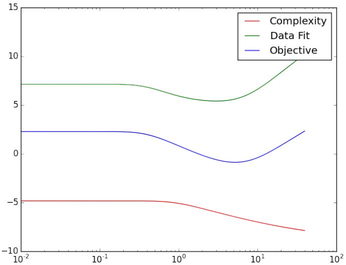


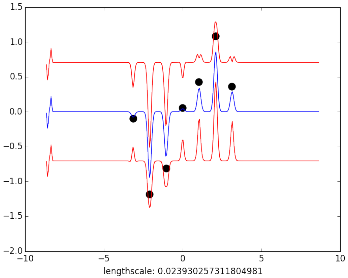


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

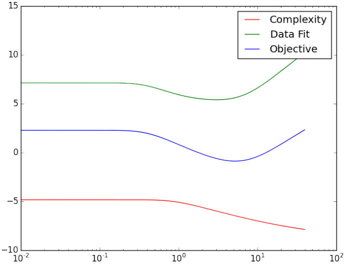


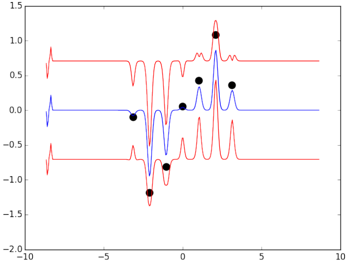


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

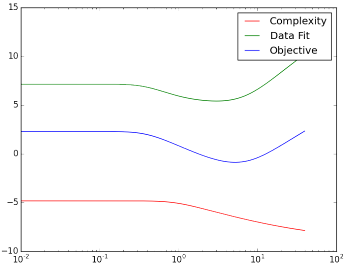


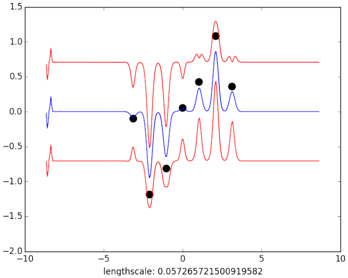


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

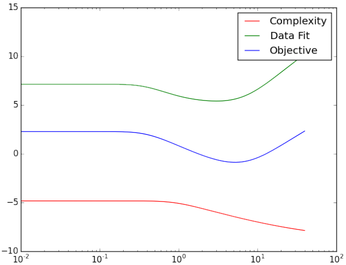


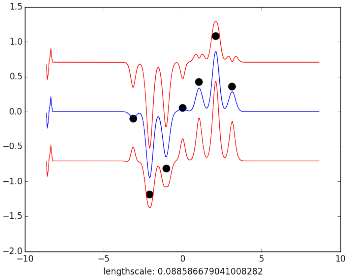


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

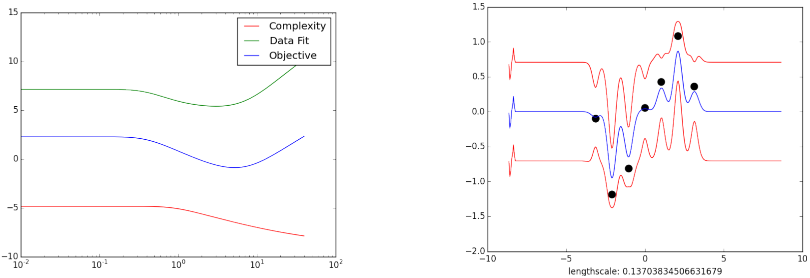




$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

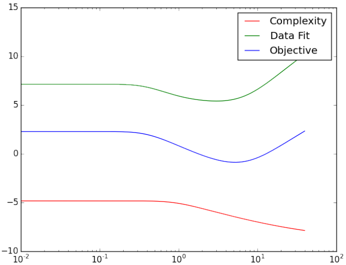
Learning in Gaussian Processes[2]

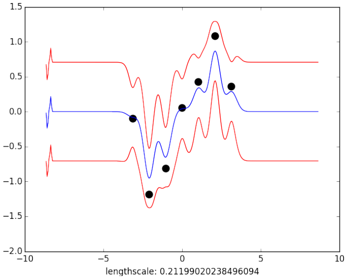


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

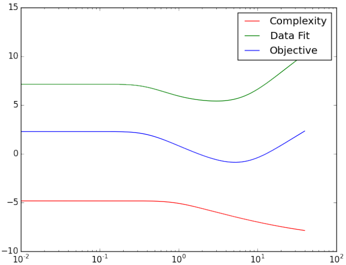


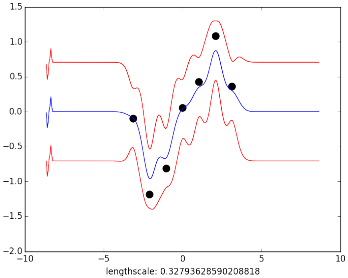


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

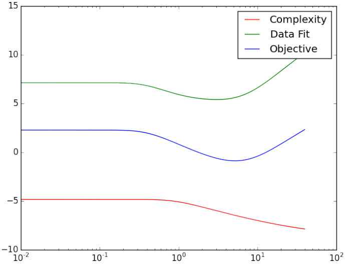


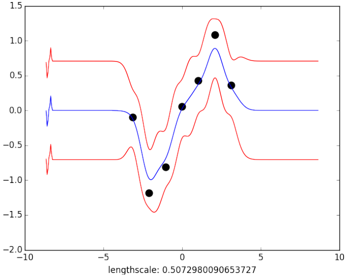


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

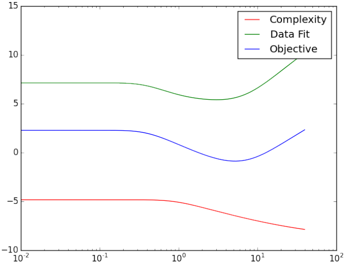


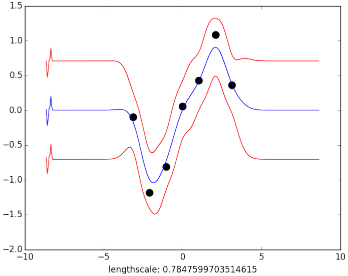


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

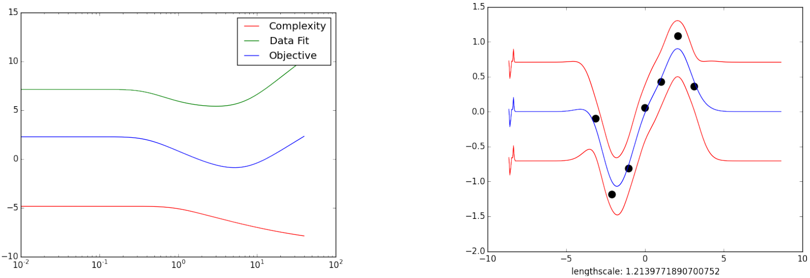




$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

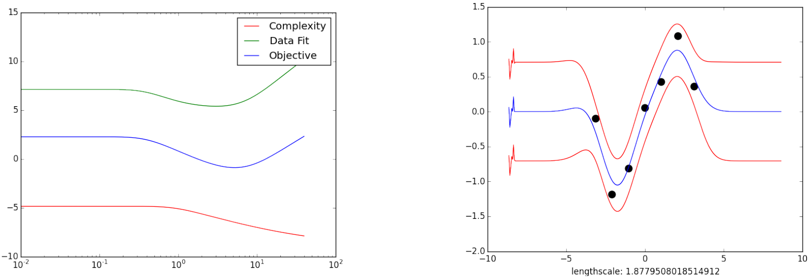
Learning in Gaussian Processes[2]



$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

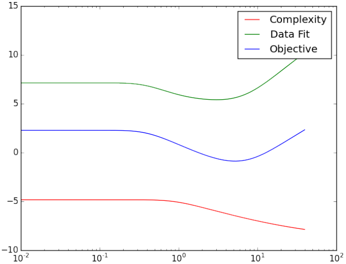
Learning in Gaussian Processes[2]

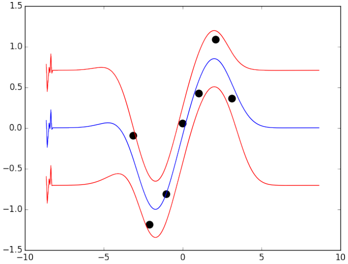


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

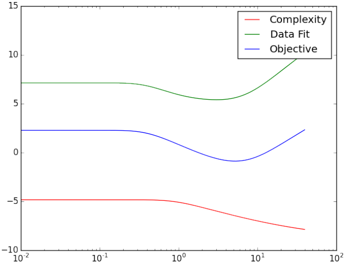


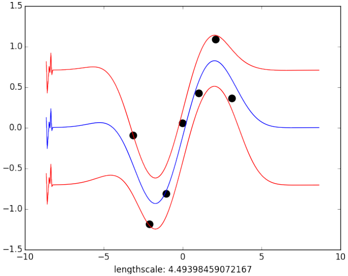


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2]Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

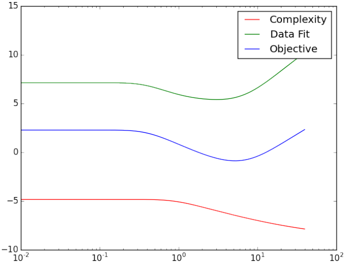


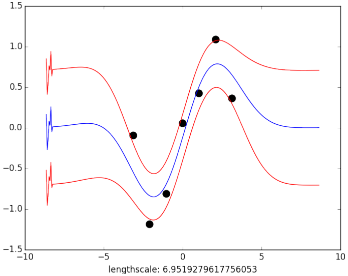


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

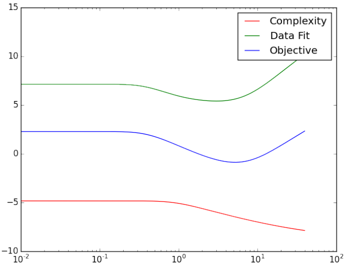


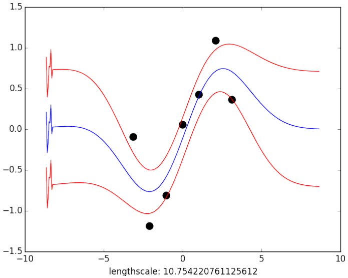


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

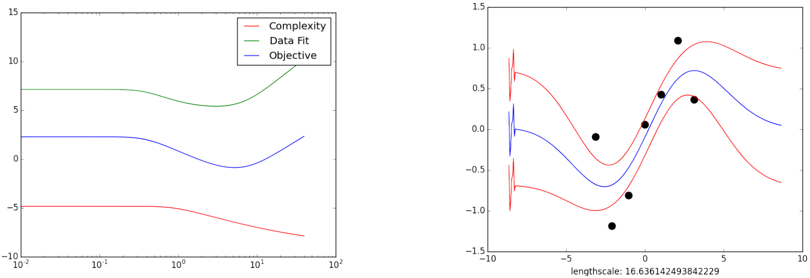




$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

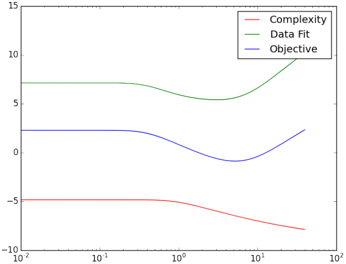
Learning in Gaussian Processes[2]

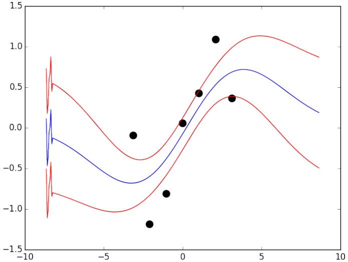


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

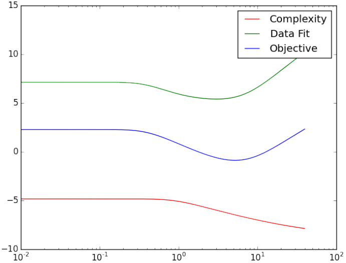


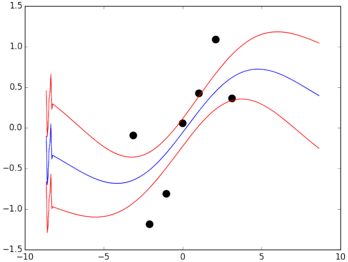


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]

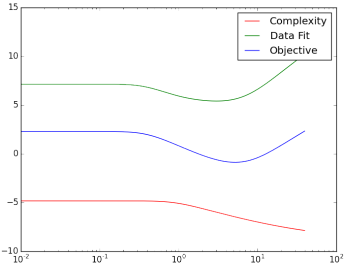


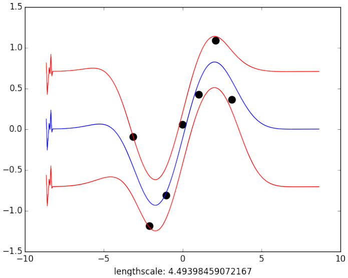


$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Learning in Gaussian Processes[2]





$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Bishop 2006, p. 6.4.3

Introduction

Recap

Representation Learning

Graphical Models[3]

$$p(x, y, z) = p(y|x, z)p(x)p(z)$$

$$p(x, y, z) = p(y|z)p(x|z)p(z)$$

$$p(\{x_i\}_{i=1}^{N}) = \prod_{i=1}^{N} p(x_i|\mathrm{pa}_i) \tag{30}$$

[3] Bishop 2006, pp. 8.0, 8.1.

Latent Variable Models[4]

Machine Learning

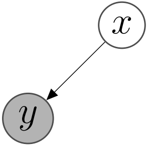
- What is our task?
- $p(y)$
- Unobservables
  - Latent variables
- Explaining away

[4] Bishop 2006, p. 364.

Latent Variable Models[4]

Machine Learning

- What is our task?
- $p(y)$
- Unobservables
  - Latent variables
- Explaining away

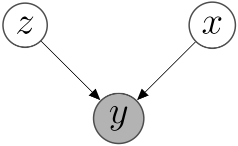


[4] Bishop 2006, p. 364.

Latent Variable Models[4]

Machine Learning

- What is our task?
- $p(y)$
- Unobservables
  - Latent variables
- Explaining away

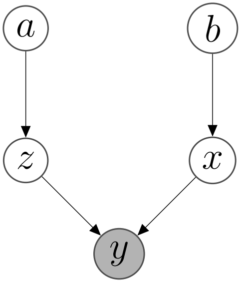


[4] Bishop 2006, p. 364.

Latent Variable Models[4]

Machine Learning

- What is our task?
- $p(y)$
- Unobservables
  - Latent variables
- Explaining away

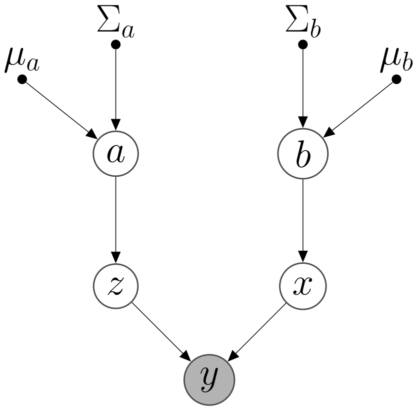


[4] Bishop 2006, p. 364.

Latent Variable Models[4]

Machine Learning

- What is our task?
- $p(y)$
- Unobservables
  - Latent variables
- Explaining away

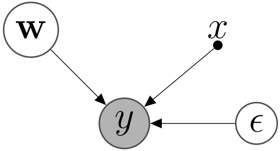


[4] Bishop 2006, p. 364.

Latent Variable Models[4]

Machine Learning

- What is our task?
- $p(y)$
- Unobservables
  - Latent variables
- Explaining away



$$\mathbf{y}_i = \mathbf{w}\mathbf{x}_i + \boldsymbol{\epsilon} \quad (31)$$

[4] Bishop 2006, p. 364.

Latent Variable Models[4]

Latent Variables[a]

[a] Bishop 2006, p. 366.

"The primary role of the latent variable is to allow a complicated distribution over the observed variables be represented in terms of a model constructed from a simpler (typically exponential family) conditional distribution."

[4] Bishop 2006, p. 364.

Latent Variable Models[4]

[4]Bishop 2006, p. 364.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5]Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5]Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5]Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5]Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

[5] Bishop 2006, p. 8.1.2.

Generative Models[5]

- We want to model this data, i.e. $p(y)$
- $\mathbf{y}_i \in \mathbb{R}^{1400\times931\times3=3910200}$
- Latent variable representation
- Conditional distribution parametrised by latent variable

[5]Bishop 2006, p. 8.1.2.

Generative Models[5]

```
t = np.sort(5*np.random.randn(30))
for i in range(0,len(t)):
    pos[0] = int(radius*np.sin(t[i])+offset[0]/2)
    pos[1] = int(radius*np.cos(t[i])+offset[1]/2)
    background_draw.paste(stella,pos,stella)
    background_draw.save('lvm_'+str(i)+'.png')
    background_draw = copy.deepcopy(background)
```

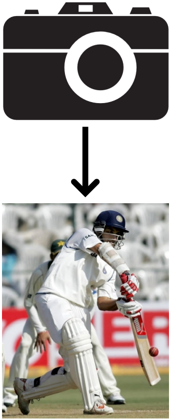
$$p(\mathbf{Y}|\mathbf{X}) \tag{32}$$

[5] Bishop 2006, p. 8.1.2.

Sensory Data

What we are doing

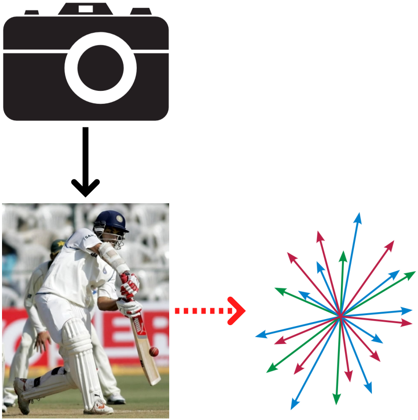
- Sensory representation
  - Capturing process
  - Pixels, Waveforms
- Degrees of freedom and dimensionality



Sensory Data

What we are doing

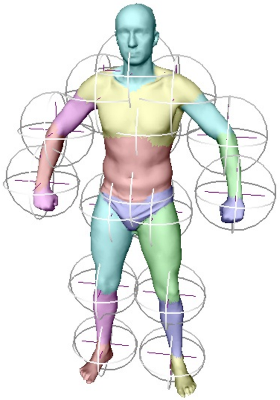
- Sensory representation
  - Capturing process
  - Pixels, Waveforms
- Degrees of freedom and dimensionality



Sensory Data

What we are doing

- Sensory representation
  - Capturing process
  - Pixels, Waveforms
- Degrees of freedom and dimensionality



Outline

- Re-visit PCA
- PCA as a Latent Variable Model
- Factor Analysis
- Example of Intractability

Re-visit: Principal Component Analysis

- Given data $\mathbf{Y}$ project to directions of maximum variance
- Provides no uncertainty

$$\text{argmax}_{\mathbf{v}} \sigma(\mathbf{Y}\mathbf{v}, \mathbf{Y}\mathbf{v}) \tag{33}$$

$$\mathbf{v}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{v} \tag{34}$$

$$\text{subject to: } :\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1 \tag{35}$$

Latent Variable Models[6]

$$p(\mathbf{Y}) \tag{36}$$

- We have observed some data $\mathbf{Y}$
- Lets assume that $\mathbf{Y} \in \mathbb{R}^{N \times d}$ have been generated from $\mathbf{X} \in \mathbb{R}^{N \times q}$
- $\mathbf{X}$ - latent variable
- f - generative mapping

[6] Bishop 2006, p. 8.1.2.

Latent Variable Models[6]

$$p(\mathbf{Y}|f, \mathbf{X}) \tag{37}$$

$$\mathbf{f} : \mathbf{X} \rightarrow \mathbf{Y} \tag{38}$$

- We have observed some data $\mathbf{Y}$
- Lets assume that $\mathbf{Y} \in \mathbb{R}^{N \times d}$ have been generated from $\mathbf{X} \in \mathbb{R}^{N \times q}$
- $\mathbf{X}$ - latent variable
- f - generative mapping

[6] Bishop 2006, p. 8.1.2.

Latent Variable Models[6]

$$p(\mathbf{Y}|f, \mathbf{X}) \tag{39}$$

$$\mathbf{f} : \mathbf{X} \rightarrow \mathbf{Y} \tag{40}$$

- We have observed some data $\mathbf{Y}$
- Lets assume that $\mathbf{Y} \in \mathbb{R}^{N \times d}$ have been generated from $\mathbf{X} \in \mathbb{R}^{N \times q}$
- $\mathbf{X}$ - latent variable
- f - generative mapping

[6] Bishop 2006, p. 8.1.2.

Latent Variable Models[6]

$$p(\mathbf{Y}|f, \mathbf{X}) \tag{41}$$

$$\mathbf{f} : \mathbf{X} \rightarrow \mathbf{Y} \tag{42}$$

- We have observed some data $\mathbf{Y}$
- Lets assume that $\mathbf{Y} \in \mathbb{R}^{N \times d}$ have been generated from $\mathbf{X} \in \mathbb{R}^{N \times q}$
- $\mathbf{X}$ - latent variable
- f - generative mapping

[6] Bishop 2006, p. 8.1.2.

Linear Latent Variable Models[7]

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) \tag{43}$$

Regression

- Regression without inputs?
- Solve the task: Given some data
  - a representation of this data
  - and a mapping that have generated the

[7] Bishop 2006, p. 12.2.0.

WTF?

The strength of Priors

- Encodes prior belief
- This can also be seen as a preference
  - Given several perfectly valid solutions which one do i prefer
  - Regularises solution space
- Latent variable models what do we prefer?

Factor Analysis[8]

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon \tag{44}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}) \tag{45}$$

- Assume the generating mapping to be linear
- Assume $\mathbf{\Psi}$ diagonal
- For regression we assumed that we knew the inputs $\mathbf{X}$
- Now we do not

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon \tag{46}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}) \tag{47}$$

- Assume the generating mapping to be linear
- Assume $\mathbf{\Psi}$ diagonal
- For regression we assumed that we knew the inputs $\mathbf{X}$
- Now we do not

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon \tag{48}$$

$$p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{W}\mathbf{x}_i, \boldsymbol{\Psi}) \tag{49}$$

$$p(\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{0}) \tag{50}$$

- Assume the generating mapping to be linear
- Assume $\boldsymbol{\Psi}$ diagonal
- For regression we assumed that we knew the inputs $\mathbf{X}$
- Now we do not $\Rightarrow$ specify a prior

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$p(\mathbf{y}_i|\boldsymbol{\theta}) = \int p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})p(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i = \int \mathcal{N}(\mathbf{W}\mathbf{x}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$
$$= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}})$$

- $\mathbf{X}$ and $\mathbf{W}$ are related
- Integrate out $\mathbf{X}$
  - pick $\boldsymbol{\mu}_0 = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- Low dimensional density model of $\mathbf{Y}$
  - rank of $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ dimensionality of $\mathbf{X}$

[8]Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$
\begin{aligned}
p(\mathbf{y}_i|\boldsymbol{\theta}) &= \int p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})p(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i = \int \mathcal{N}(\mathbf{W}\mathbf{x}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
&= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}}) \\
&= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}) \qquad (51)
\end{aligned}
$$

- $\mathbf{X}$ and $\mathbf{W}$ are related
- Integrate out $\mathbf{X}$
  - pick $\boldsymbol{\mu}_0 = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- Low dimensional density model of $\mathbf{Y}$
  - rank of $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ dimensionality of $\mathbf{X}$

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$
\begin{aligned}
p(\mathbf{y}_i|\boldsymbol{\theta}) &= \int p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta})p(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i = \int \mathcal{N}(\mathbf{W}\mathbf{x}_i+\boldsymbol{\mu},\boldsymbol{\Psi})\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\
&= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0+\boldsymbol{\mu},\boldsymbol{\Psi}+\mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}}) \\
&= \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Psi}+\mathbf{W}\mathbf{W}^{\mathrm{T}}) \qquad (52)
\end{aligned}
$$

- $\mathbf{X}$ and $\mathbf{W}$ are related
- Integrate out $\mathbf{X}$
  - pick $\boldsymbol{\mu}_0 = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- Low dimensional density model of $\mathbf{Y}$
  - rank of $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ dimensionality of $\mathbf{X}$

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$\tilde{\mathbf{W}} = \mathbf{W}\mathbf{R} \tag{53}$$

$$p(\mathbf{y}_i|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}) \tag{54}$$

$$= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}) \tag{55}$$

$$\tag{56}$$

Identifiability

- The marginal likelihood is invariant to a rotation
  - no unique solution
  - model is the same but interpretation tricky

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

$$\mathbf{W}_{ML} = \text{argmax}_{\mathbf{W}} p(\mathbf{Y}|\boldsymbol{\theta}) \tag{57}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{58}$$

$$\mathbf{W}_{ML} = \mathbf{U}_q (\Lambda - \sigma^2 \mathbf{I})^{\frac{1}{2}} \tag{59}$$

$$\mathbf{S} = \mathbf{U} \Lambda \mathbf{U}^{\mathrm{T}} \tag{60}$$

Probabilistic PCA

- Dimensions of $\mathbf{Y}$ independent given $\mathbf{X}$
  - $\mathbf{W}$ orthogonal matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

Summary

- Factor Analysis is a linear continous latent variable model
- Solution not unique
- PCA is Factor Analysis with two assumptions
  - factor loadings orthogonal $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$
  - noise free case $\epsilon = \lim_{\sigma^2 \to 0} \sigma^2 \mathbf{I}$
- PCA is incredibly useful but its important to know what you are assuming, the probabilistic formulation allows you to do just that

[8] Bishop 2006, p. 12.2.4.

Factor Analysis[8]

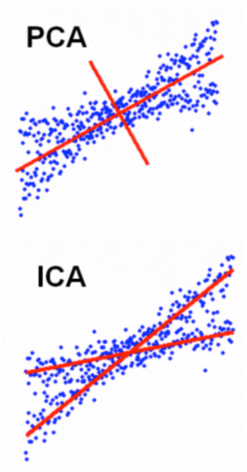
Summary

- Factor Analysis is a linear continous latent variable model
- Solution not unique
- PCA is Factor Analysis with two assumptions
  - factor loadings orthogonal $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$
  - noise free case $\epsilon = \lim_{\sigma^2 \to 0} \sigma^2 \mathbf{I}$
- PCA is incredibly useful but its important to know what you are assuming, the probabilistic formulation allows you to do just that

[8] Bishop 2006, p. 12.2.4.

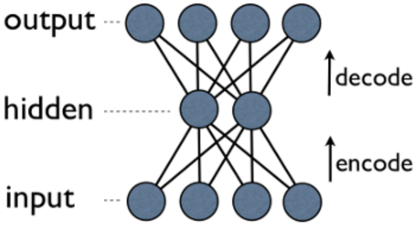
Other Factor Analysis Models

- Independent Component Analysis
  - $p(\mathbf{X}) = \prod_i^q p(x_i)$
  - Coctail party problem
- Auto-associative models
  - $p(\mathbf{X}|\mathbf{Y}) = \prod_i^N p(\mathbf{x}_i|\mathbf{y}_i)$
  - strange
- Lots and lots of different models differing by prior



Other Factor Analysis Models

- Independent Component Analysis
  - $p(\mathbf{X}) = \prod_i^q p(x_i)$
  - Coctail party problem
- Auto-associative models
  - $p(\mathbf{X}|\mathbf{Y}) = \prod_i^N p(\mathbf{x}_i|\mathbf{y}_i)$
  - strange
- Lots and lots of different models differing by prior



Other Factor Analysis Models

- Independent Component Analysis
  - $p(\mathbf{X}) = \prod_i^q p(x_i)$
  - Coctail party problem
- Auto-associative models
  - $p(\mathbf{X}|\mathbf{Y}) = \prod_i^N p(\mathbf{x}_i|\mathbf{y}_i)$
  - strange
- Lots and lots of different models differing by prior

$$p(\mathbf{X})$$

Assignment

You should now be able to do Task 2.3 and 2.4 in the assignment

Gaussian Process Latent Variable Models

History repeats itself

- In PPCA we assumed no uncertainty in the form of mapping
- We can use $\mathcal{GP}$s over mapping
- Gaussian Process Latent Variable Model [Lawrence 2005]

Gaussian Process Latent Variable Models

History repeats itself

- In PPCA we assumed no uncertainty in the form of mapping
- We can use $\mathcal{GP}$s over mapping
- Gaussian Process Latent Variable Model [Lawrence 2005]

Gaussian Process Latent Variable Models

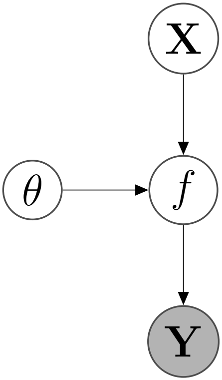
$$p(\mathbf{Y}|\mathbf{f}, \mathbf{X}, \theta) \tag{61}$$

- In PPCA we marginalised out $\mathbf{X}$ and optimised for $\mathbf{W}$
- Not possible for a general $\mathcal{GP}$

Gaussian Process Latent Variable Models

GP-LVM

- General co-variance function (Ex. SE)
- $\mathbf{X}$ appears non-linearly in relation to $\mathbf{Y}$
- Marginalisation of $\mathbf{X}$ intractable



Gaussian Process Latent Variable Models

$$\text{argmax}_{\mathbf{X},\theta} p(\mathbf{Y}|\mathbf{X},\theta)p(\mathbf{X}) \tag{62}$$

$$p(\mathbf{Y}|\mathbf{X},\theta) = \int p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\theta)\mathrm{d}\mathbf{f} \tag{63}$$

$$p(\mathbf{X}) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{64}$$

- $\mathbf{GP}$-prior sufficiently regularises objective
- Need to set dimensionality of $\mathbf{X}$

Demo

Bayesian PCA[9]

- PPCA We have no prior on $\mathbf{W}$
- GP-LVM We have no prior on $\mathbf{X}$
- Likelihood always increases with number of free parameters
  - larger dimensionality always better
- Cross-validation expensive
- Should learn distributions rather than point estimates

[9] Bishop 2006, p. 12.2.3

Bayesian PCA[9]

- PPCA We have no prior on $\mathbf{W}$
- GP-LVM We have no prior on $\mathbf{X}$
- Likelihood always increases with number of free parameters
  - larger dimensionality always better
- Cross-validation expensive
- Should learn distributions rather than point estimates

[9] Bishop 2006, p. 12.2.3

Bayesian PCA[9]

- PPCA We have no prior on $\mathbf{W}$
- GP-LVM We have no prior on $\mathbf{X}$
- Likelihood always increases with number of free parameters
  - larger dimensionality always better
- Cross-validation expensive
- Should learn distributions rather than point estimates

[9] Bishop 2006, p. 12.2.3

Bayesian PCA[9]

- PPCA We have no prior on **W**
- GP-LVM We have no prior on **X**
- Likelihood always increases with number of free parameters
  - larger dimensionality always better
- Cross-validation expensive
- Should learn distributions rather than point estimates

[9] Bishop 2006, p. 12.2.3

Bayesian PCA[9]

$$
\begin{aligned}
p(\mathbf{Y}) &= \int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})p(\mathbf{W})\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{W} \qquad (65)\\
&= \int p(\mathbf{Y}|\mathbf{W})p(\mathbf{W})\mathrm{d}\mathbf{W} = \int \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Psi}+\mathbf{W}\mathbf{W}^{\mathrm{T}})p(\mathbf{W})\mathrm{d}\mathbf{W}\\
&\propto \int \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{Y}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{\Psi}+\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right)\right)p(\mathbf{W})\mathrm{d}\mathbf{W}
\end{aligned}
$$

[9] Bishop 2006, p. 12.2.3

Bayesian PCA[9]

Inference

- Bayesian inference maximise $p(\mathbf{Y})$
  - requires integrating out $\mathbf{W}$ and $\mathbf{X}$
  - intractable in PCA (and most models)
- Elementary functions **not** closed under integration

$$\int e^{e^{ax}} \mathrm{d}x \tag{66}$$

- Solution through approximate inference (tomorrow & Friday & Hedvig)

[9] Bishop 2006, p. 12.2.3

Summary

- Data often represented based on what we can measure
- Implicit representation & degrees of freedom of data
- Generative models
- Priors as preference
- Probabilistic PCA
- Bayesian inference

Next Time

Lecture 4

- November 12th 13-15 V1
- Summary of my part of the course
- Approximative Inference
  - Variational Bayes
- Complete assignment Task 2.3 and 2.4

Next Time

Lecture 4

- November 12th 13-15 V1
- Summary of my part of the course
- Approximative Inference
  - Variational Bayes
- Complete assignment Task 2.3 and 2.4

e.o.f.

References I

Christopher M Bishop. *Pattern recognition and machine learning*. 2006. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf.

Neil D Lawrence. "Probabilistic non-linear principal component analysis with Gaussian process latent variable models". In: *The Journal of Machine Learning Research* 6 (2005), pp. 1783–1816. URL: http://dl.acm.org/citation.cfm?id=1194904.