DD2434 - Advanced Machine Learning

Representation Learning

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

Royal Institute of Technology

November 27th, 2014

Last Lecture

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

Regression

Regression model,

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

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

Introduce f_i as *instansiation* of function,

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

as a new random variable.

Regression

Model,

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

Want to "push" **X** through a mapping *f* of which we are uncertain,

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

prior over instansiations of function.

Gaussian Processes[1]

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

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] Murphy 2012, p. 15.2

Gaussian Processes[1]

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

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

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

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

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] Murphy 2012, p. 15.2

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] Murphy 2012, p. 15.2

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] Murphy 2012, p. 15.2

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] Murphy 2012, p. 15.2

Gaussian Processes[1]

The Prior

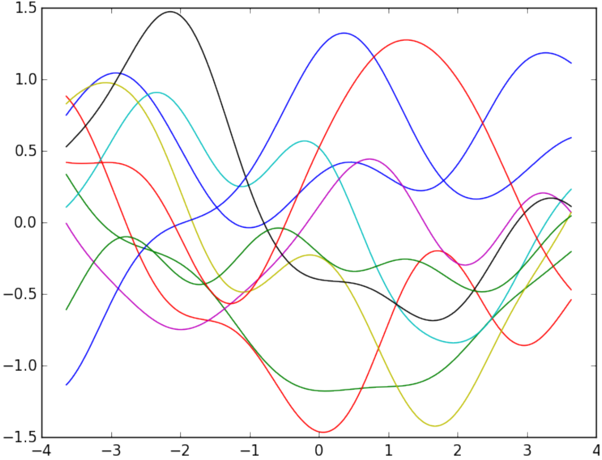
$$p(f|\mathbf{X},\boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}')) \tag{12}$$

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

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

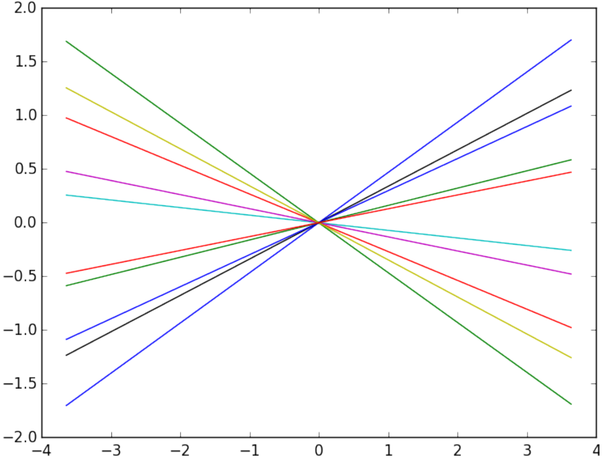
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



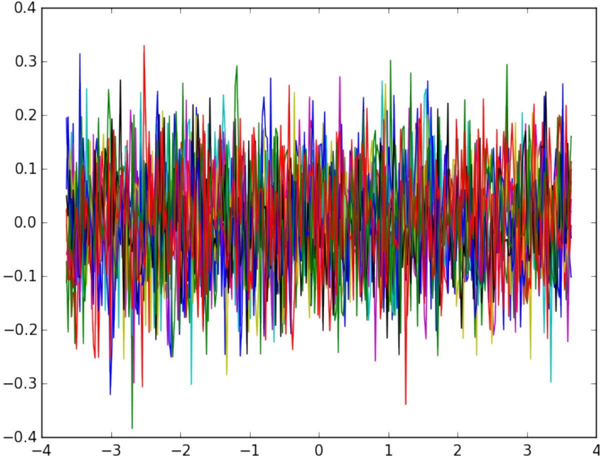
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



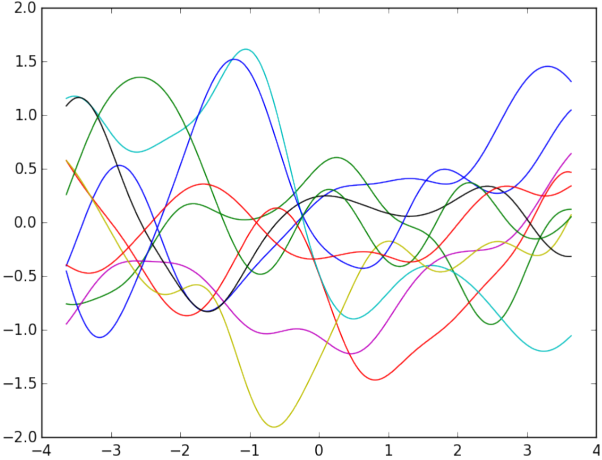
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



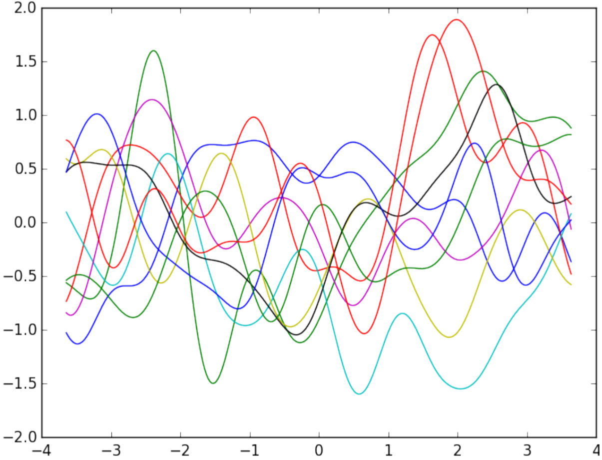
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



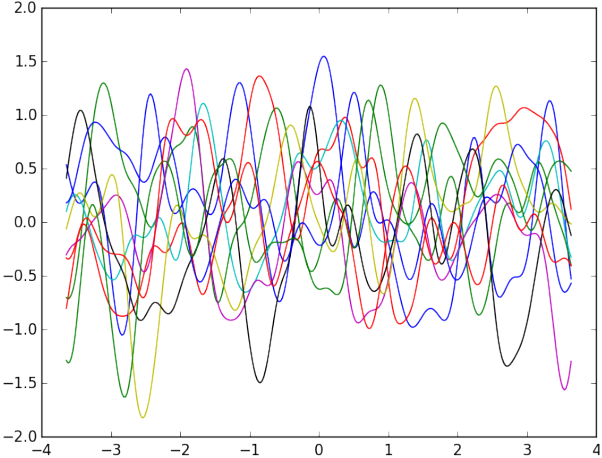
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



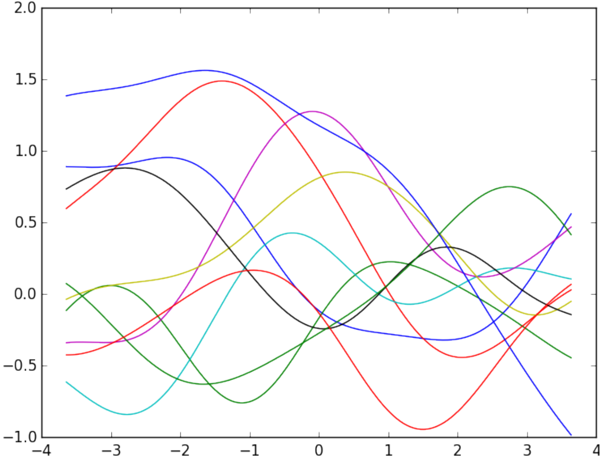
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



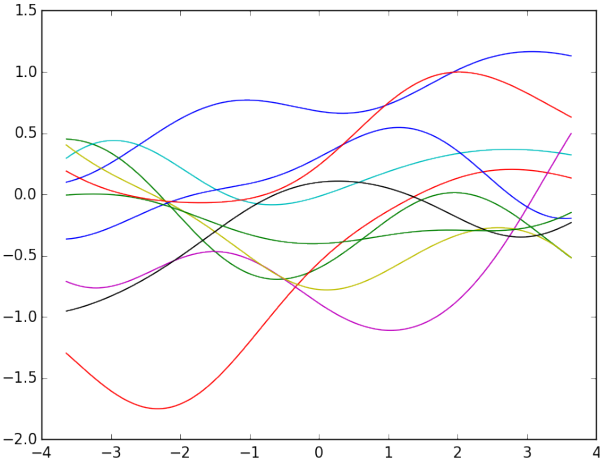
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



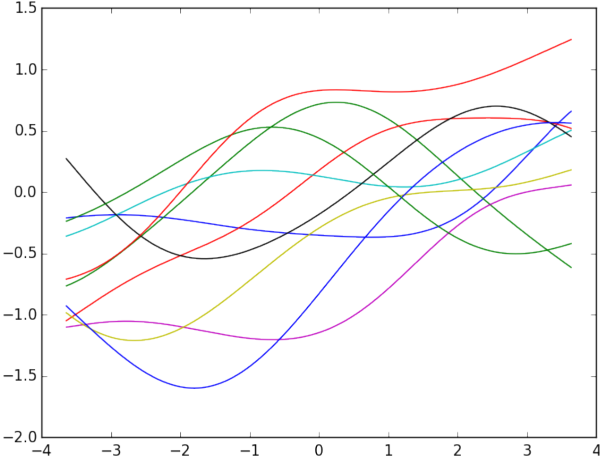
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



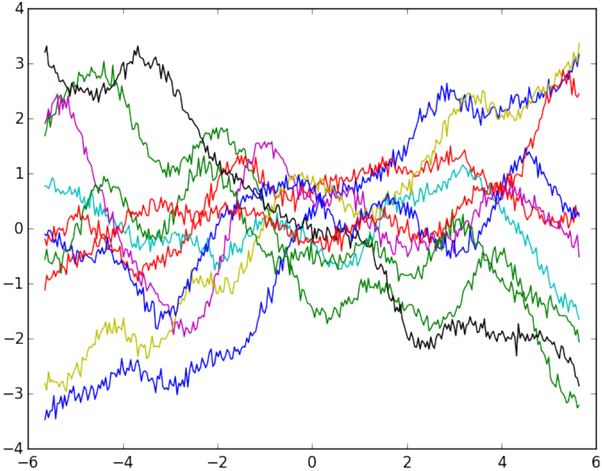
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



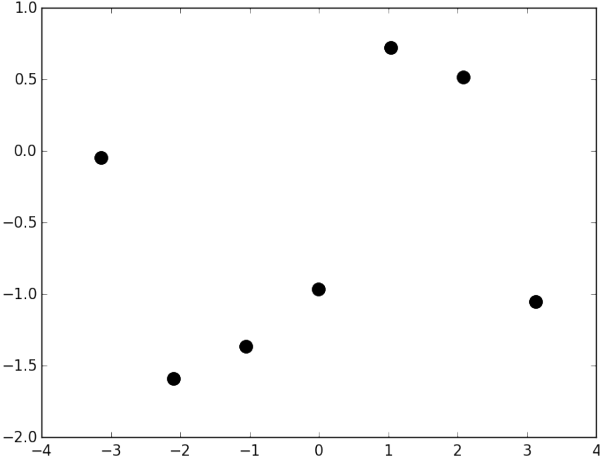
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



[1] Murphy 2012, p. 15.2

Gaussian Processes[1]

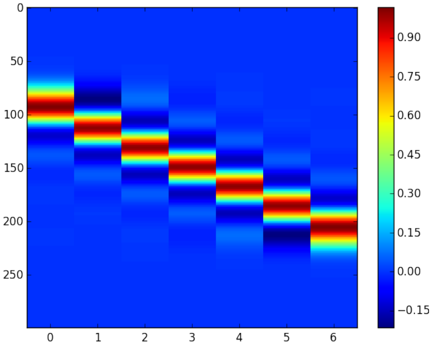
The (predictive) Posterior

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

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

[1] Murphy 2012, p. 15.2

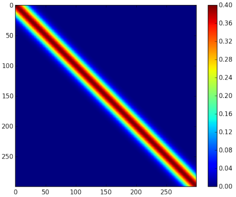
Gaussian Processes[1]



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

[1] Murphy 2012, p. 15.2

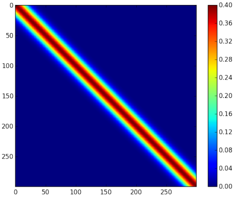
Gaussian Processes[1]

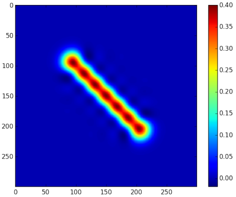


$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}, \mathbf{x}_*) \quad (18)$$

[1] Murphy 2012, p. 15.2

Gaussian Processes[1]

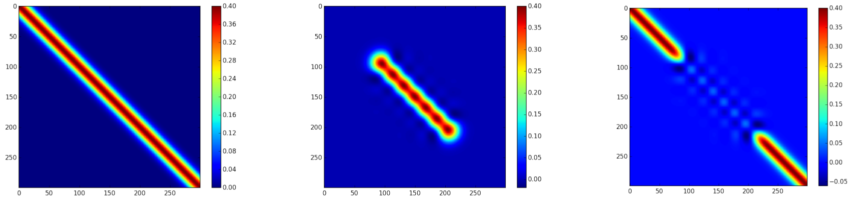




$$k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, \mathbf{X})^{\mathrm{T}} K(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}, \mathbf{x}_*) \tag{19}$$

[1] Murphy 2012, p. 15.2

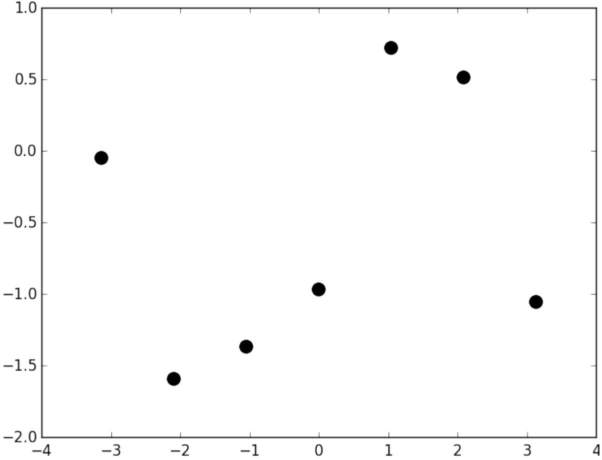
Gaussian Processes[1]



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

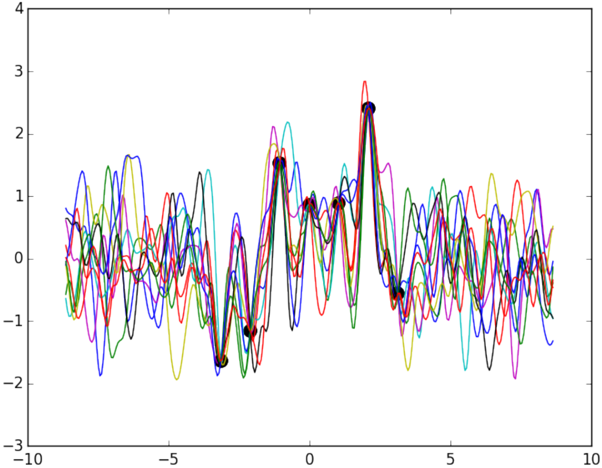
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



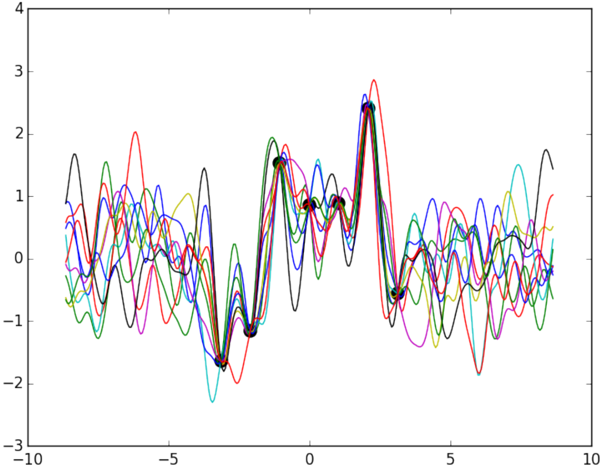
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



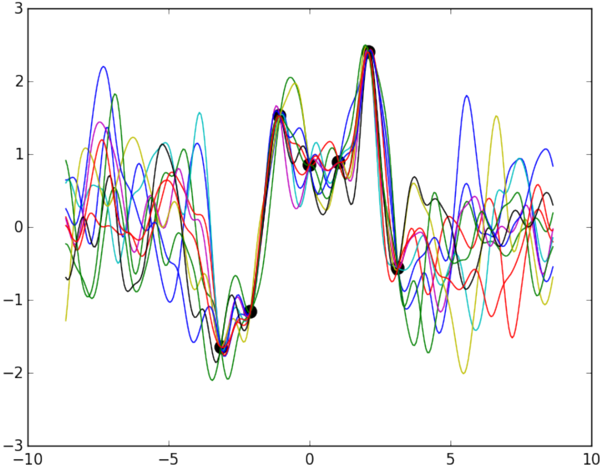
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



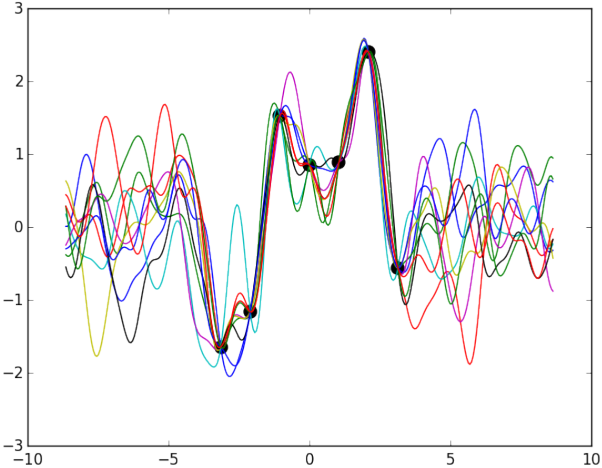
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



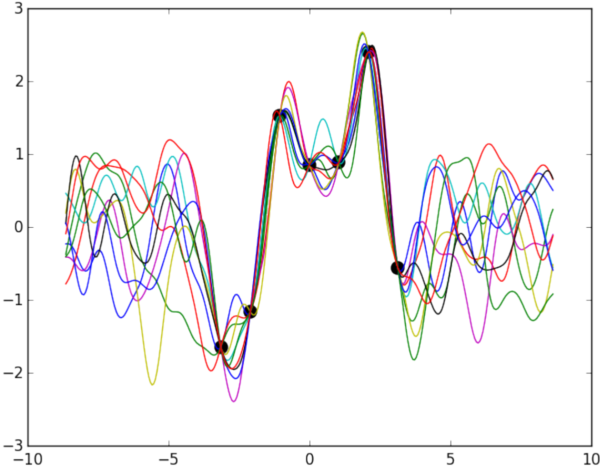
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



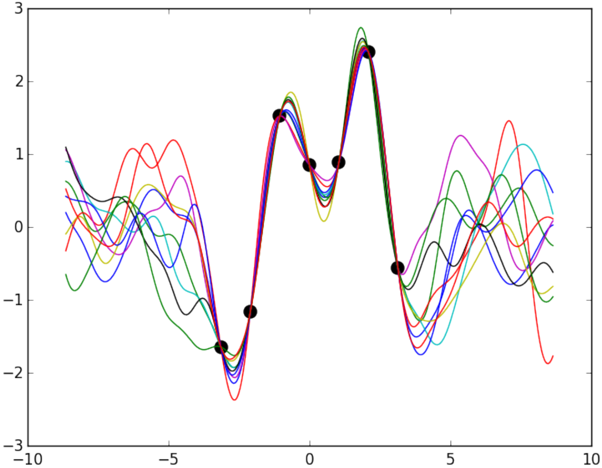
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



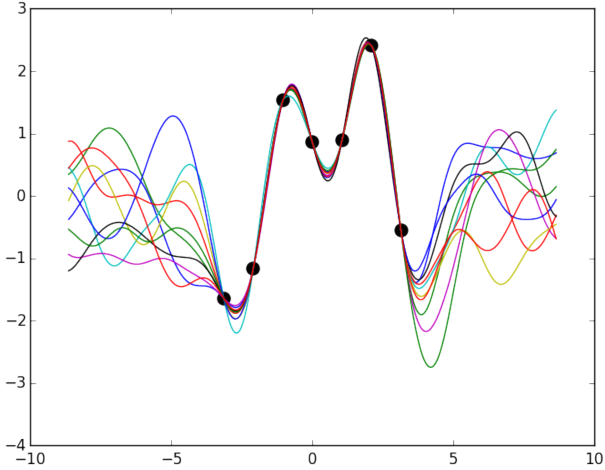
[1]Murphy 2012, p. 15.2

Gaussian Processes[1]



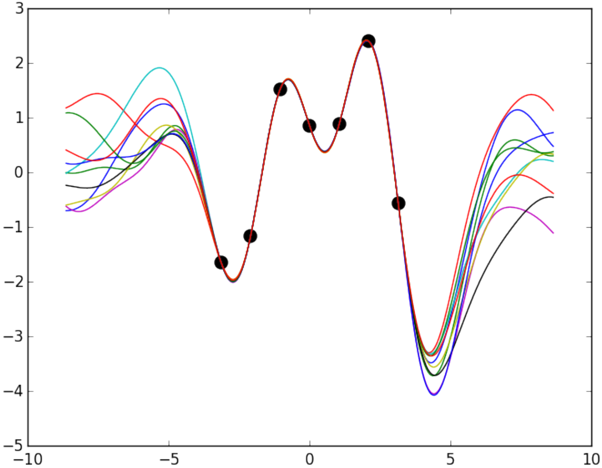
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



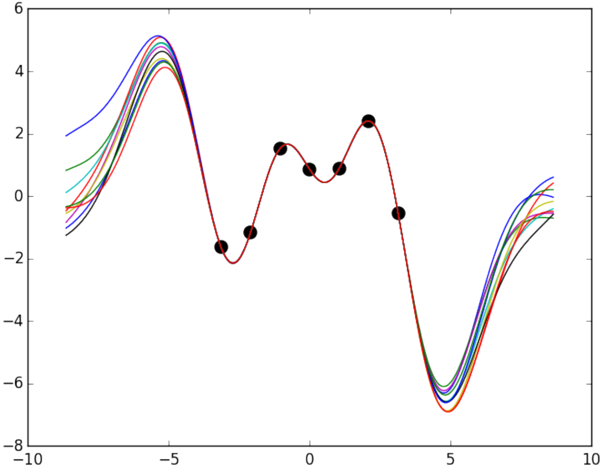
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



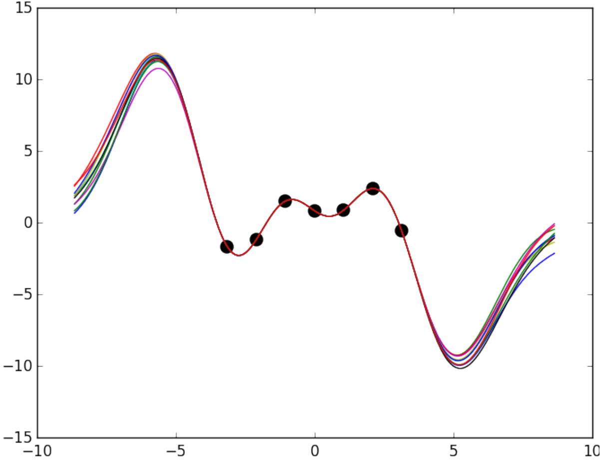
[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



[1] Murphy 2012, p. 15.2

Gaussian Processes[1]



[1] Murphy 2012, p. 15.2

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] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

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

Marginal Likelihood

- We are not interested in **f** directly
- Marginalise out **f**!

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

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

Marginal Likelihood

- We are not interested in **f** directly
- Marginalise out **f**!

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

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

Marginal Likelihood

- We are not interested in **f** directly
- Marginalise out **f**!

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

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

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

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] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

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

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

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] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

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

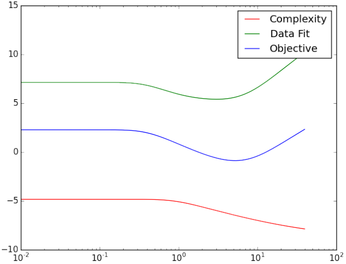
$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi) \quad (29)$$

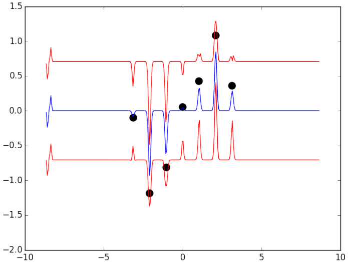
Type-II Maximum Likelihood

- 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] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

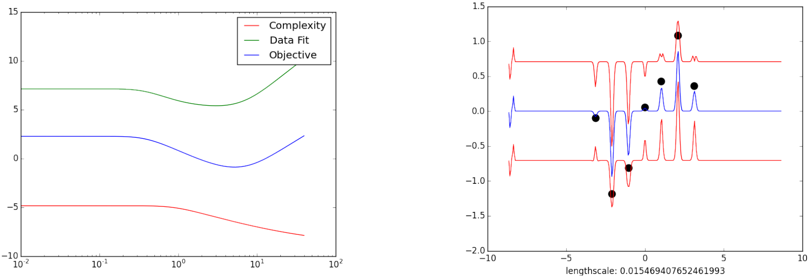




$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

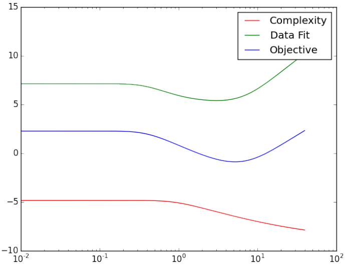
Learning in Gaussian Processes[2]

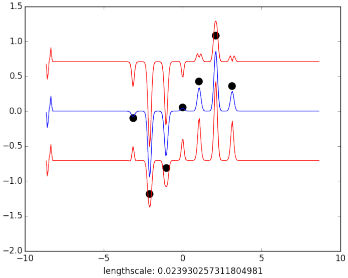


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

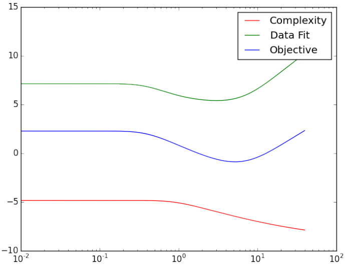


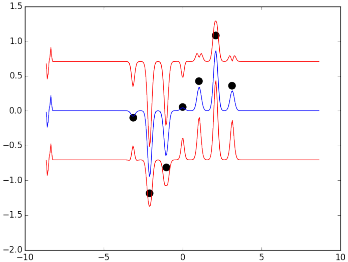


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

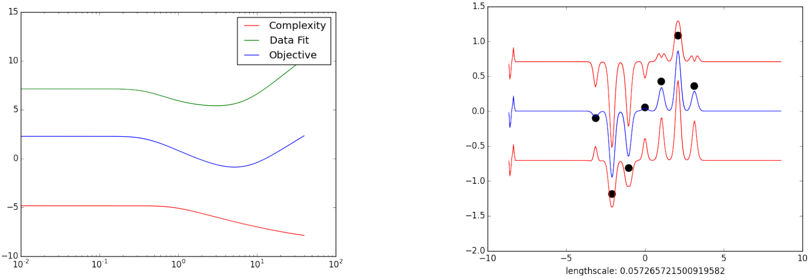




$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

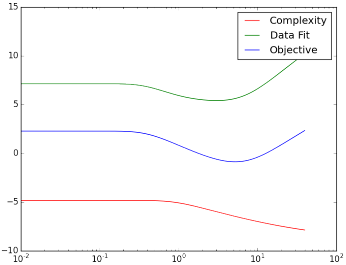
Learning in Gaussian Processes[2]

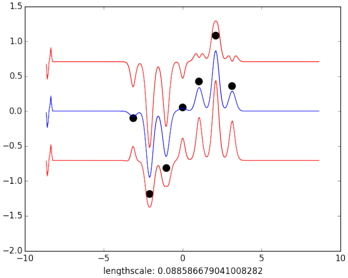


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2]Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

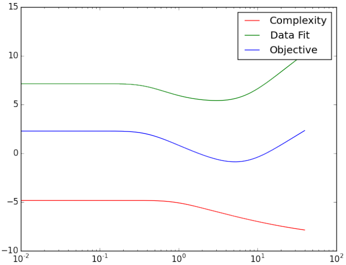


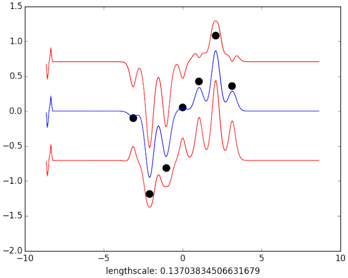


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

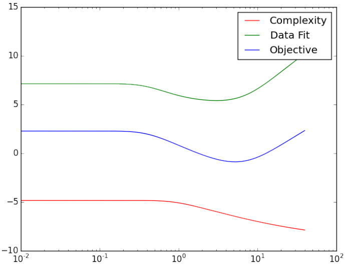


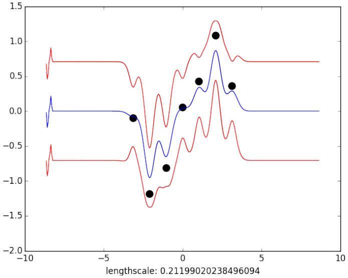


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

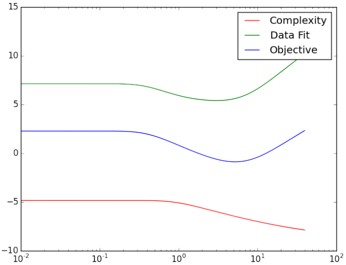


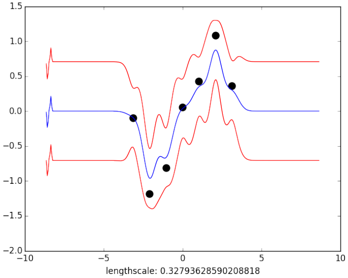


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

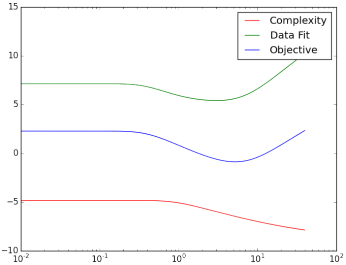


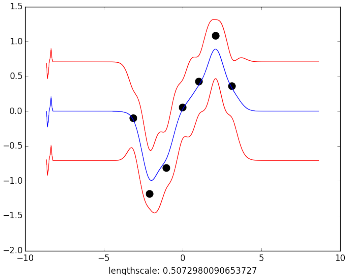


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

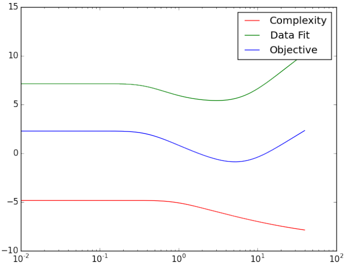


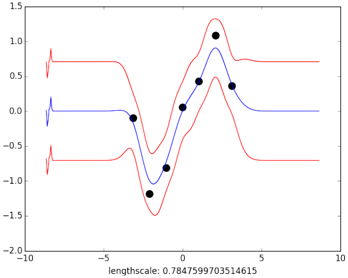


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

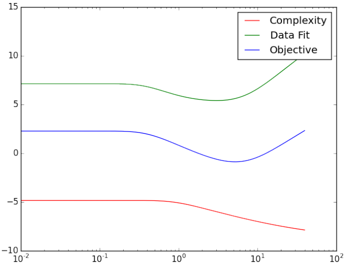


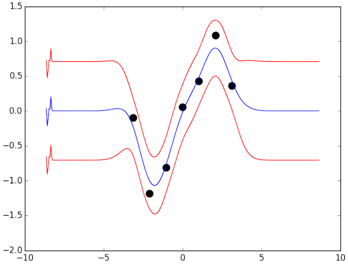


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

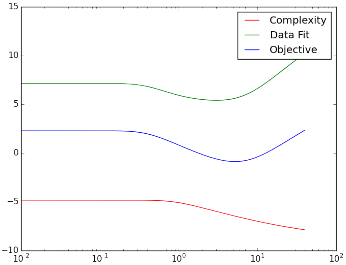


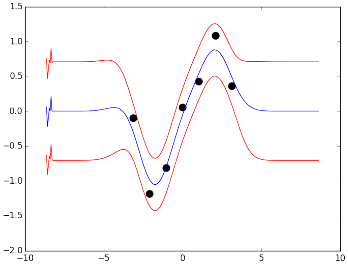


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

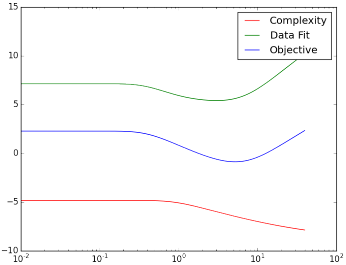


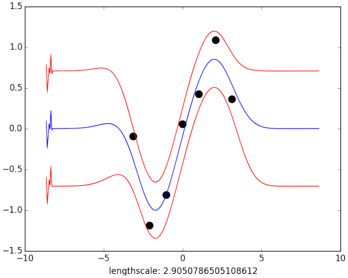


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

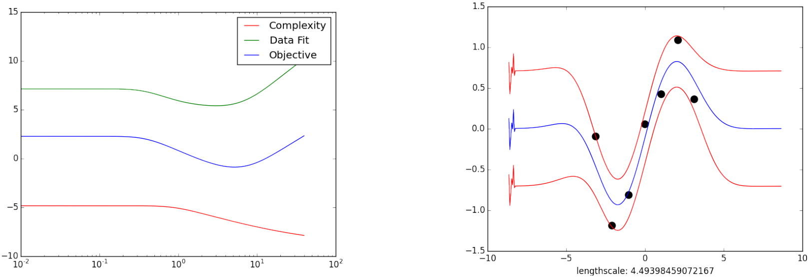




$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

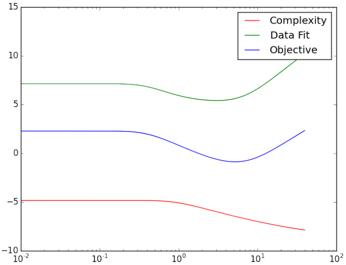
Learning in Gaussian Processes[2]

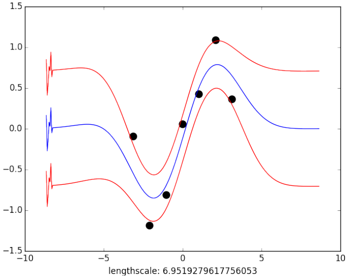


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

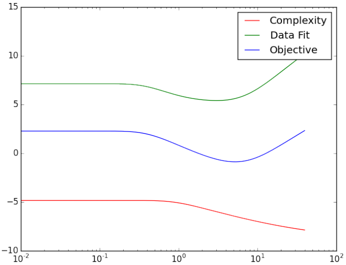


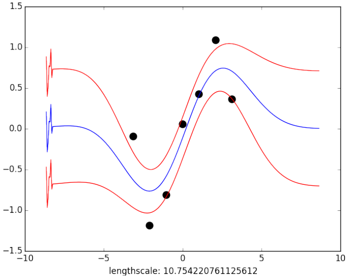


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

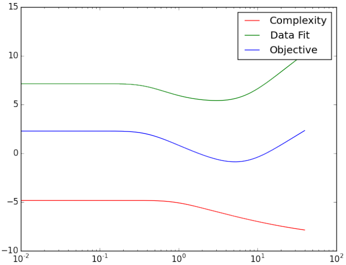


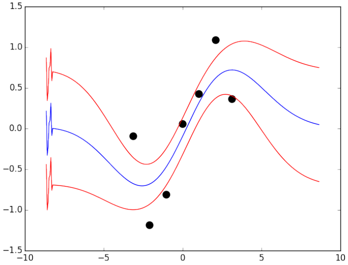


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

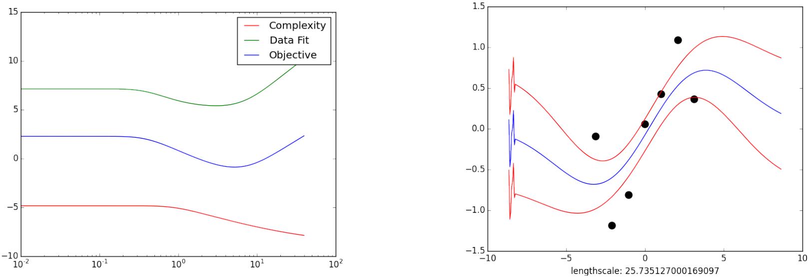




$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

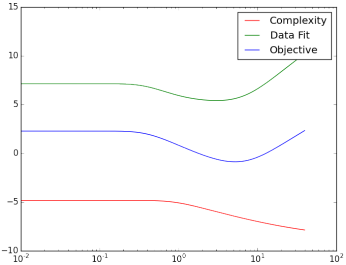
Learning in Gaussian Processes[2]

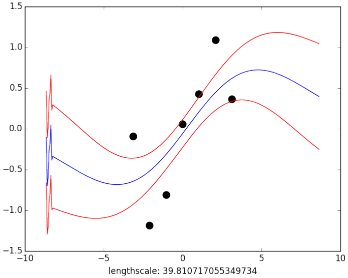


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]

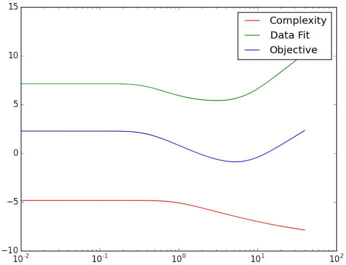


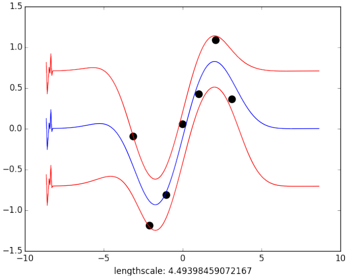


$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

[2] Murphy 2012, p. 15.2.4

Learning in Gaussian Processes[2]





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

[2]Murphy 2012, p. 15.2.4

Introduction

Recap

Representation Learning

Spectral Methods

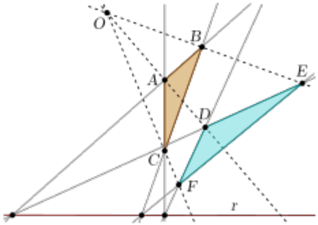
Outline

- Representation Learning
- Why is this important?
- Generative modeling
- Geometry

Reasoning

Geometry

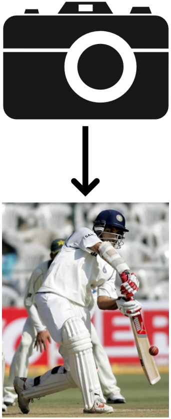
- Intuative conceptual proxy
- Distance
- "Structure"
- Translation, Scaling, ..



Sensory Data

What we are doing

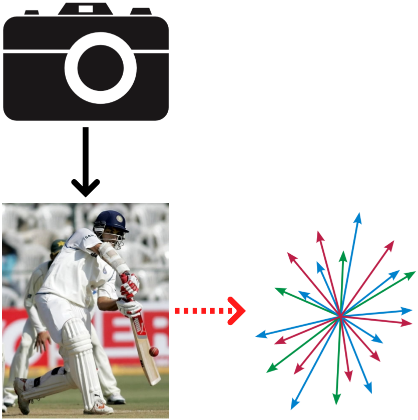
- Sensory representation
  - Capturing process
  - Pixels, Waveforms
- Degrees of freedom and dimensionality



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

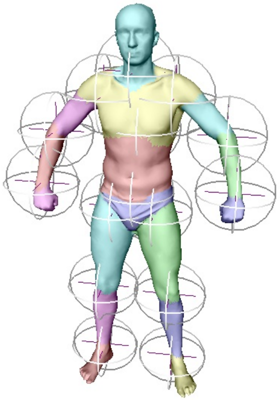


Image data

Image data


Carlsberg

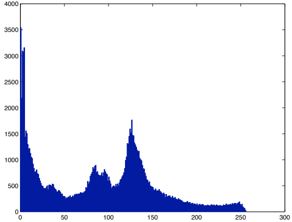

4000
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300

Image data

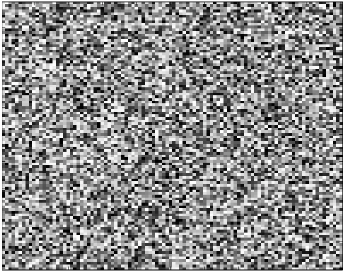


Image data

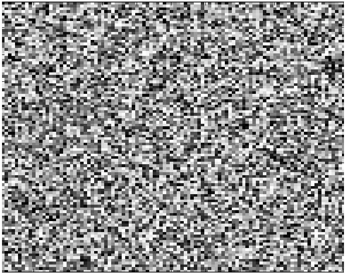


Image data

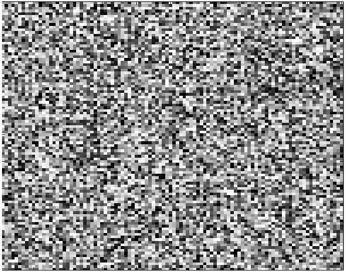
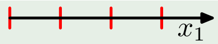


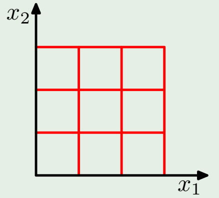
Image data

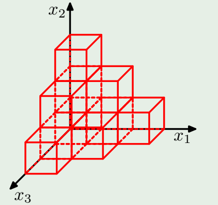
- Parametrisation
- Degrees of Freedom
- Generating parameters

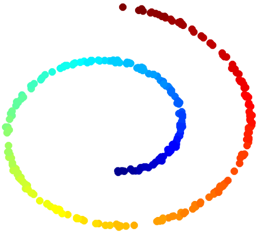
Motivation

- Want to re-parametrise data
- Computational efficiency
- Discover “data-driven” degrees of freedom
  - Unravel data-manifold
- Interpretability
- Generalisation


x_1


x_2
x_1


x_2
x_1
x_3



Re-visit: Principal Component Analysis

- Given data **X** project to directions of maximum variance
- Provides no uncertainty
- How do we compare with other approaches?

$$\text{argmax}_{\mathbf{v}} \sigma(\mathbf{X}\mathbf{v}, \mathbf{X}\mathbf{v}) \quad (30)$$

$$\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v} \quad (31)$$

$$\text{subject to: } :\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1 \quad (32)$$

Re-visit: Principal Component Analysis

- Given data $\mathbf{X}$ project to directions of maximum variance
- Provides no uncertainty
- How do we compare with other approaches?

$$\text{argmax}_{\mathbf{v}}\sigma(\mathbf{X}\mathbf{v}, \mathbf{X}\mathbf{v}) \quad (33)$$

$$\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v} \quad (34)$$

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

Re-visit: Principal Component Analysis

- Given data $\mathbf{X}$ project to directions of maximum variance
- Provides no uncertainty
- How do we compare with other approaches?

$$\text{argmax}_{\mathbf{v}}\sigma(\mathbf{X}\mathbf{v},\mathbf{X}\mathbf{v}) \tag{36}$$

$$\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v} \tag{37}$$

$$\text{subject to: }:\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1 \tag{38}$$

Latent Variable Models[3]

$$p(\mathbf{X}) \tag{39}$$

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

[3] Murphy 2012, p. 12.

Latent Variable Models[3]

$$p(\mathbf{X}|f,\mathbf{Z}) \quad (40)$$

$$\mathbf{f} : \mathbf{Z} \rightarrow \mathbf{X} \quad (41)$$

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

[3] Murphy 2012, p. 12.

Latent Variable Models[3]

$$p(\mathbf{X}|f,\mathbf{Z}) \tag{42}$$

$$\mathbf{f}:\mathbf{Z}\rightarrow\mathbf{X} \tag{43}$$

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

[3] Murphy 2012, p. 12.

Latent Variable Models[3]

$$p(\mathbf{X}|f, \mathbf{Z}) \tag{44}$$

$$\mathbf{f} : \mathbf{Z} \rightarrow \mathbf{X} \tag{45}$$

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

[3] Murphy 2012, p. 12.

Linear Latent Variable Models[4]

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) \tag{46}$$

Regression

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

[4] Murphy 2012, pp. 12.1-12.1.3.

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[5]

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \epsilon \quad (47)$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}) \quad (48)$$

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

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \epsilon \tag{49}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \Psi) \tag{50}$$

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

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \epsilon \tag{51}$$

$$p(\mathbf{X}|\mathbf{Z},\theta) = \mathcal{N}(\mathbf{W}\mathbf{Z},\Psi)) \tag{52}$$

$$p(\mathbf{Z}) = \mathcal{N}(\mu_{\mathbf{0}},\Sigma_{\mathbf{0}}) \tag{53}$$

- Assume the generating mapping to be linear
- For regression we assumed that we knew the inputs **Z**
- Now we do not $\Rightarrow$ specify a prior

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$p(\mathbf{X}|\theta) = \int p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z})\mathrm{d}\mathbf{Z} = \quad (54)$$

$$= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}}) \quad (55)$$

- **Z** and **W** are related
- Integrate out **Z**
  - pick $\mu_0 = 0$, $\Sigma_0 = \mathbf{I}$
- Low dimensional density model of **X**
  - $\mathcal{O}(QD)$ compared to $\mathcal{O}(D^2)$

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$p(\mathbf{X}|\theta) = \int p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z})\mathrm{d}\mathbf{Z} = \quad (56)$$

$$= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}}) \quad (57)$$

$$= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}) \quad (58)$$

- **Z** and **W** are related
- Integrate out **Z**
  - pick $\boldsymbol{\mu}_0 = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- Low dimensional density model of **X**
  - $\mathcal{O}(QD)$ compared to $\mathcal{O}(D^2)$

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$p(\mathbf{X}|\theta) = \int p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z})\mathrm{d}\mathbf{Z} = \quad (59)$$

$$= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^{\mathrm{T}}) \quad (60)$$

$$= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}) \quad (61)$$

- **Z** and **W** are related
- Integrate out **Z**
  - pick $\boldsymbol{\mu}_0 = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- Low dimensional density model of **X**
  - $\mathcal{O}(QD)$ compared to $\mathcal{O}(D^2)$

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$\tilde{\mathbf{W}} = \mathbf{WR} \tag{62}$$

$$p(\mathbf{X}|\theta) = \mathcal{N}(\mu, \Psi + \mathbf{WRR}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}) \tag{63}$$

$$= \mathcal{N}(\mu, \Psi + \mathbf{WW}^{\mathrm{T}}) \tag{64}$$

$$\tag{65}$$

Identifiability

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

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$\mathbf{W}_{ML} = \mathrm{argmax}_{\mathbf{W}} p(\mathbf{X}|\theta) \tag{66}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{67}$$

Probabilistic PCA

- Dimensions of **X** independent given **Z**
  - **W** orthogonal matrix
- Closed form solution Murphy 2012, p. 12.2.2

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

$$\mathbf{W}_{ML} = \operatorname{argmax}_{\mathbf{W}} p(\mathbf{X}|\theta) \tag{68}$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{69}$$

$$\mathbf{W}_{ML} = \mathbf{U}_q(\Lambda - \sigma^2\mathbf{I})^{\frac{1}{2}} \tag{70}$$

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} \tag{71}$$

Probabilistic PCA

- Dimensions of **X** independent given **Z**
  - **W** orthogonal matrix
- Closed form solution Murphy 2012, p. 12.2.2

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

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

[5] Murphy 2012, p. 12.1.1.

Factor Analysis[5]

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 $\epsilon = \lim_{\sigma^2 \rightarrow 0} \sigma^2 \mathbf{I}$
- PCA is incredibly useful but its important to know what you are assuming, the probabilistic formulation allows you to do just that

[5] Murphy 2012, p. 12.1.1.

Gaussian Process Latent Variable Models

History repeats itself

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

Gaussian Process Latent Variable Models

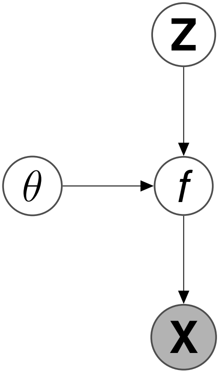
$$p(\mathbf{X}|\mathbf{f}, \mathbf{Z}, \theta) \quad (72)$$

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

Gaussian Process Latent Variable Models

GP-LVM

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



Gaussian Process Latent Variable Models

$$\text{argmax}_{\mathbf{Z},\theta} p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}) \tag{73}$$

$$p(\mathbf{X}|\mathbf{Z},\theta) = \int p(\mathbf{X}|\mathbf{f})p(\mathbf{f}|\mathbf{Z},\theta)\mathrm{d}\mathbf{f} \tag{74}$$

$$p(\mathbf{Z}) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{75}$$

- **GP**-prior sufficiently regularises objective
- Need to set dimensionality of **Z**

Gaussian Process Latent Variable Models

- You can add different priors on latent representations
  - Topological
  - Dynamic GP and a GP
  - Classification
- Any preference you can formulate as a prior

Gaussian Process Latent Variable Models

$$\mathbf{z}_{t+1} = g(\mathbf{z}_t) + \epsilon_z \quad (76)$$

$$g \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{z}_i, \mathbf{z}_j)) \quad (77)$$

- You can add different priors on latent representations
  - Topological
  - Dynamic GP and a GP
  - Classification
- Any preference you can formulate as a prior

Gaussian Process Latent Variable Models

- You can add different priors on latent representations
  - Topological
  - Dynamic GP and a GP
  - Classification
- Any preference you can formulate as a prior

Gaussian Process Latent Variable Models

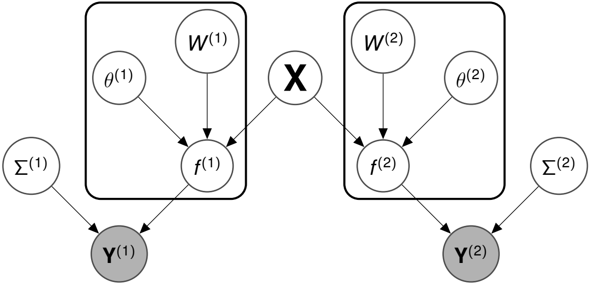
- You can add different priors on latent representations
  - Topological
  - Dynamic GP and a GP
  - Classification
- Any preference you can formulate as a prior

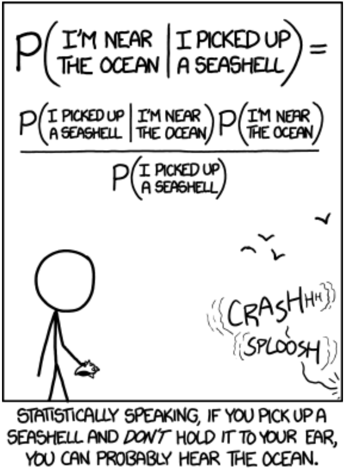
Assignment

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

Grochow *et al.* 2004

My Research


$W^{(1)}$
$\theta^{(1)}$
$\mathbf{X}$
$W^{(2)}$
$\theta^{(2)}$
$\Sigma^{(1)}$
$f^{(1)}$
$f^{(2)}$
$\Sigma^{(2)}$
$\mathbf{Y}^{(1)}$
$\mathbf{Y}^{(2)}$


P(I'M NEAR THE OCEAN | I PICKED UP A SEASHELL) =
P(I PICKED UP A SEASHELL | I'M NEAR THE OCEAN) P(I'M NEAR THE OCEAN)
P(I PICKED UP A SEASHELL)
CRASHHH
SPLOOSH
STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Introduction

Recap

Representation Learning

Spectral Methods

Linear Mapping

Linear Mapping:

$$T: \quad U \rightarrow V$$

- Carries elements from vector space U to vector space V
- Can be expressed by matrices

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

Vector Bases

A Basis is a linearly independent spanning set for a vector space.

- Linearly independent

$$\mathbf{0} = \sum_{i=1}^{D} \alpha_i \mathbf{v}_i$$

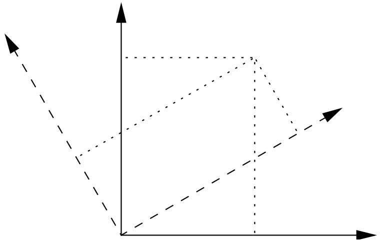
$$\text{only solution} \qquad \alpha = \mathbf{0}$$

- Spanning

$$\langle \mathbf{V} \rangle = \left\{ \sum_{i=1}^{D} \alpha_i \mathbf{v}_i | \alpha_i \in \Re \right\}$$

Vector Bases: Change of Basis

- Element in vector space explained relative a reference
- Change of Basis is a linear transform



Matrix Fundamentals

$$T \quad : \quad V \rightarrow W$$

- dim(V) number of dimensions in representation of V
- image(T) The set of *all* values the map can take, image(T)$\subseteq$ W
$$\mathrm{image}(T) = \{T(\mathbf{x}) : \mathbf{x} \in V\}$$
- kernel(T) The set of *all* values that T maps to zero, kernel(T)$\subseteq$ V
$$\mathrm{kernel}(T) = \{\mathbf{x} : \mathbf{x} \in V | T(\mathbf{x}) = \mathbf{0}\}$$

Rank-Nullity Theorem

$$T \quad : \quad V \rightarrow W$$

$$\text{T Linear Transform} \quad \Rightarrow \quad T: \ \mathbf{Ax} = \mathbf{y}, \begin{cases} \mathbf{x} \in V \\ \mathbf{y} \in W \end{cases}$$

- $\text{rank}(\mathbf{A})$ Number of independent columns
- $\text{rank}(\mathbf{A}) = \dim\left(\text{image}(\mathbf{A})\right)$
- Rank-Nullity Theorem

$$\dim(V) = \underbrace{\dim(\text{image}(\mathbf{A}))}_{\text{rank}(\mathbf{A})} + \dim(\text{kern}(\mathbf{A}))$$

Similarity Transform

$$
\begin{aligned}
T &: \quad V \rightarrow V \\
\mathbf{M}_A &: \quad T: \; V_A \rightarrow V_A \\
\mathbf{M}_B &: \quad T: \; V_B \rightarrow V_B \\
\mathbf{P} &: \quad \text{Change of basis } A \rightarrow B
\end{aligned}
$$

$$
\begin{array}{ccccc}
 & & \mathbf{M}_A & & \\
 & \mathbf{x}_A & \rightarrow & \mathbf{M}_A\mathbf{x}_A & \\
\mathbf{P} & \downarrow & & \uparrow & \mathbf{P}^{-1} \\
 & \mathbf{x}_B & \rightarrow & \mathbf{M}_B\mathbf{x}_B & \\
 & & \mathbf{M}_B & &
\end{array}
$$

$$
\begin{aligned}
\mathbf{M}_B\mathbf{x}_B &= \mathbf{P}\left(\mathbf{M}_A\mathbf{x}_A\right) = \mathbf{P}\left(\mathbf{M}_A\left(\mathbf{P}^{-1}\mathbf{x}_B\right)\right) = \left(\mathbf{P}\mathbf{M}_A\mathbf{P}^{-1}\right)\mathbf{x}_B \\
&\Rightarrow \quad \mathbf{M}_B = \mathbf{P}\mathbf{M}_A\mathbf{P}^{-1} \\
&\Rightarrow \quad \mathbf{M}_A \sim \mathbf{M}_B
\end{aligned}
$$

Similar Matrices

$$\mathbf{A} \sim \mathbf{A}$$ Proof

$$\mathbf{A} \sim \mathbf{B} \Rightarrow \det(\mathbf{A}) = \det(\mathbf{B})$$ Proof

$$\mathbf{A} \sim \mathbf{B} \Rightarrow \mathrm{trace}(\mathbf{A}) = \mathrm{trace}(\mathbf{B})$$ Proof

$$\mathbf{A} \sim \mathbf{B} \Rightarrow \mathbf{A}^m \sim \mathbf{B}^m, m \in \mathbf{Z}^+$$ Proof

$\mathbf{A} \sim \mathbf{B} \Rightarrow$ **A** invertible if and only if **B** invertible Proof

Determinant, trace and invertibility are *invariant* under similarity.

Similarity Transform: Spectral Decomposition

- Special Similarity Transform

$$\begin{aligned} \mathbf{A} &= \mathbf{V}\Lambda\mathbf{V}^{-1} \\ \Lambda_{ij} &= \begin{cases} 0 & i \neq j \\ \lambda_i & i = j \end{cases} \\ \mathbf{V}\mathbf{V}^T &= \mathbf{I} \Rightarrow \mathbf{V}^{-1} = \mathbf{V}^T \end{aligned}$$

- $\Rightarrow$ *Spectral Decomposition*
-

$$\begin{aligned} \text{Eigenvectors} \qquad & \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] \\ \text{Eigenvalues} \qquad & \text{diag}(\Lambda) = \{\lambda_1, \ldots, \lambda_N\} \end{aligned}$$

- If $\lambda_i \geq 0, \ \forall i \Rightarrow \mathbf{A}$ *Positive Semidefinite*

Spectral Factorization

- A matrix can be written as a linear combination of rank one matrices ▸ Proof

$$\begin{aligned} \mathbf{A} &= \sum_{k=1}^{N} \lambda_k \mathbf{v}_k \mathbf{v}_k^T \\ \mathbf{A}_i &= \lambda_i \mathbf{v}_i \mathbf{v}_i^T \end{aligned}$$

- Best rank i approximation to $\mathbf{A}$

$$\mathbf{A}_{\rightarrow i} = \sum_{k=1}^{i} \lambda_k \mathbf{v}_k \mathbf{v}_k^T, \; \lambda_i \geq \lambda_j, \; i \leq j$$

▸ Proof

Multidimensional Scaling

- Visualization of proximities
- Proximity: $\sim$ Dissimilarity Measure
- "Find a geometric configuration which conserves a given proximity relation"

Multidimensional Scaling

- N entities with proximity relations δ_{ij}
- Must be metric
- Find embedding $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$ to minimize

$$
E_{MDS} = ||\mathbf{D} - \Delta||_F \quad \begin{cases} \mathbf{D}_{ij} = ||\mathbf{y}_i - \mathbf{y}_j||_{L2} \\ \Delta_{ij} = \delta_{ij} \end{cases}
$$

$$||\mathbf{A}||_F = \sqrt{\mathrm{trace}\left(\mathbf{A}\mathbf{A}^T\right)} = \sqrt{\sum_{i=1}^{N} \lambda_i^2}$$

$$||\mathbf{D} - \Delta||_F = \left\{ \Delta = \mathbf{V}\Lambda\mathbf{V}^T \Rightarrow \Delta = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \right\} =$$

$$= \quad ||\mathbf{D} - \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T||_F = ||\sum_{i=1}^{d} q_i \mathbf{v}_i \mathbf{v}_i^T - \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T||_F =$$

$$= \quad ||\sum_{i=1}^{d}(q_i - \lambda_i)\mathbf{v}_i \mathbf{v}_i^T - \sum_{i=d+1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T||_F$$

Choose $\mathbf{D} = \mathbf{A}_{\rightarrow d} \Rightarrow E_{MDS} = \sqrt{\sum_{i=d+1}^{N} \lambda_i^2}$

Multidimensional Scaling

Generate geometrical configuration **Y** that could generate **D**

1. Convert distance matrix D to Gram matrix $\mathbf{G} = \mathbf{Y}\mathbf{Y}^T$

Proof

2. Diagonalise Gram matrix G

$$\begin{aligned}\mathbf{G} &= \mathbf{Y}\mathbf{Y}^T = \mathbf{V}\Lambda\mathbf{V}^T = \left(\mathbf{V}\Lambda^{\frac{1}{2}}\right)\left(\Lambda^{\frac{1}{2}}\mathbf{V}^T\right) = \\ &= \left(\mathbf{V}\Lambda^{\frac{1}{2}}\right)\left(\mathbf{V}\left(\Lambda^{\frac{1}{2}}\right)^T\right)^T = \left(\mathbf{V}\Lambda^{\frac{1}{2}}\right)\left(\mathbf{V}\Lambda^{\frac{1}{2}}\right)^T\end{aligned}$$

3. Chose $\mathbf{Y} = \mathbf{V}\Lambda^{\frac{1}{2}}$
4. Dimension of **Y**: $\mathrm{rank}(\mathbf{Y}\mathbf{Y}^T) = \mathrm{rank}(\mathbf{G}) = \mathrm{rank}(\mathbf{D}) = d$

PCA Equivalence

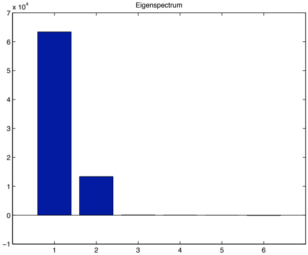
Multidimensional Scaling: Example

[6]	*Man*	*Ox*	*Lon*	*Bristol*	*LFC*	*Bir*
Manchester	0	203	262	224	46	114
Oxford	203	0	83	95	217	91
London	262	83	0	170	285	161
Bristol	224	95	170	0	217	122
Liverpool	46	217	285	217	0	126
Birmingham	114	91	161	122	126	0

[6] http://www.geobytes.com/CityDistanceTool.htm

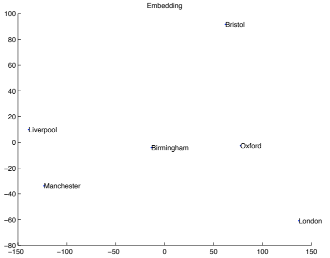
Multidimensional Scaling: Example[7]

- Two significant non-zero eigenvalues
- ⇒ Even though we know earth is a sphere ...



[7]/example/mds_city.m

Multidimensional Scaling: Example[8]



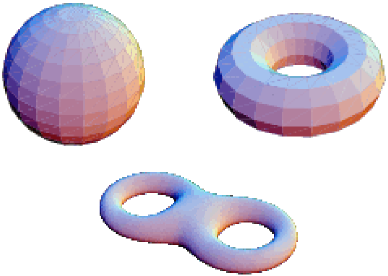
- 1^{st} $\sim$ North-South
- 2^{nd} $\sim$ West-East

[8] /example/mds_city.m

Non linearities[9]

Manifold

- Generalisation of low dimensional object embedded in high dimensional space
- Similarity?
- Local similarity
- Extend local similarity to global



[9] /src/nonlinear.m

Non linearities[9]

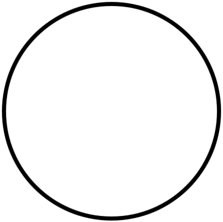
Definition

"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[a]

[a] http://en.wikipedia.org/wiki/Manifold

[9] /src/nonlinear.m

Non linearities[9]



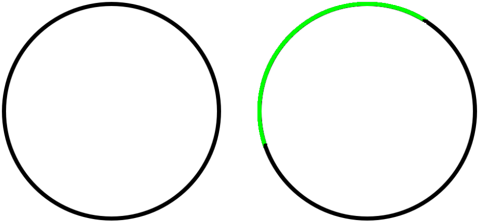
Definition

"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[a]

[a] http://en.wikipedia.org/wiki/Manifold

[9] /src/nonlinear.m

Non linearities[9]



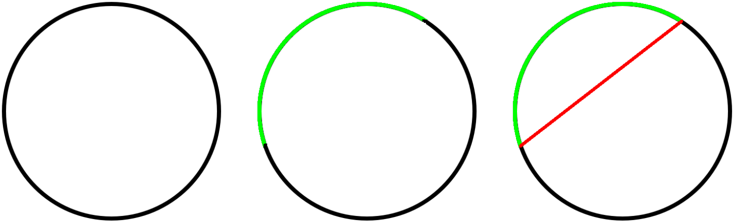
Definition

"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[a]

[a] http://en.wikipedia.org/wiki/Manifold

[9] /src/nonlinear.m

Non linearities[9]



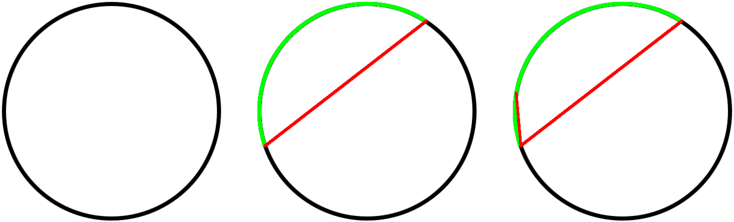
Definition

"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[a]

[a] http://en.wikipedia.org/wiki/Manifold

[9] /src/nonlinear.m

Non linearities[9]



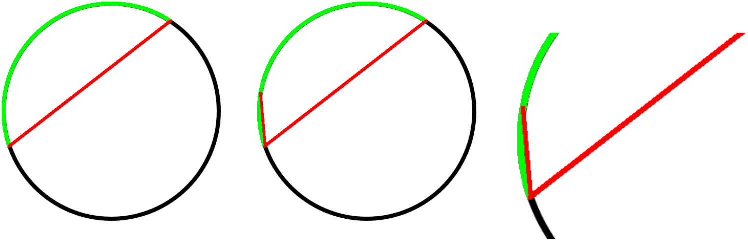
Definition

“In mathematics, a manifold is a topological space that near each point resembles Euclidean space”[a]

[a] http://en.wikipedia.org/wiki/Manifold

[9] /src/nonlinear.m

Non linearities[9]



Definition

"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[a]

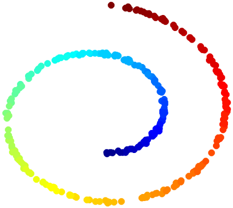
[a] http://en.wikipedia.org/wiki/Manifold

[9] /src/nonlinear.m

Non linearities[9]

Manifold

- Generalisation of low dimensional object embedded in high dimensional space
- Similarity?
- Local similarity
- Extend local similarity to global

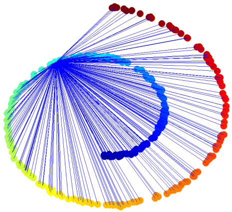


[9] /src/nonlinear.m

Non linearities[9]

Manifold

- Generalisation of low dimensional object embedded in high dimensional space
- Similarity?
- Local similarity
- Extend local similarity to global

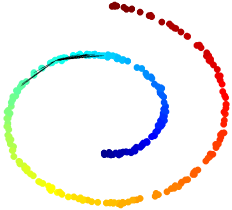


[9]/src/nonlinear.m

Non linearities[9]

Manifold

- Generalisation of low dimensional object embedded in high dimensional space
- Similarity?
- Local similarity
- Extend local similarity to global

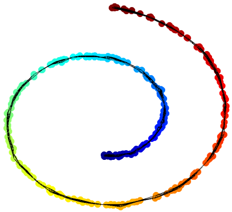


[9] /src/nonlinear.m

Non linearities[9]

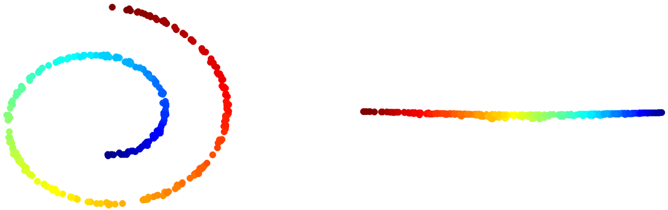
Manifold

- Generalisation of low dimensional object embedded in high dimensional space
- Similarity?
- Local similarity
- Extend local similarity to global



[9] /src/nonlinear.m

Non linearities[9]



[9] /src/nonlinear.m

Proximity Graph

1. Identify neighbors of each data point $\mathbf{x}_i \in N(\mathbf{x_j})$
2. Build graph $\mathbf{P} = \left\{ \underbrace{\mathbf{X}}_{\textit{vertexset}}, \underbrace{\mathbf{W}}_{\textit{edgeset}} \right\}$
   - Put edges between vertices's in neighborhood
   - Assume **P** connected (and in most cases symmetric)
3. **Objective:** *Complete* **P** *to make it fully connected*
4. Different algorithms have different strategies
   - What are the edge weights?
   - How to complete **P**

Maximum Variance Unfolding

- *Weinberg, Sha, Saul* - ICML & CVPR 2004
- First presented as Semi-Definite Embeddings
- Formulate dimensionality reduction in terms of Gram matrix

Maximum Variance Unfolding

- Want to keep local structure $(\mathbf{x}_i, \mathbf{x}_j) \in W$

$$
\begin{aligned}
||\mathbf{y}_i - \mathbf{y}_j||_{L2}^2 &= ||\mathbf{x}_i - \mathbf{x}_j||_{L2}^2 \\
\Rightarrow \quad \mathbf{K}_{ii} + \mathbf{K}_{jj} - \mathbf{K}_{ij} - \mathbf{K}_{ji} &= \mathbf{G}_{ii} + \mathbf{G}_{jj} - \mathbf{G}_{ij} - \mathbf{G}_{ji}
\end{aligned}
$$

Proof

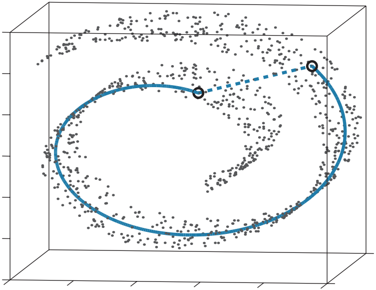
- Remove Translational Invariance

$$
||\sum_{i=1}^{N} \mathbf{y_i}||_{L2}^2 = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{K}_{ij} = 0
$$

Proof

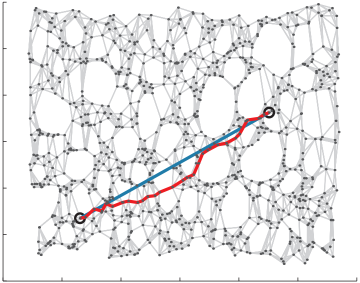
- Need to be valid Gram matrix $\Rightarrow \ \mathbf{K} \succcurlyeq \ 0$

Maximum Variance Unfolding



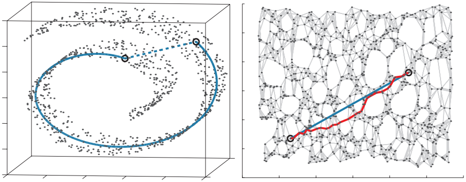
Any “fold” of the manifold between two points will **decrease** the *Euclidean* distance between the points while the *Manifold* distance remains **constant**

Maximum Variance Unfolding



If manifold is **maximally** stretched between two points the *Euclidean* distance will **equal** the *Manifold* distance

Maximum Variance Unfolding



Maximise all pairwise distance outside local neighborhood (upper bound)

$$\max \sum_{i=1}^{N} \sum_{j=1}^{N} ||\mathbf{y}_i - \mathbf{y}_j||_{L2}^2$$

$$\Rightarrow \quad \max(\text{trace}(\mathbf{K}))$$

Proof

Maximum Variance Unfolding: Algorithm

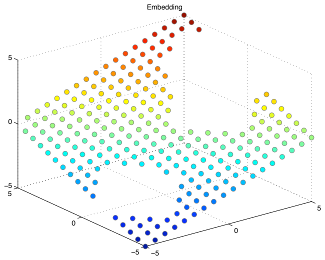
1. Compute Proximity Graph
2. Compute Local Gram Matrix $\mathbf{G}$
3. Compute Global Gram Matrix $\mathbf{K}$

$$\begin{aligned}
& \max(\text{trace}(\mathbf{K})) \\
\text{subject to :} \quad & \mathbf{K} \succcurlyeq 0 \\
& \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{K}_{ij} = 0 \\
& \mathbf{K}_{ii} + \mathbf{K}_{jj} - \mathbf{K}_{ij} - \mathbf{K}_{ji} = \mathbf{G}_{ii} + \mathbf{G}_{jj} - \mathbf{G}_{ij} - \mathbf{G}_{ji}
\end{aligned}$$

Instance of *Semidefinite Programming*

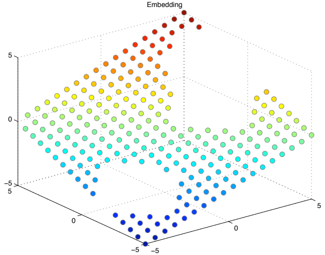
4. Apply MDS to $\mathbf{K}$

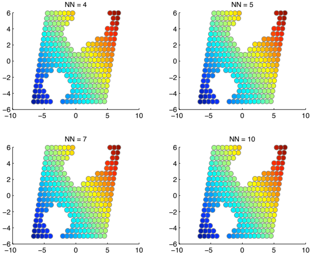
Maximum Variance Unfolding: Example[10]



[10]/algos/mvu_embed.m

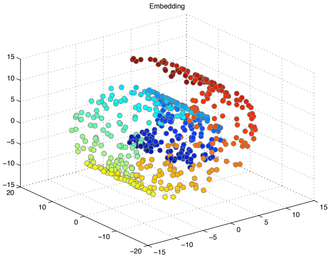
Maximum Variance Unfolding: Example[10]





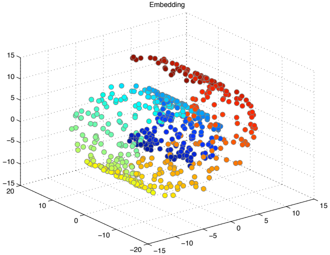
[10]/algos/mvu_embed.m

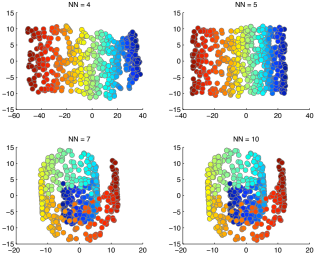
Maximum Variance Unfolding: Example[10]



[10] /algos/mvu_embed.m

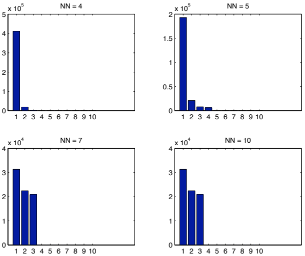
Maximum Variance Unfolding: Example[10]

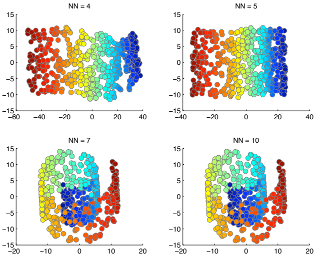




[10] /algos/mvu_embed.m

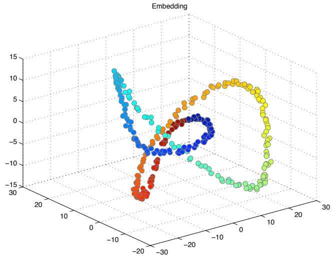
Maximum Variance Unfolding: Example[10]





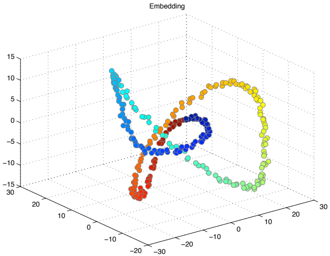
[10] /algos/mvu_embed.m

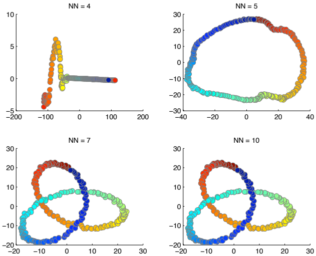
Maximum Variance Unfolding: Example[10]



[10] /algos/mvu_embed.m

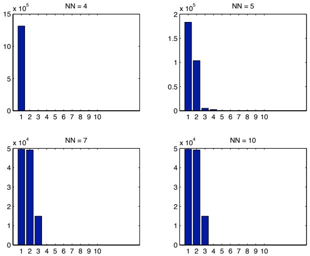
Maximum Variance Unfolding: Example[10]

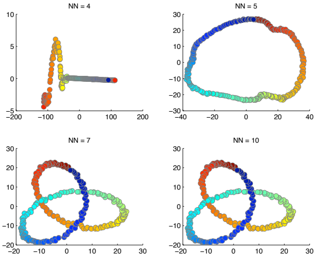




[10] /algos/mvu_embed.m

Maximum Variance Unfolding: Example[10]





[10] /algos/mvu_embed.m

Maximum Variance Unfolding: Summary

- MDS on optimised constrained Gram Matrix
- \+ Dimensionality through eigen spectra
- \+ Convex optimisation problem
- \+ Handles holes and non-convex manifolds
- \- Expensive

Next Time

Lecture 9

- December 1st 10-12 Q34
- Hierarchical Models
  - priors
  - models
- Neural Networks
- Summary of my part of the course
  - what to do next
- Complete assignment Task 2.3 and 2.4

Next Time

Lecture 9

- December 1st 10-12 Q34
- Hierarchical Models
  - priors
  - models
- Neural Networks
- Summary of my part of the course
  - what to do next
- Complete assignment Task 2.3 and 2.4

e.o.f.

References I

- Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029.
- 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.

References II

Keith Grochow *et al.* “Style-based inverse kinematics”. In: *SIGGRAPH ’04: SIGGRAPH 2004 Papers* (Aug. 2004). DOI: 10.1145/1186562.1015755. URL: http://portal.acm.org/citation.cfm?id=1186562.1015755&coll=DL&dl=ACM&CFID=19928546&CFTOKEN=59187189.

Appendix

Similar Matrices: Self-Similarity

$$\mathbf{A} = \mathbf{I}\mathbf{A}\mathbf{I}^{-1} = \mathbf{I}^{-1}\mathbf{A}\mathbf{I}$$

Return

Similar Matrices: Symmetry

$$
\begin{aligned}
\mathbf{A} \quad \sim \quad & \mathbf{B} \Rightarrow \mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \\
\det \mathbf{B} \quad & = \det\left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right) = \det(\mathbf{P}^{-1})\det(\mathbf{A})\det(\mathbf{P}) = \\
& = \det(\mathbf{A})\det(\mathbf{P}^{-1})\det(\mathbf{P}) = \det(\mathbf{A})\frac{1}{\det(\mathbf{P})}\det(\mathbf{P}) = \\
& \det(\mathbf{B})
\end{aligned}
$$

◂ Return

Similar Matrices: Trace

$$
\begin{aligned}
\mathbf{A} \quad \sim \quad & \mathbf{B} \Rightarrow \mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \\
\text{trace}(\mathbf{B}) \quad & = \text{trace}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) = \{\text{trace}(\mathbf{A}\mathbf{B}) = \text{trace}(\mathbf{A}\mathbf{B})\} = \\
& = \text{trace}\left(\left(\mathbf{P}\mathbf{P}^{-1}\right)\mathbf{A}\right) = \text{trace}(\mathbf{A})
\end{aligned}
$$

◂ Return

Similar Matrices: Power

$$\begin{aligned}
\mathbf{A} &\sim \mathbf{B} \Rightarrow \mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \\
\mathbf{B}^2 &= \left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right)^2 = \left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right)\left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right) = \\
&= \left(\mathbf{P}^{-1}\mathbf{A}\right)\left(\underbrace{\mathbf{P}\mathbf{P}^{-1}}_{=\mathbf{I}}\right)(\mathbf{A}\mathbf{P}) = \\
&= \mathbf{P}^{-1}\mathbf{A}\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}^2\mathbf{P}
\end{aligned}$$

Prove further powers by induction over exponent

Return

Similar Matrices: Invertability

$$\begin{aligned} \mathbf{A} &\sim \mathbf{B} \Rightarrow \mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \\ &\Rightarrow \det(\mathbf{A}) = \det(\mathbf{B}) \end{aligned}$$

$\mathbf{A}^{-1}$ Exists if $\det(\mathbf{A}) \neq 0$

$$\det(\mathbf{B}) \neq 0 \iff \det(\mathbf{A}) \neq 0$$

Return

$$\mathbf{A}_{ij} = \sum_{k=1}^{N} \mathbf{V}_{ik}\mathbf{D}_{kk}\left(\mathbf{V}^{T}\right)_{kj} = \sum_{k=1}^{N} (\mathbf{v}_k)_i \, \lambda_k \, (\mathbf{v}_k)_j$$

$$= \sum_{k=1}^{N} \left(\lambda_k \mathbf{v}_k \mathbf{v}_k^{T}\right)_{ij}$$

Return

Rank Approximation

$$
\begin{aligned}
||\mathbf{A}-\mathbf{B}||_F &= ||\sum_{i=1}^{N}\lambda_i \mathbf{v}_i\mathbf{v}_i^T - \sum_{i=1}^{N} q_i \mathbf{v}_i\mathbf{v}_i^T||_F = \\
&= ||\sum_{i=1}^{N}(\lambda_i - q_i)\mathbf{v}_i\mathbf{v}_i^T|| = \\
&= \left\{ ((\lambda_i - q_i)\mathbf{v}_i \underbrace{\mathbf{v}_i^T)\mathbf{v}_i}_{=1} = (\lambda_i - q_i)\mathbf{v}_i \right\} = \\
&= \sqrt{\sum_{i=1}^{N}(\lambda_i - q_i)^2}
\end{aligned}
$$

Return

Multidimensional Scaling

Define:

$$
\begin{aligned}
d_{ij}^2 &= \sum_{k=1}^{d}(x_{ki} - x_{kj})^2 = \mathbf{x}_i^T\mathbf{x}_i + \mathbf{x}_j^T\mathbf{x}_j - 2\mathbf{x}_i\mathbf{x}_j \\
g_{ij} &= \sum_{k=1}^{d} x_{ki}x_{kj} = \mathbf{x}_i^T\mathbf{x}_j \\
&\Rightarrow \quad d_{ij}^2 = g_{ii} + g_{jj} - 2g_{ij} \\
\text{Centering:} \quad & \sum_{i=1}^{N} g_{ij} = \sum_{i=1}^{N} \mathbf{x}_i^T\mathbf{x}_j = \underbrace{(\sum_{i=1}^{N}\mathbf{x}_i^T)}_{=\mathbf{0}}\mathbf{x}_j = 0
\end{aligned}
$$

Multidimensional Scaling

Want to Express **G** in terms of **D**

$$g_{ij} = \frac{1}{2}(g_{ii} + g_{jj} - d_{ij}^2)$$

$$\frac{1}{N}\sum_{i=1}^{N} d_{ij}^2 = g_{jj} + \frac{1}{N}\sum_{i=1}^{N} g_{ii}$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = \frac{2}{N}\sum_{i=1}^{N} g_{ii}$$

$$\Rightarrow g_{ij} = \frac{1}{2}\left(\frac{1}{N}\left(\sum_{k=1}^{N} d_{kj}^2 + \sum_{k=1}^{N} d_{ik}^2 - \frac{1}{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2\right) - d_{ij}^2\right)$$

◂ Return: MDS ◂ Return: MVU

PCA MDS Equivalence

$$
\begin{aligned}
\mathbf{G} &= \mathbf{X}\mathbf{X}^T = \mathbf{V}\Lambda\mathbf{V}^T \\
&\Rightarrow (\mathbf{X}\mathbf{X}^T)\mathbf{v}_i = \lambda_i \mathbf{v}_i \\
&\Rightarrow \frac{1}{N-1}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)\mathbf{v}_i = \lambda_i \frac{1}{N-1}\mathbf{X}^T\mathbf{v}_i \\
&\Rightarrow \underbrace{\frac{1}{N-1}\mathbf{X}^T(\mathbf{X}}_{\mathbf{S}}\mathbf{X}^T)\mathbf{v}_i = \lambda_i \frac{1}{N-1}\mathbf{X}^T\mathbf{v}_i \\
&\Rightarrow \mathbf{S}\underbrace{(\mathbf{X}^T\mathbf{v}_i)}_{\text{eigenvectors?}} = \underbrace{\frac{\lambda_i}{N-1}}_{\text{eigenvalue?}}\underbrace{(\mathbf{X}^T\mathbf{v}_i)}_{\textit{eigenvector}?}
\end{aligned}
$$

PCA MDS Equvalence

Enforce orthogonality

$$\left(\mathbf{X}^T\mathbf{v}_i\right)^T\left(\mathbf{X}^T\mathbf{v}_i\right) = \mathbf{v}_i^T\mathbf{X}\mathbf{X}^T\mathbf{v}_i = \lambda_i$$

$$\Rightarrow \quad \frac{1}{\sqrt{\lambda_i}}\mathbf{v}_i^T\mathbf{X}\mathbf{X}^T\mathbf{v}_i\frac{1}{\sqrt{\lambda_i}} = \left(\frac{1}{\sqrt{\lambda_i}}\right)^2\lambda_i = 1$$

$$\left(\mathbf{X}^T\mathbf{v}_i\right)\frac{1}{\sqrt{\lambda_i}})^T\left(\mathbf{X}^T\mathbf{v}_i\frac{1}{\sqrt{\lambda_i}}\right) = 1$$

PCA MDS Equivalence

$$\begin{aligned}
\text{Define: } \mathbf{v}_i^{\text{PCA}} &= \mathbf{X}^T\mathbf{v}_i\frac{1}{\sqrt{\lambda_i}} \\
\mathbf{y}_i^{\text{PCA}} &= \mathbf{X}\mathbf{v}_i^{\text{PCA}} = \mathbf{X}\mathbf{X}^T\mathbf{v}_i\frac{1}{\sqrt{\lambda_i}} = \\
&= \lambda_i\mathbf{v}_i\frac{1}{\sqrt{\lambda_i}} = \sqrt{\lambda_i}\mathbf{v}_i \\
\mathbf{y}_i^{\text{MDS}} &= \mathbf{v}_i\sqrt{\lambda_i} = \sqrt{\lambda_i}\mathbf{v}_i \\
&\Rightarrow \mathbf{y}_i^{\text{PCA}} = \mathbf{y}_i^{\text{MDS}}
\end{aligned}$$

◂ PCA

Maximum Variance Unfolding: Objective

$$
\begin{aligned}
\sum_{i=1}^{N} g_{ii} &= \sum_{i=1}^{N} \frac{1}{2}\left(\frac{1}{N}\left(\sum_{k=1}^{N} d_{kj}^2 + \sum_{k=1}^{N} d_{ik}^2 - \frac{1}{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2\right) - d_{ii}^2\right) = \\
&= \underbrace{\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{N} d_{ki}^2 + \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{N} d_{ik}^2}_{\text{symmetry } = \frac{1}{2N} 2 \sum_{i=1}^{N}\sum_{k=1}^{N} d_{ki}^2} - \\
&- \frac{1}{2N^2} N \sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2 - \frac{1}{2}\sum_{i}^{N} \underbrace{d_{ii}^2}_{=0} =
\end{aligned}
$$

Maximum Variance Unfolding: Objective

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} d_{ki}^2 - \frac{1}{2N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2 = \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 \\
\mathrm{trace}(\mathbf{G}) &= \sum_{i=1}^{N} g_{ii} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} ||\mathbf{y}_i - \mathbf{y}_j||_{L2}^2
\end{aligned}
$$

◂ Return

Maximum Variance Unfolding: Centering

$$
\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ii} &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{2}\left(\frac{1}{N}\left(\sum_{k=1}^{N} d_{kj}^2 + \sum_{k=1}^{N} d_{ik}^2 - \right.\right. \\
&- \left.\left.\frac{1}{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2\right) - d_{ij}^2\right) = \\
&= \frac{1}{2N}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} d_{kj}^2}_{=N\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2} + \frac{1}{2N}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} d_{ik}^2}_{=N\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2} -
\end{aligned}
$$

Maximum Variance Unfolding: Centering

$$- \quad \underbrace{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2}_{=N^2\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 =$$

$$= \quad \underbrace{(\frac{1}{2}+\frac{1}{2}-\frac{1}{2}-\frac{1}{2})}_{=0}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = 0$$

$$||\sum_{i=1}^{N}\mathbf{y}_i||_{L2}^2 \quad \Rightarrow \quad \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{K}_{ij} = 0$$

Return

Spectral Theorem

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \qquad \mathbf{A} = \mathbf{V}\Delta\mathbf{V}^T,\ ||\mathbf{x}||_{L2} = 1$$

$$\mathbf{x} = 1\sum_{i=1}^{N}\alpha_i\mathbf{v}_i$$

$$||\alpha|| = 1$$

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \left(\sum_{i=1}^{N}\alpha_i\mathbf{v}_i\right)^T\mathbf{A}\left(\sum_{i=1}^{N}\alpha_i\mathbf{v}_i\right) =$$

$$= \left(\sum_{i=1}^{N}\alpha_i\mathbf{v}_i\right)^T\left(\sum_{i=1}^{N}\lambda_i\mathbf{v}_i\mathbf{v}_i^T\right)\left(\sum_{i=1}^{N}\alpha_i\mathbf{v}_i\right) =$$

Spectral Theorem

$$= \left(\sum_{i=1}^{N}\alpha_i\mathbf{v}_i\right)^T\left(\sum_{i=1}^{N}\lambda_i\mathbf{v}_i\mathbf{v}_i^T\right)\left(\sum_{i=1}^{N}\alpha_i\mathbf{v}_i\right) =$$

$$= \left\{\mathbf{v}_i^T\mathbf{v}_j = \begin{cases} 1 & i=j \\ 0 & \text{otherwise} \end{cases}\right\} =$$

$$= \sum_{i=1}^{N}\alpha_i^2\lambda_i\underbrace{\mathbf{v}_i^T\mathbf{v}_i}_{=1}\underbrace{\mathbf{v}_i^T\mathbf{v}_i}_{=1} =$$

$$= \sum_{i=1}^{N}\alpha_i^2\lambda_i\begin{cases} \max & : \quad \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_1 \quad \mathbf{x} = \mathbf{v}_1 \\ \min & : \quad \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_N \quad \mathbf{x} = \mathbf{v}_N \end{cases}$$

◂ Return LLE ◂ Return Laplacian

Maximum Variance Unfolding: Objective

$$
\begin{aligned}
\sum_{i=1}^{N} g_{ii} &= \sum_{i=1}^{N} \frac{1}{2}\left(\frac{1}{N}\left(\sum_{k=1}^{N} d_{kj}^2 + \sum_{k=1}^{N} d_{ik}^2 - \frac{1}{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2\right) - d_{ii}^2\right) = \\
&= \underbrace{\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{N} d_{ki}^2 + \frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{N} d_{ik}^2}_{\text{symmetry } = \frac{1}{2N} 2 \sum_{i=1}^{N}\sum_{k=1}^{N} d_{ki}^2} - \\
&- \frac{1}{2N^2} N \sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2 - \frac{1}{2}\sum_{i}^{N} \underbrace{d_{ii}^2}_{=0} =
\end{aligned}
$$

Maximum Variance Unfolding: Objective

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} d_{ki}^2 - \frac{1}{2N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2 = \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 \\
\mathrm{trace}(\mathbf{G}) &= \sum_{i=1}^{N} g_{ii} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = \\
&= \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} ||\mathbf{y}_i - \mathbf{y}_j||_{L2}^2
\end{aligned}
$$

◂ Return

Maximum Variance Unfolding: Centering

$$
\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ii} &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{2}\left(\frac{1}{N}\left(\sum_{k=1}^{N} d_{kj}^2 + \sum_{k=1}^{N} d_{ik}^2 - \right.\right. \\
&- \left.\left.\frac{1}{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2\right) - d_{ij}^2\right) = \\
&= \frac{1}{2N}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} d_{kj}^2}_{=N\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2} + \frac{1}{2N}\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} d_{ik}^2}_{=N\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2} -
\end{aligned}
$$

Maximum Variance Unfolding: Centering

$$- \underbrace{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{p=1}^{N} d_{kp}^2}_{=N^2\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 =$$

$$= \underbrace{(\frac{1}{2}+\frac{1}{2}-\frac{1}{2}-\frac{1}{2})}_{=0}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = 0$$

$$||\sum_{i=1}^{N}\mathbf{y}_i||_{L2}^2 \Rightarrow \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{K}_{ij} = 0$$

◂ Return