# Machine Learning

## Gaussian Processes and Unsupervised Learning

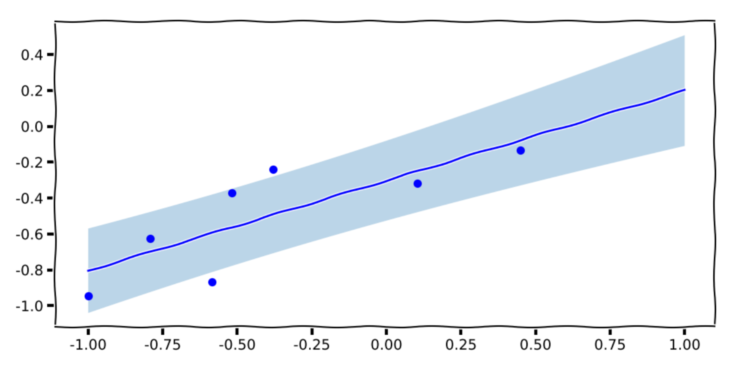
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 22, 2018

http://www.carlhenrik.com

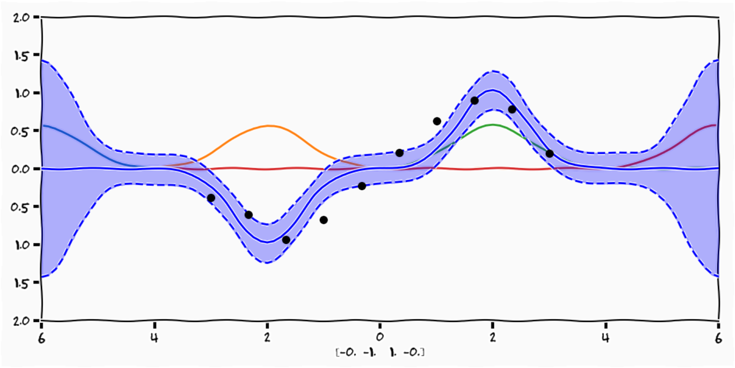
# Introduction

# Regression: Linear



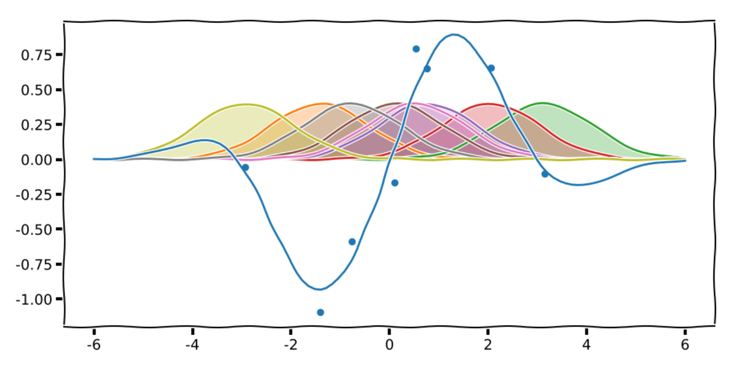

$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$$

# Regression: Linear Basis




$$y_i = \sum_{i=1}^{K} w_k \Phi_k(\mathbf{x}_i)$$

# Regression: Kernel

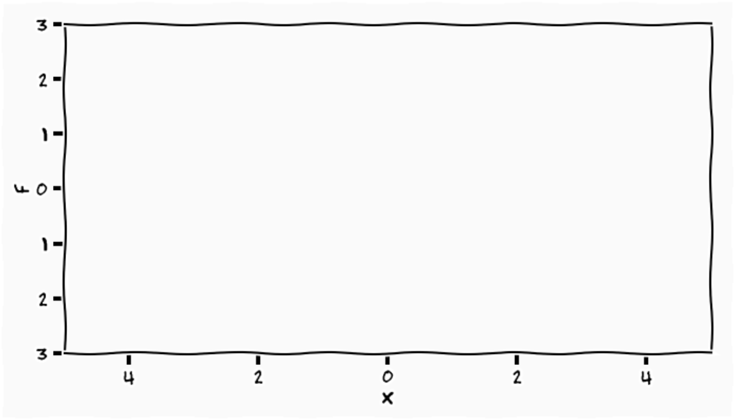


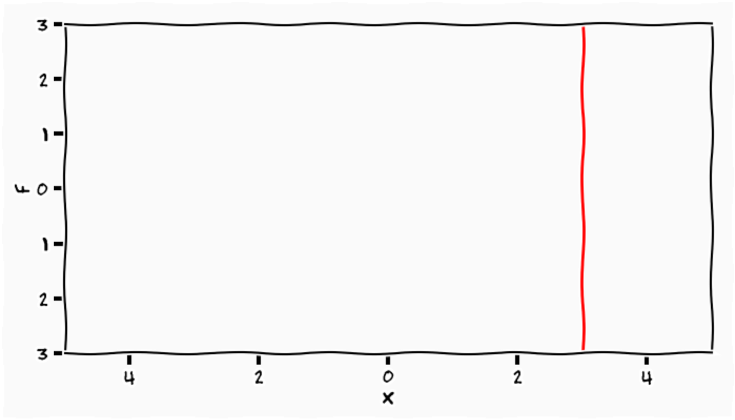

$$y_i = k(\mathbf{x}_i, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1} \mathbf{y}$$

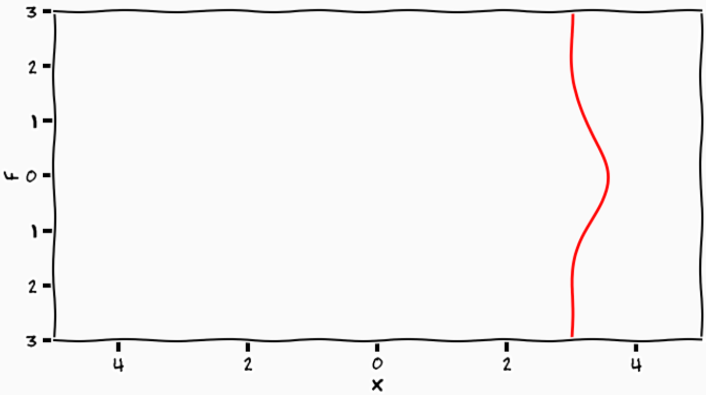
# Regression

- Linear
  - − we are limited by lines
- Basis functions
  - + nonlinear functions
  - − how many basis functions should I have, what should they look like?
  - − prior hard to interpret
- Kernel
  - + complexity set by data
  - − no uncertainty in our estimate

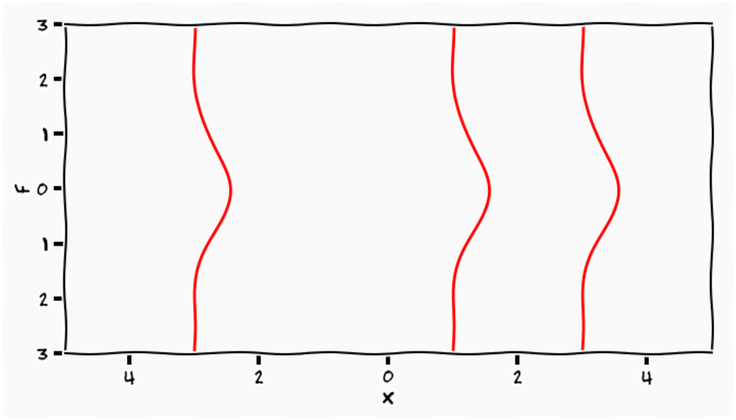
# Gaussian Processes




3
2
1
0
1
2
3
f
4
2
0
2
4
x




$$p(f|x) = \mathcal{N}(\mu(x), \Sigma(x))$$




$$p(f_1, f_2, f_3 | x_1, x_2, x_3)$$

# Gaussian Process: definition

$$
\begin{aligned}
p(f_1, f_2, \ldots, f_N, \ldots | \mathbf{x}, \boldsymbol{\theta}) &= \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x})) \\
&= \mathcal{N}\left( \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) & \cdots \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)
\end{aligned}
$$

# Gaussian Identities

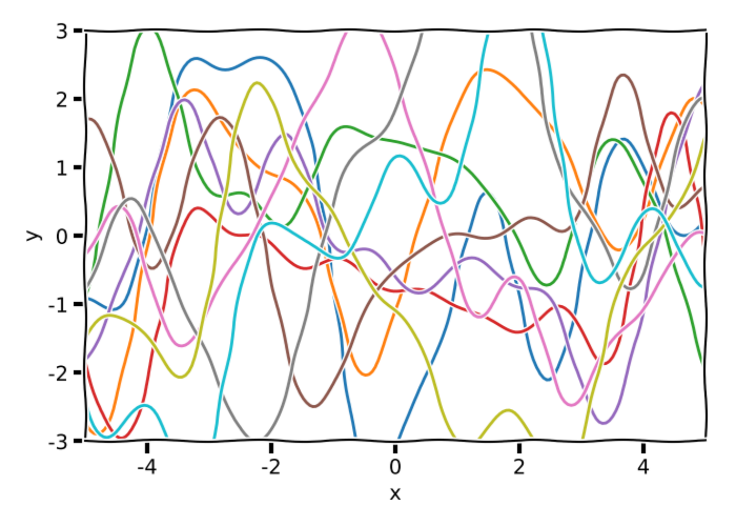
## Marginal

$$p(f_1, f_2 | x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}\right)$$

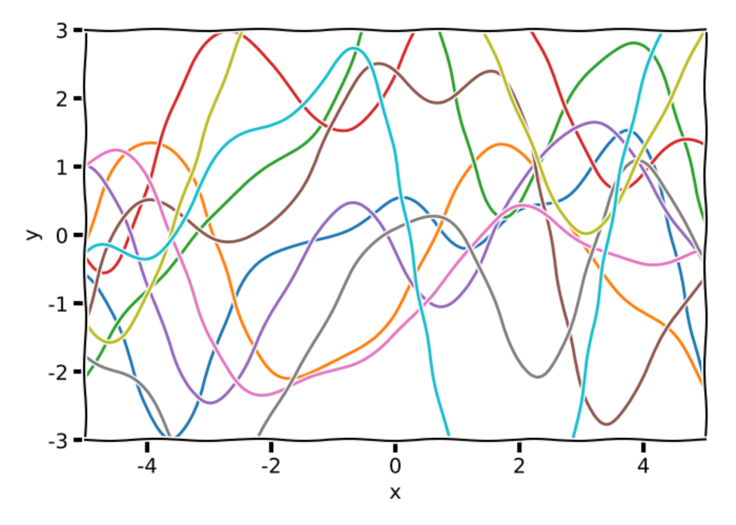
## Conditional

$$\begin{aligned} p(f_1 | f_2, x_1, x_2) = \mathcal{N}(&\mu(x_1) + k(x_1, x_2) k(x_2, x_2)^{-1} (f_2 - \mu(x_2)), \\ &k(x_1, x_1) - k(x_1, x_2) k(x_2, x_2)^{-1} k(x_2, x_1)) \end{aligned}$$

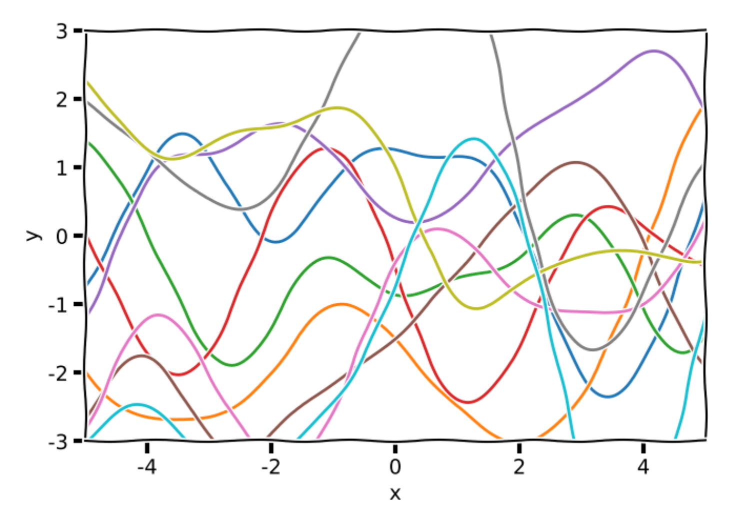
# Sampling



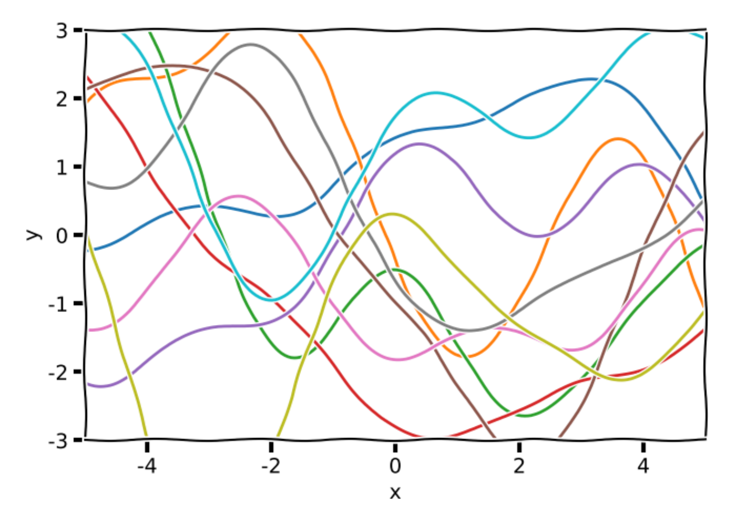
# Sampling



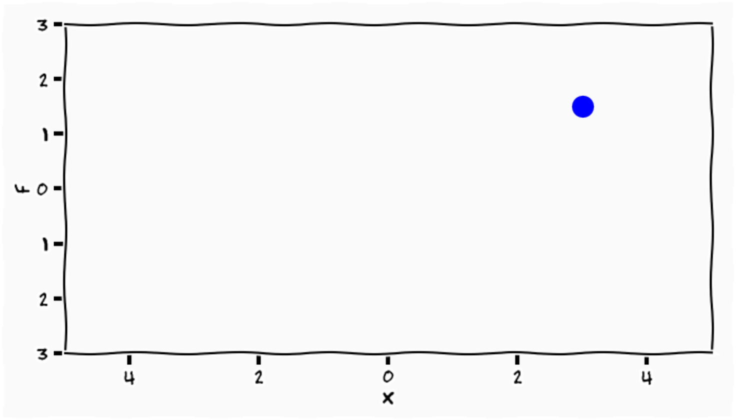
# Sampling



# Sampling



# Gaussian Processes




$$p(f_2|x_2, f_1, x_2) = \mathcal{N}(\mu(x_2, x_1, f_1), \Sigma(x_2, x_1, f_1))$$

# Gaussian Process: definition

$$p(f_1, f_2, \ldots, f_N, \ldots | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$$

$$= \mathcal{N}\left( \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) & \cdots \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)$$

# Gaussian Identities

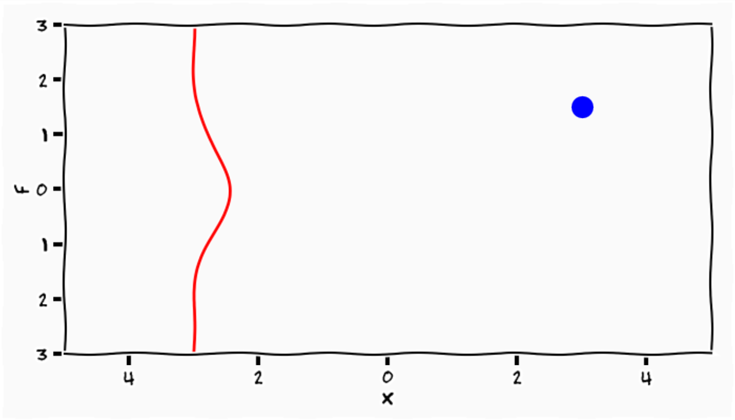
## Marginal

$$p(f_1, f_2 | x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}\right)$$

## Conditional

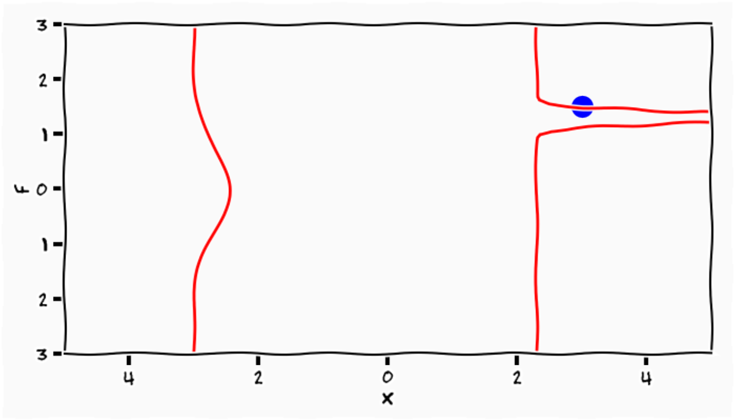
$$\begin{aligned} p(f_1 | f_2, x_1, x_2) = \mathcal{N}(&\mu(x_1) + k(x_1, x_2) k(x_2, x_2)^{-1} (f_2 - \mu(x_2)), \\ &k(x_1, x_1) - k(x_1, x_2) k(x_2, x_2)^{-1} k(x_2, x_1)) \end{aligned}$$

# Gaussian Processes



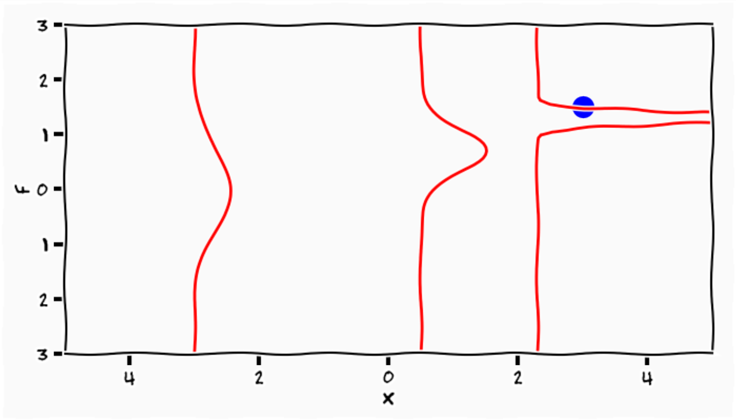

$$p(f_2|x_2, y_1, x_2) = \mathcal{N}(\mu(x_2, x_1, f_1), \Sigma(x_2, x_1, f_1))$$

# Gaussian Processes



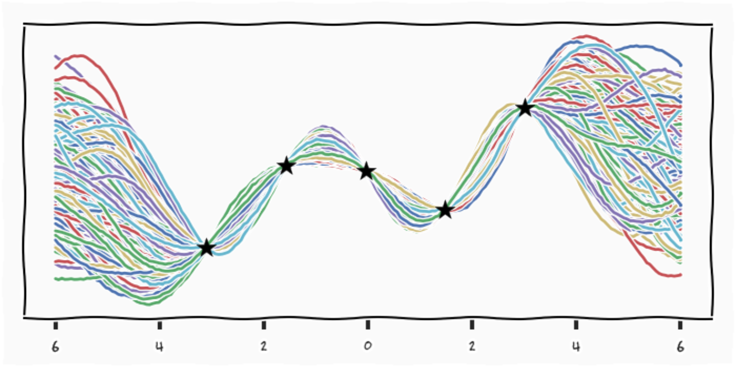

$$p(f_2|x_2, f_1, x_2) = \mathcal{N}(\mu(x_2, x_1, f_1), \Sigma(x_2, x_1, f_1))$$

# Gaussian Processes

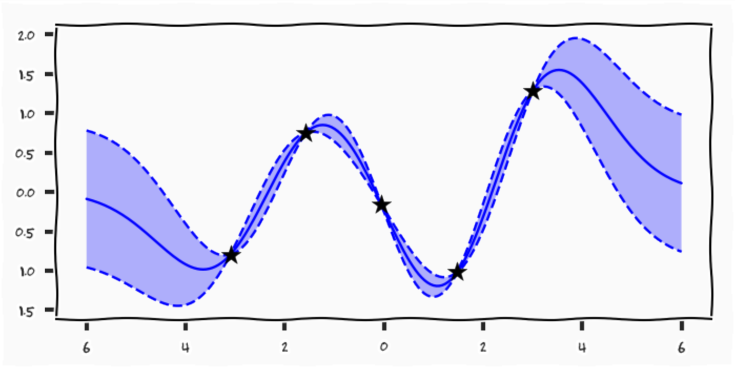



$$p(f_2|x_2, f_1, x_2) = \mathcal{N}(\mu(x_2, x_1, f_1), \Sigma(x_2, x_1, f_1))$$

# Gaussian Processes Posterior



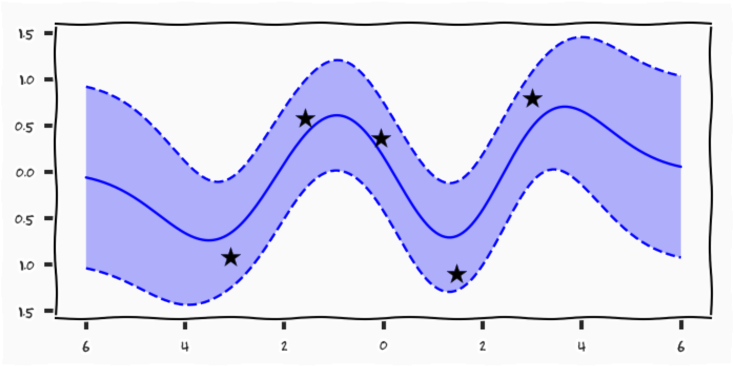
# Gaussian Processes Posterior



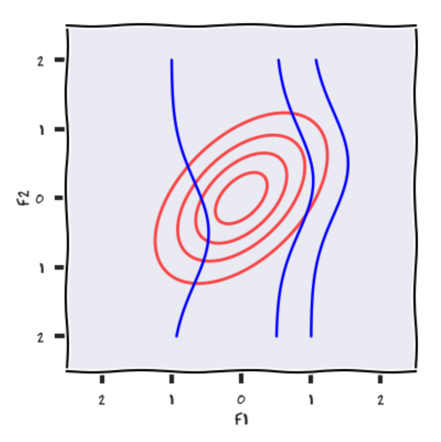
# Gaussian Processes: Noisy observations

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X},\mathbf{X}) + \sigma^2 \mathbf{I} & k(\mathbf{X},\mathbf{x}_*) \\ k(\mathbf{x}_*,\mathbf{X}) & k(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix} \right)$$

$$\begin{aligned} p(f_*|\mathbf{x}_*,\mathbf{X},\mathbf{f}) = \mathcal{N}(& k(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}}(K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I})^{-1}\mathbf{f}, \\ & k(\mathbf{x}_*,\mathbf{x}_*) - k(\mathbf{x}_*,\mathbf{X})^{\mathrm{T}}(K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I})^{-1}K(\mathbf{X},\mathbf{x}_*)) \end{aligned}$$

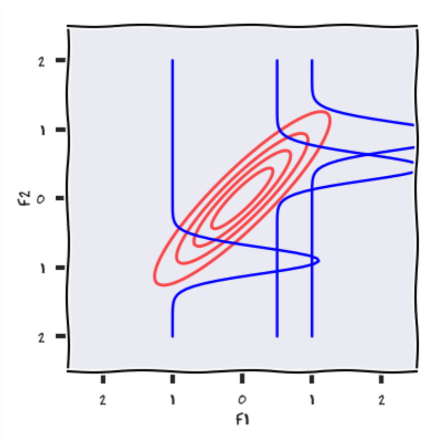

1.5
1.0
0.5
0.0
0.5
1.0
1.5
6
4
2
0
2
4
6

# Conditional Gaussians



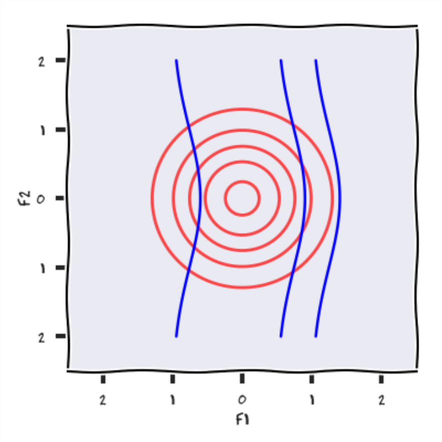

$$\mathcal{N}\left(\underbrace{\begin{bmatrix}0\\0\end{bmatrix}}_{\begin{bmatrix}\mu(x_1)\\ \mu(x_2)\end{bmatrix}},\ \underbrace{\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}}_{\begin{bmatrix}k(x_1,x_1) & k(x_1,x_2)\\ k(x_2,x_1) & k(x_2,x_2)\end{bmatrix}}\right)$$

# Conditional Gaussians




$$\mathcal{N}\left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}}, \underbrace{\begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}}_{\begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}}\right)$$

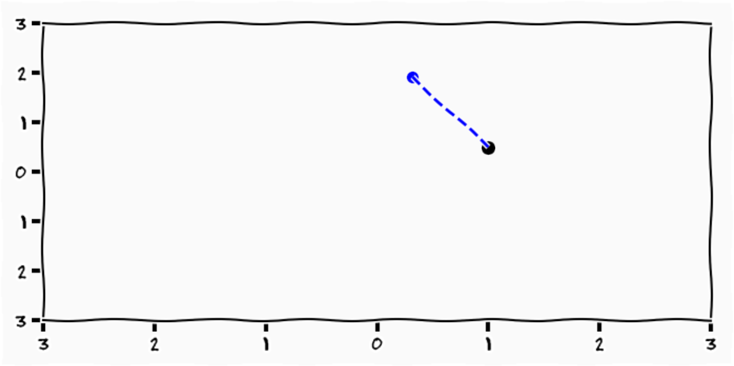
# Conditional Gaussians

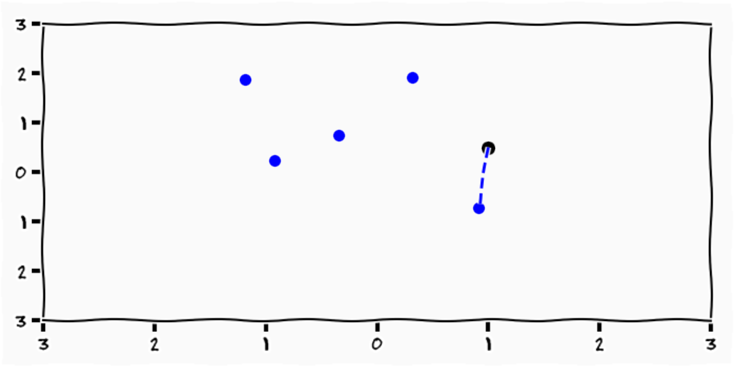



$$\mathcal{N}\left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}}, \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}}\right)$$

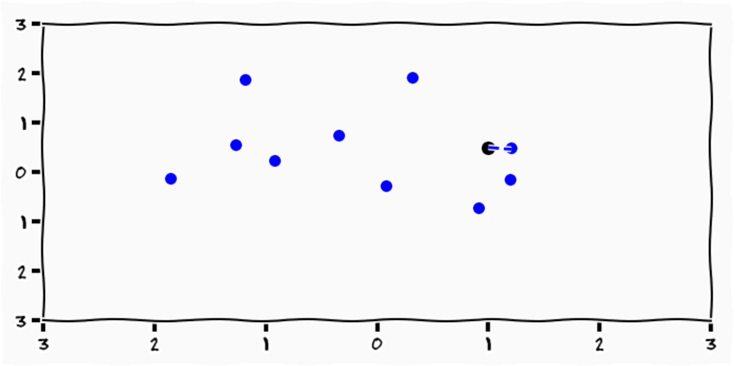
Non-Parametrics??

# Nearest Neighbour

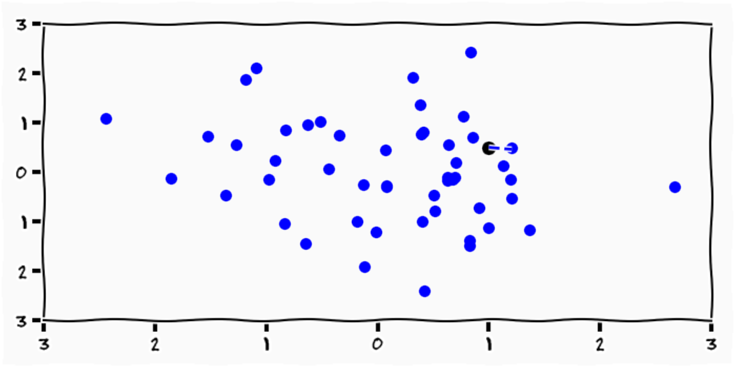



3
2
1
0
1
2
3
3
2
1
0
1
2
3

# Nearest Neighbour



# Nearest Neighbour



# Non-parametrics

- Task of Machine Learning, describe models of data $\mathbf{Y}$

# Non-parametrics

- Task of Machine Learning, describe models of data $\mathbf{Y}$
- A model is a subset of all probability measures on the sample space $\mathcal{Y}$

$$M \subset PM(\mathcal{Y})$$

# Non-parametrics

- Task of Machine Learning, describe models of data $\mathbf{Y}$
- A model is a subset of all probability measures on the sample space $\mathcal{Y}$

$$M \subset PM(\mathcal{Y})$$

- each model is indexed by $\theta$ from the parameter space $\mathcal{T}$

$$M = \{p(\mathbf{Y}|\theta)|\theta \in \mathcal{T}\}$$

# Non-parametrics

- Task of Machine Learning, describe models of data $\mathbf{Y}$
- A model is a subset of all probability measures on the sample space $\mathcal{Y}$

$$M \subset PM(\mathcal{Y})$$

- each model is indexed by $\theta$ from the parameter space $\mathcal{T}$

$$M = \{p(\mathbf{Y}|\theta)|\theta \in \mathcal{T}\}$$

- If $\mathcal{T}$ is
    - finite dimensional space we call this a parametric
    - infinite dimensional space we call this a non-parametric

# Learning

$$
\begin{aligned}
p(f|\mathbf{x}) &= \mathcal{N}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \\
k(\mathbf{x}_i, \mathbf{x}_j) &= \sigma^2 e^{-\frac{2}{\ell^2}\sin^2\left(\pi\frac{|\mathbf{x}_i - \mathbf{x}_j|}{p}\right)} \\
k(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j) \\
k(\mathbf{x}_i, \mathbf{x}_j) &= \frac{2}{\pi}\sin^{-1}\left(\frac{2\mathbf{x}_i^{\mathrm{T}}\Sigma\mathbf{x}_j}{\sqrt{(1 + 2\mathbf{x}_i^{\mathrm{T}}\Sigma\mathbf{x}_i)(1 + 2\mathbf{x}_j^{\mathrm{T}}\Sigma\mathbf{x}_j)}}\right) \\
k(\mathbf{x}_i, \mathbf{x}_j) &= \sigma^2 e^{||\mathbf{x}_i - \mathbf{x}_j||^2/\ell^2}
\end{aligned}
$$

- how do we set the parameters of the co-variance function?

# Marginal Likelihood

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

- We are not interested in $f$ directly
- Marginalise out $f$
- Gaussian likelihood and Gaussian prior $\rightarrow$ Gaussian marginal

# Marginalisation

- Deterministic world

$$\mathbb{E}[y] = \int y p(y) \mathrm{d}y$$

# Marginalisation

- Deterministic world

$$\mathbb{E}[y] = \int y p(y) \mathrm{d}y$$

- Stochastic world

$$\mathbb{E}[p(y)] = \int p(y|x) p(x) \mathrm{d}x$$

# Marginal Likelihood

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

- We are not interested in $f$ directly
- Marginalise out $f$
- Gaussian likelihood and Gaussian prior $\rightarrow$ Gaussian marginal

# Learning

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

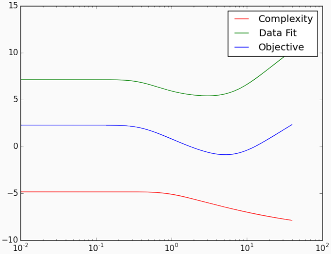
- Type-II Maximum likelihood [1] 3.5.0
- minimise logarithm of marginal likelihood

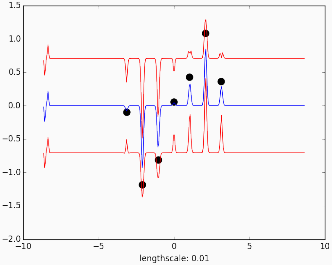
# Learning

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

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

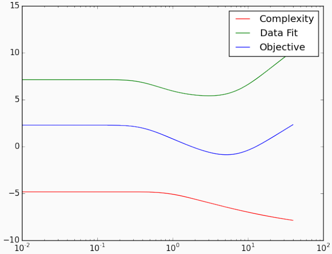
# Learning

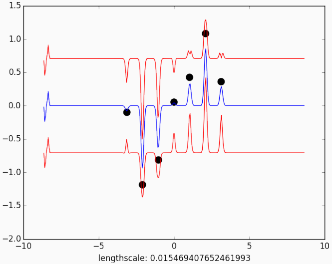



$$\color{lime}{\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}} \quad \color{red}{\frac{1}{2}\log|\mathbf{K}|}$$

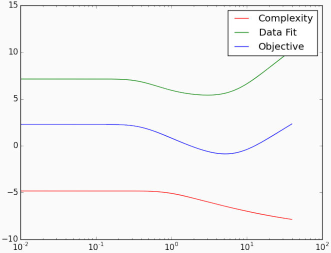
# Learning

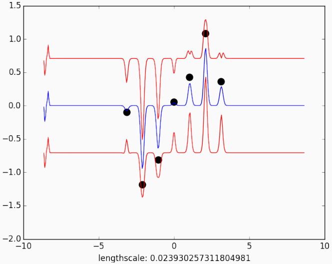



$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$ $\quad$ $\frac{1}{2}\log|\mathbf{K}|$

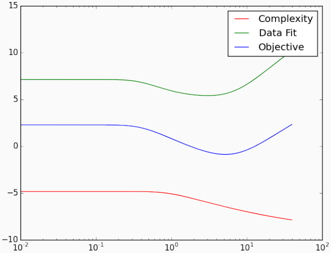
# Learning

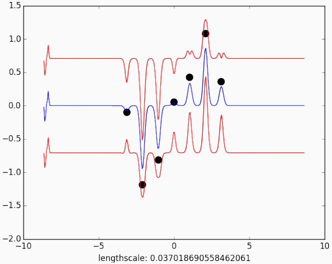



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

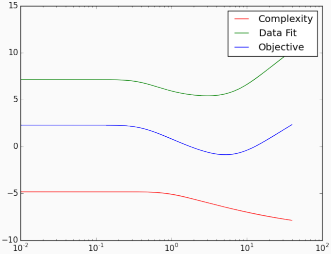
# Learning

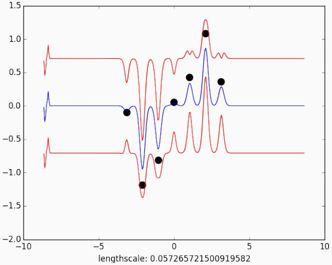



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

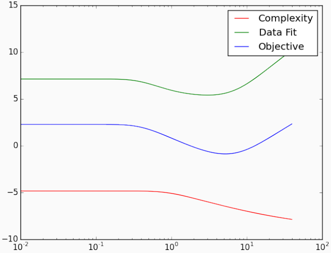
# Learning

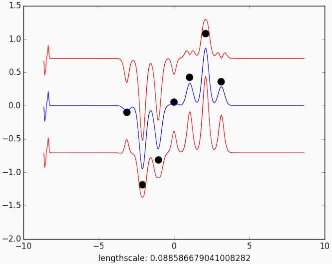



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

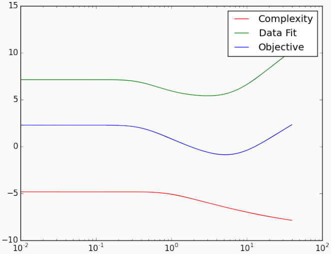
# Learning

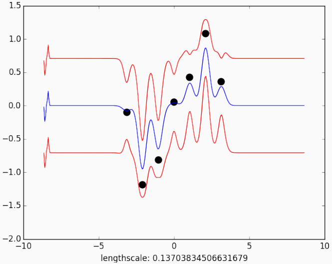



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

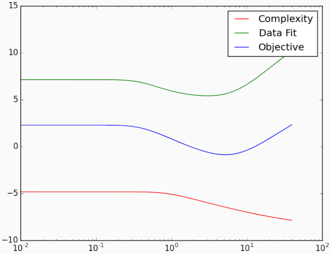
# Learning

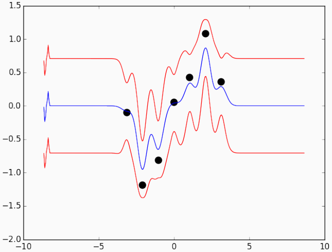



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

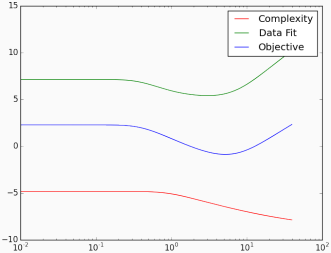
# Learning

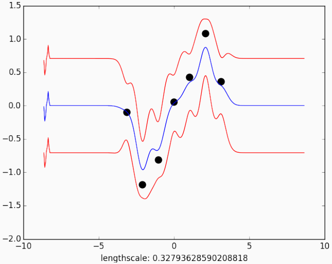



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

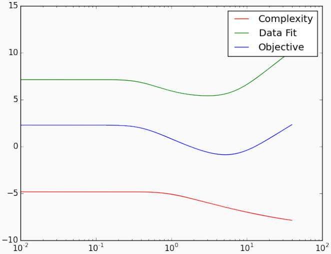
# Learning

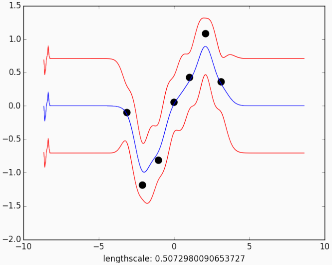



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

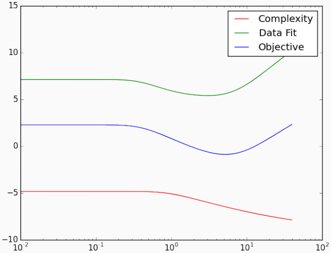
# Learning

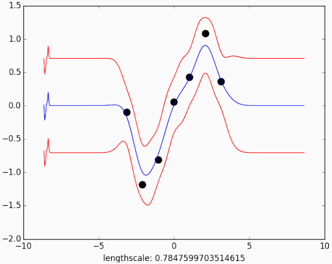


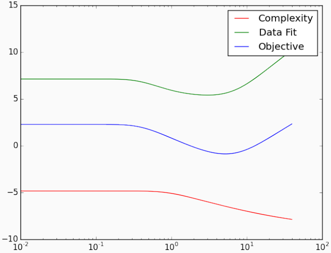

$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log\left|\mathbf{K}\right|$$

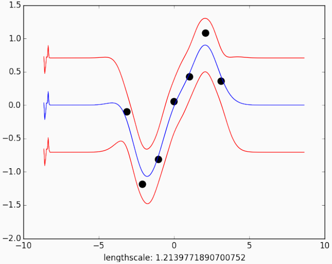
# Learning



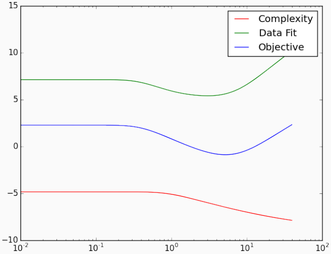

$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

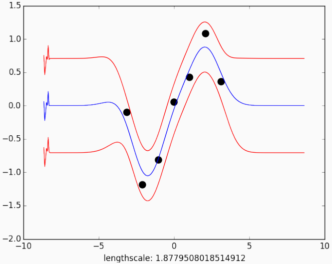



$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$

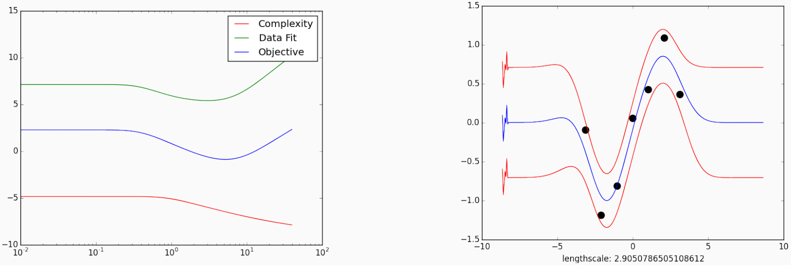
# Learning



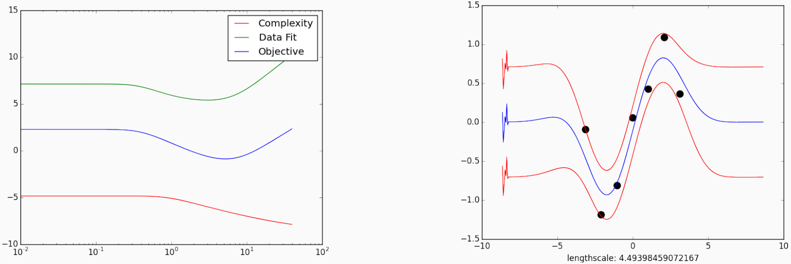

$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

# Learning



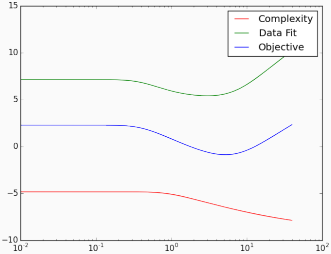

$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

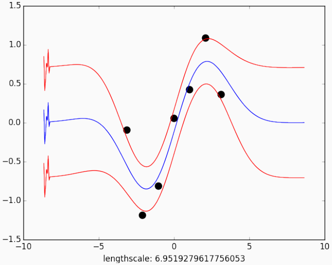
# Learning




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

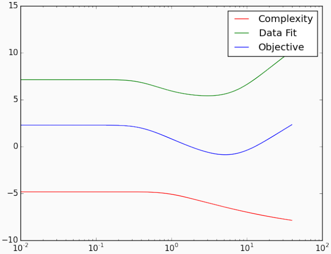
# Learning

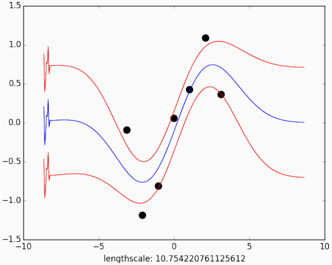



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

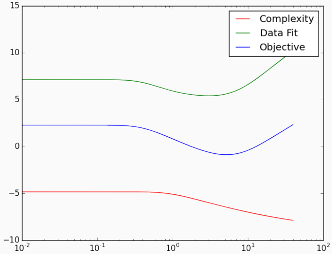
# Learning

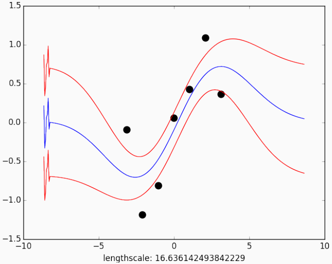



$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$

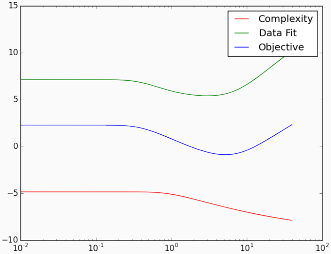
# Learning

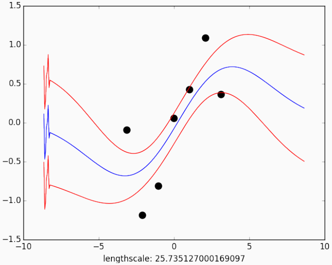



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

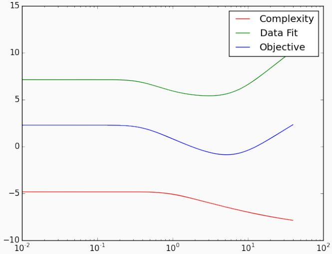
# Learning

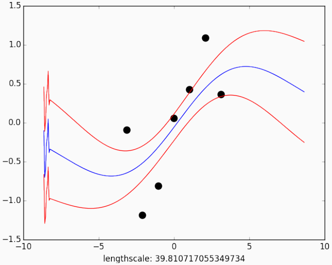



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

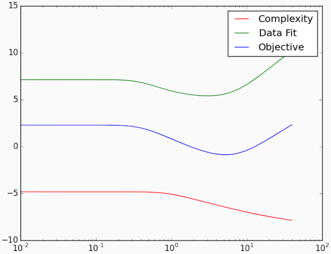
# Learning

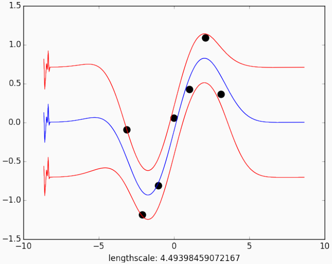



$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

# Learning



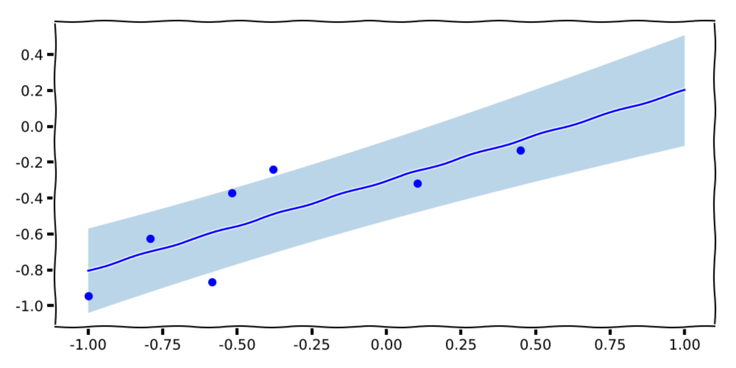

$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$ $\frac{1}{2}\log|\mathbf{K}|$

# Gaussian Processes

- completely specified by mean and covariance function
- mean and covariance are functions of input variable
- every instantiation of the function is jointly Gaussian
    - conditional and marginal distribution trivial
- very flexible
    - covariance function can encode any behaviour
- infer parameters through Type-II maximum likelihood

# Unsupervised Learning

# Regression: Linear




$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$$

## Supervised Learning

$$y_i = f(x_i)$$

- learn relationship $f(\cdot)$ between pairs of data $x_i$ and $y_i$

# Machine Learning

## Supervised Learning

$$y_i = f(x_i)$$

- learn relationship $f(\cdot)$ between pairs of data $x_i$ and $y_i$

## Unsupervised Learning

$$y_i = f(x_i)$$

- learn a representation $\mathbf{X}$ from data $\mathbf{Y}$

# Strength of Priors

$$y = f(x)$$

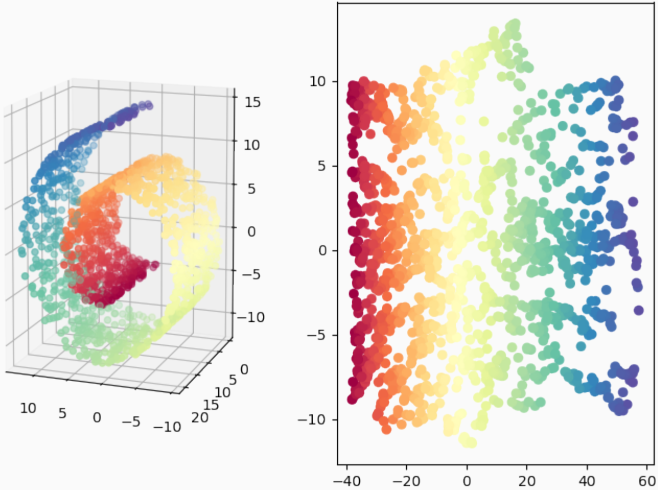
- given input output pairs we have made assumptions about f
- from data we can update our assumption
- can we push this further?

# Unsupervised learning

$$y = f(x)$$

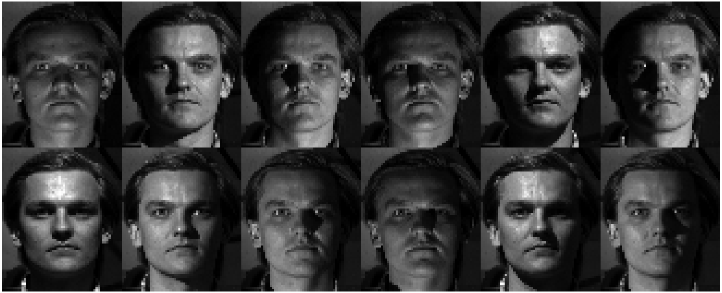
- In unsupervised learning we are given only output
- Input is *latent*
- Task: recover both $f$ and $x$

# Manifold



# Latent Variable Models

# Latent Variable Models



**output data** $\mathbf{y} \in \mathbb{R}^{256 \times 256} \rightarrow 65536$ dimensions

**input** location on sphere $\rightarrow 3$ dimensions

**manifold** images lie on a 3 dimensional surface in 65536 dimensions

# Linear Latent Variable Models [1] 12.2

- Observed data

$$\mathbf{x} \in \mathbb{R}^D$$

- Latent variable

$$\mathbf{z} \in \mathbb{R}^M$$

- Mapping

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

- Likelihood: make noise assumption $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

$$p(\mathbf{x}|\mathbf{z}, \mathbf{W}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

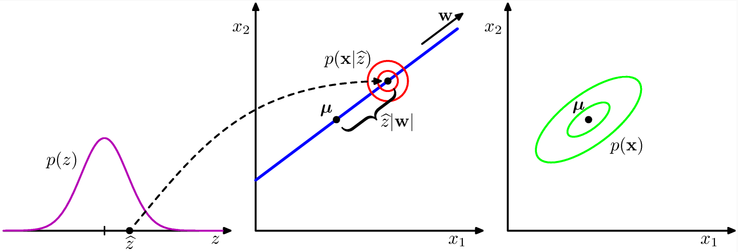
- Prior ?

# Linear Latent Variable Models

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

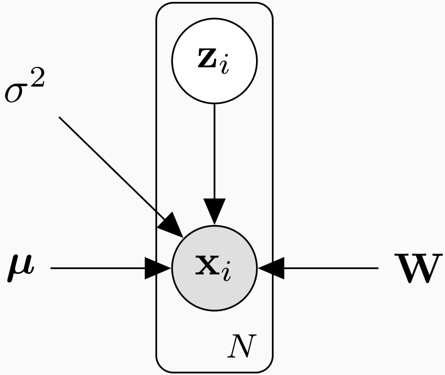
- marginalise out both $\mathbf{W}$ and $\mathbf{z}$ is intractable
- marginalise out one and optimise the other
- $\mathbf{W} \in \mathbb{R}^{D \times M}$ and $\mathbf{z} \in \mathbb{R}^{M \times N}$
- $N$ commonly larger than $D \Rightarrow$ integrate out $\mathbf{z}$

# Principal Component Analysis [1] Figure 12.9



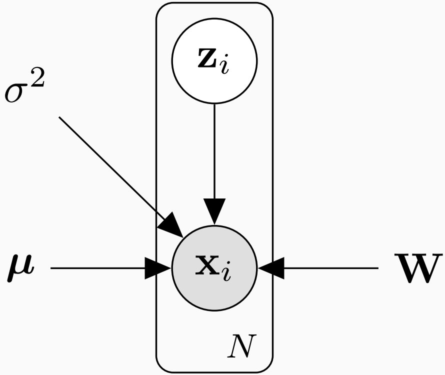

$$p(z) = \mathcal{N}(z|\mathbf{0}, \boldsymbol{I})$$

# Graphical Model




$$p(\mathsf{x}, \mathsf{z}) = p(\mathsf{x}|\mathsf{z})p(\mathsf{z})$$

# Graphical Model




$$p(\mathsf{x}, \mathsf{z}|\mathsf{W}, \boldsymbol{\mu}, \sigma^2) = p(\mathsf{x}|\mathsf{z}, \mathsf{W}, \boldsymbol{\mu}, \sigma^2)p(\mathsf{z})$$

# Marginal distribution

$$p(\mathsf{x}|\mathsf{W}) = \int p(\mathsf{x}|\mathsf{z}, \mathsf{W})p(\mathsf{z})\mathrm{d}\mathsf{z} = \mathcal{N}(\mathsf{x}|\boldsymbol{\mu}, \mathsf{C})$$

- Gaussian distribution closed under linear transformation (interesting proof)

# Marginal distribution

$$p(\mathsf{x}|\mathsf{W}) = \int p(\mathsf{x}|\mathsf{z},\mathsf{W})p(\mathsf{z})\mathrm{d}\mathsf{z} = \mathcal{N}(\mathsf{x}|\boldsymbol{\mu},\mathsf{C})$$
$$= \mathcal{N}(\mathsf{x}|\boldsymbol{\mu},\mathsf{W}\mathsf{W}^{\mathrm{T}} + \sigma^2\mathsf{I})$$

- Gaussian distribution closed under linear transformation (interesting proof)

# Maximum Likelihood [1] Ch. 12.2.1

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2)$$

- find stationary with respect to each variable gives Maximum likelihood solution to $\mathbf{W}$, $\boldsymbol{\mu}$ and $\sigma^2$
- *In the assignment we make it easier and take derivatives instead and optimise*

# Maximum Likelihood [1] Ch. 12.2.1

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$$

- find stationary with respect to each variable gives Maximum likelihood solution to $\mathbf{W}$, $\boldsymbol{\mu}$ and $\sigma^2$
- *In the assignment we make it easier and take derivatives instead and optimise*

# Maximum Likelihood [1] Ch. 12.2.1

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$$

$$= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}|$$

$$- \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I})^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- find stationary with respect to each variable gives Maximum likelihood solution to $\mathbf{W}$, $\boldsymbol{\mu}$ and $\sigma^2$
- *In the assignment we make it easier and take derivatives instead and optimise*

# Posterior

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}), \sigma^2(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I})^{-1})$$

- Gaussian likelihood and Gaussian prior $\rightarrow$ Gaussian posterior

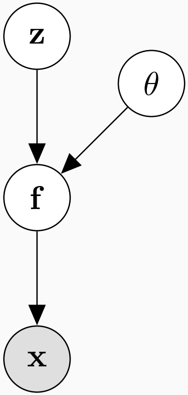
# Principal Component Analysis

- You might have seen this explained in a different way
  - *Retain variance*
  - *Error minimisation*
- These provides the same solution as the maximum likelihood but solved by an eigenvalue problem
- Do not provide intuition as it doesn't state assumptions

# Question 15-22

You now have all the material to finish the assignment!

# Non-linear Latent variable model




$$p(\mathbf{x}|\mathbf{z},\theta) = \int p(\mathbf{x}|\mathbf{f})p(\mathbf{f}|\mathbf{z},\theta)\mathrm{d}\mathbf{f}$$

# Demo

Font Demo

# Summary

# Summary

- Type II Maximum likelihood
- As long as I make assumptions I can learn from data
- Unsupervised learning, just the same, just a prior instead of observations
- Tomorrow and next three lectures

eof

# References

Christopher M. Bishop.
***Pattern Recognition and Machine Learning (Information Science and Statistics)***.
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.