Machine Learning

Unsupervised Learning

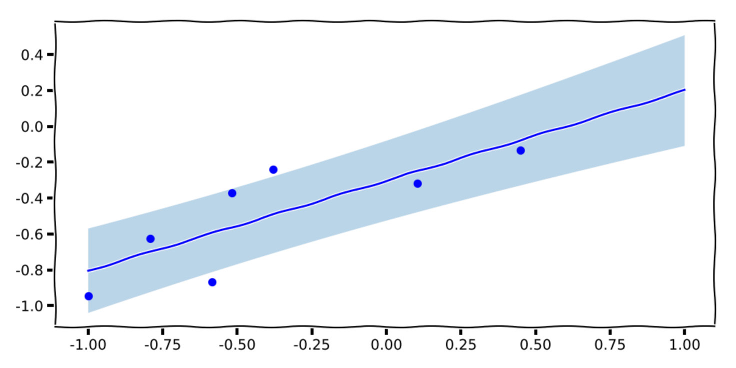
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 22, 2019

http://www.carlhenrik.com

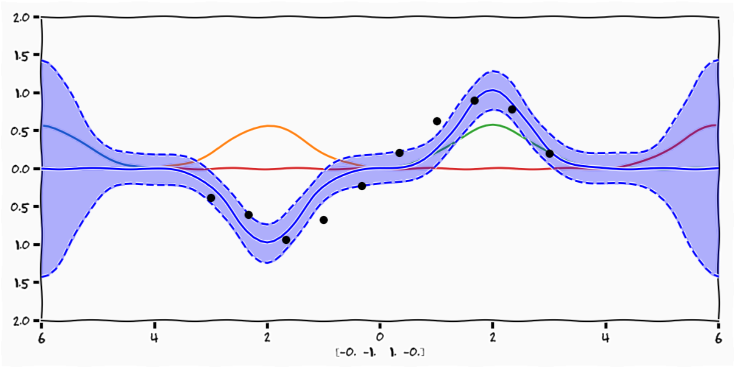
Introduction

Regression: Linear



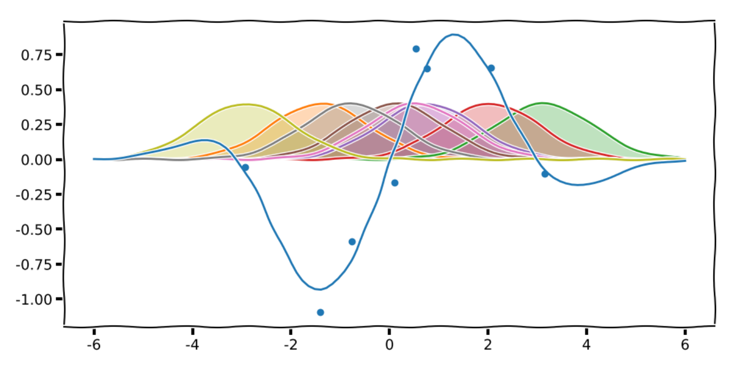
$$y_i = \mathbf{w}^{\mathrm{T}} \mathbf{x}_i$$

Regression: Linear Basis



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

Regression: Kernel

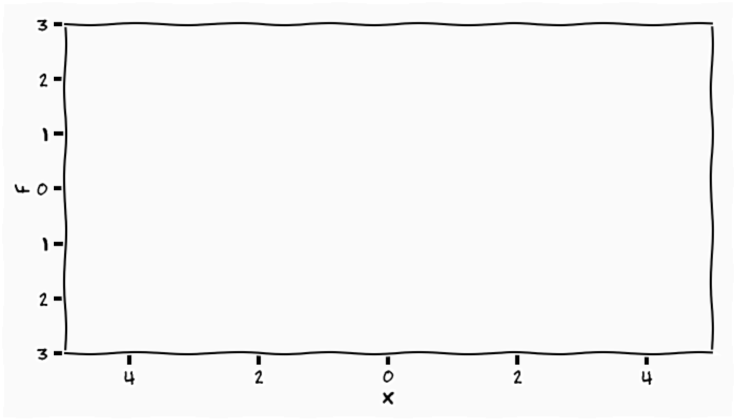


$$y_i = k(\mathbf{x}_i, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \lambda \mathbf{I})^{-1}\mathbf{y}$$

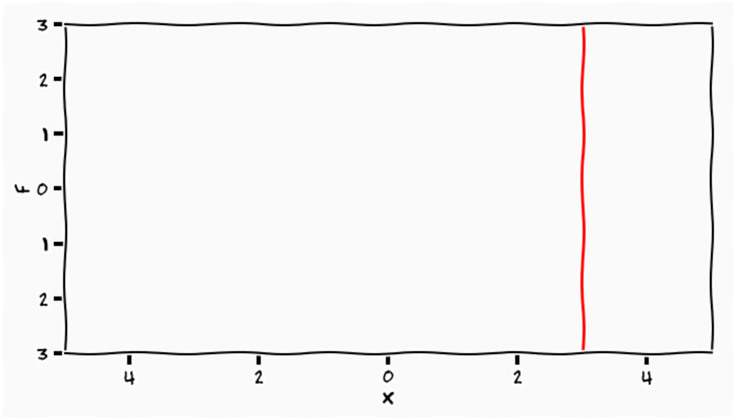
Regression

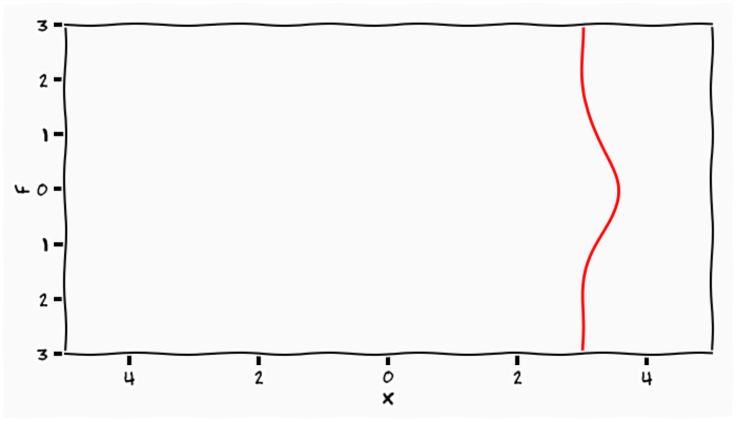
- Linear
  - – we are limited by lines
- Basis functions
  - + nonlinear functions
  - – how many basis functions should I have, what should they look like?
  - – prior hard to interpret
- Kernel
  - + complexity set by data
  - – no uncertainty in our estimate

Gaussian Processes



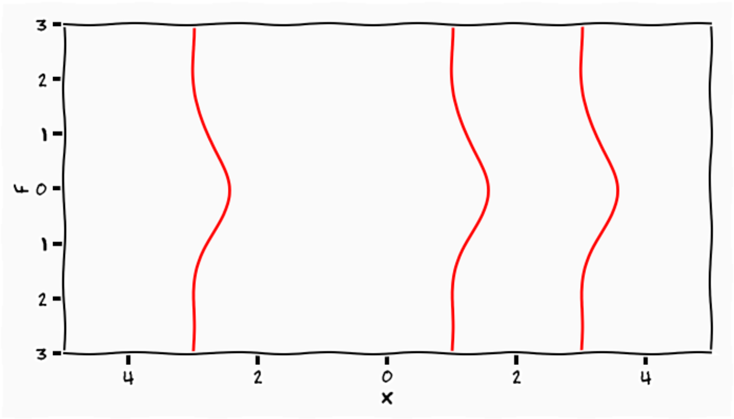
Gaussian Processes





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

Gaussian Processes



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

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

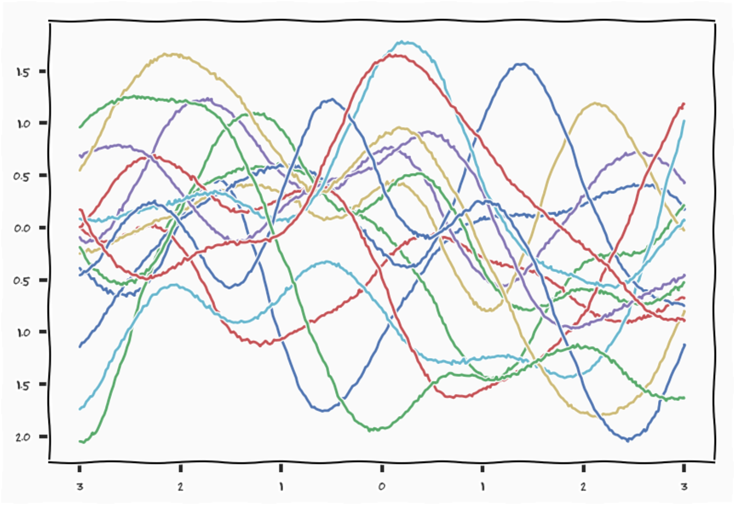
Marginal

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

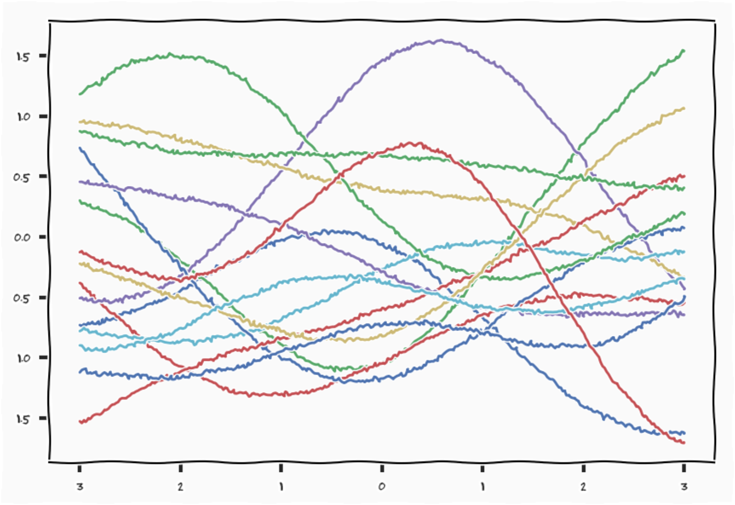
Conditional

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

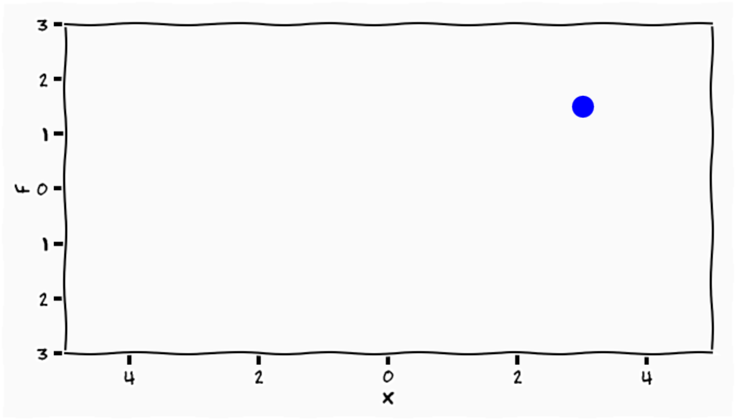
Sampling



Sampling

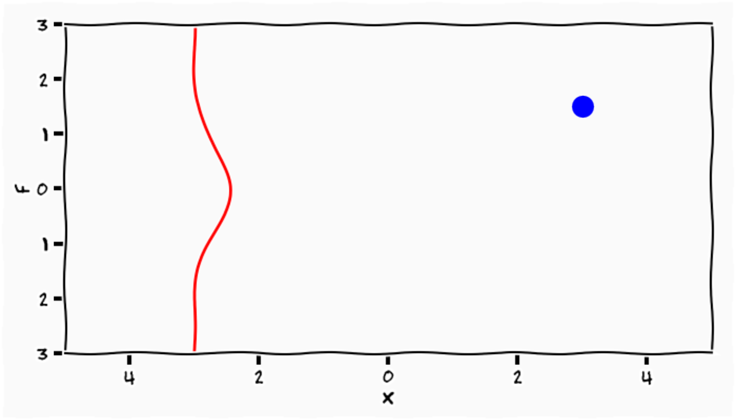


Gaussian Processes



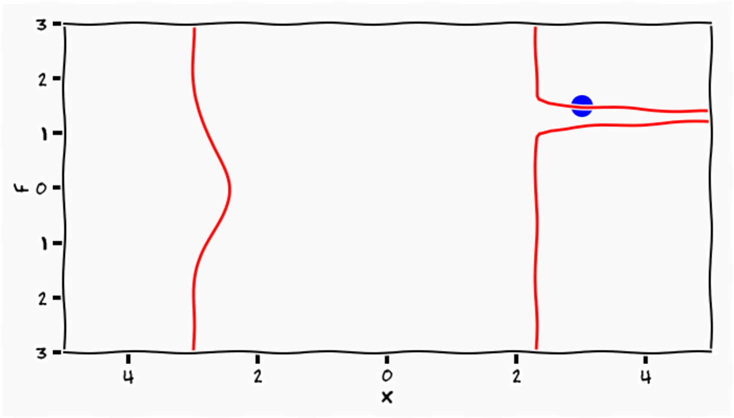
$$p(f_2|x_2, f_1, x_1) = \mathcal{N}(\mu(x_2, x_1, f_1), \Sigma(x_2, x_1, f_1))$$

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

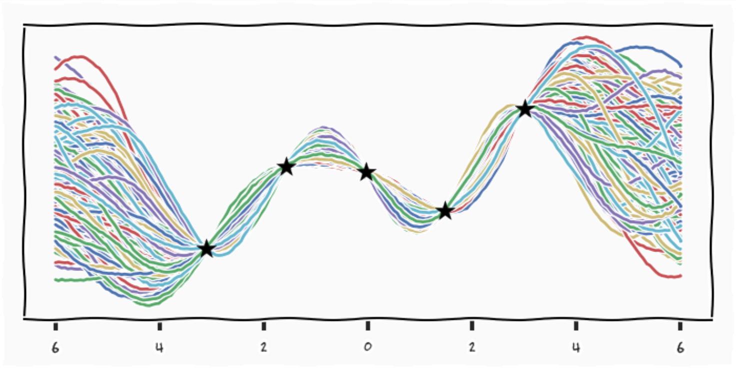


$$p(f_2 | x_2, f_1, x_2) = \mathcal{N}(\mu(x_2, x_1, f_1), \Sigma(x_2, x_1, f_1))$$

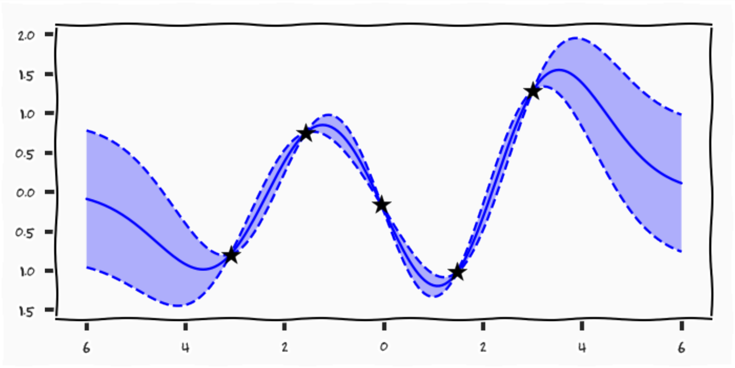
Gaussian Processes

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

Gaussian Processes Posterior



Gaussian Processes Posterior

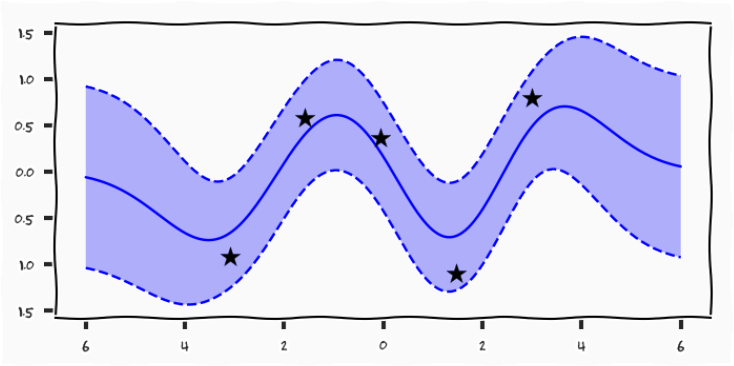


Gaussian Processes: Noisy observations

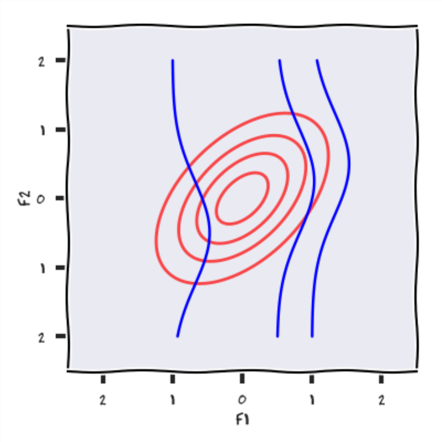
$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k(\mathbf{X},\mathbf{X}) + \sigma^2 \mathbf{I} & k(\mathbf{X},\mathbf{x}_*) \\ k(\mathbf{x}_*,\mathbf{X}) & k(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix}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

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

Gaussian Processes

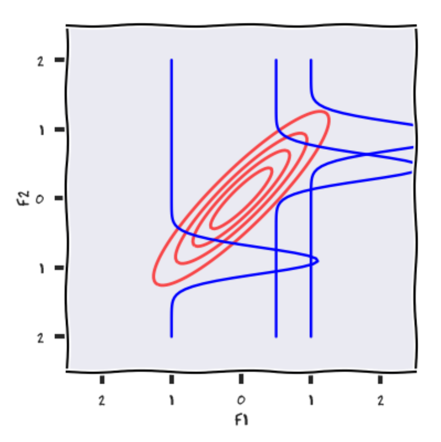


Conditional Gaussians



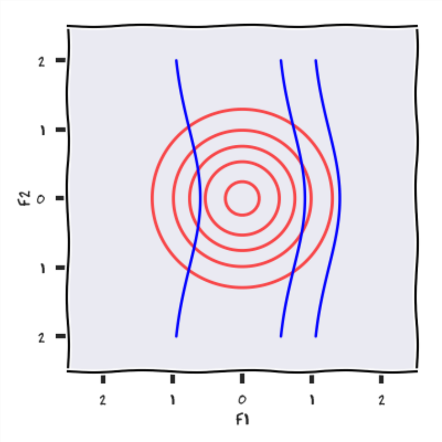
$$\mathcal{N}\left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}}, \underbrace{\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}}_{\begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}} \right)$$

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

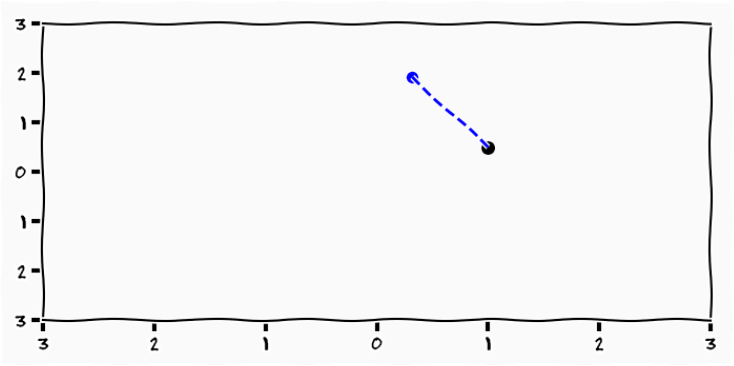


$$\mathcal{N}\left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \end{bmatrix}}, \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{bmatrix}}\right)$$

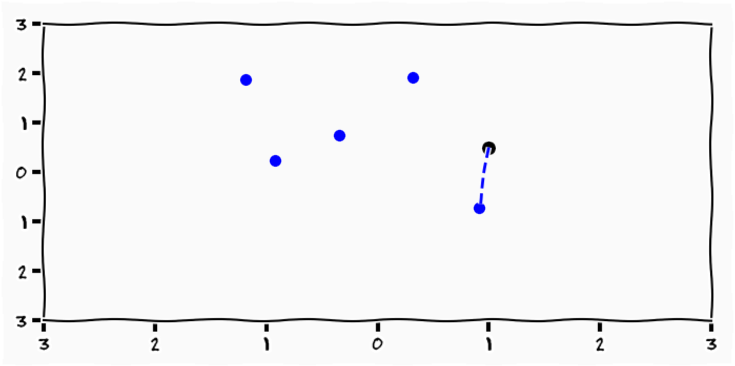
Non-parametrics

Non-Parametrics??

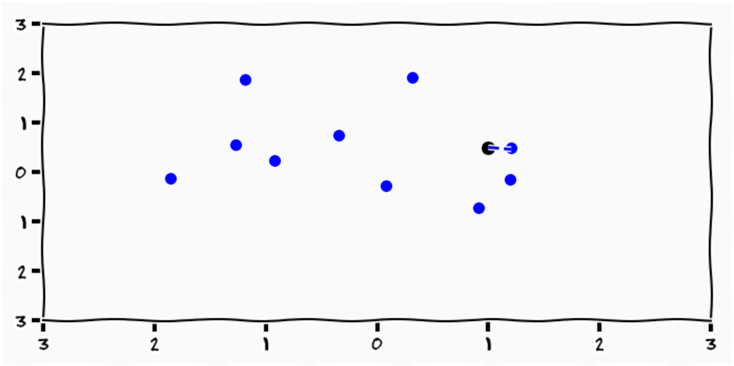
Nearest Neighbour



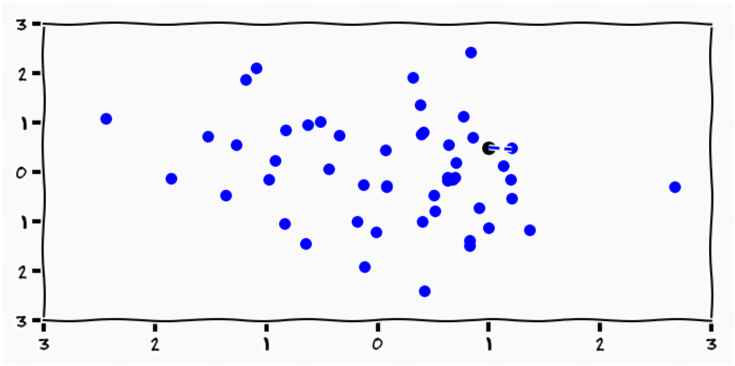
Nearest Neighbour



Nearest Neighbour



Nearest Neighbour



Learning

$$p(f|\mathsf{x}) = \mathcal{N}(\mu(\mathsf{x}), k(\mathsf{x}, \mathsf{x}'))$$

$$k(\mathsf{x}_i, \mathsf{x}_j) = \sigma^2 e^{-\frac{2}{\ell^2}\sin^2\left(\pi \frac{|\mathsf{x}_i - \mathsf{x}_j|}{p}\right)}$$

$$k(\mathsf{x}_i, \mathsf{x}_j) = (\mathsf{x}_i^{\mathrm{T}} \mathsf{x}_j)$$

$$k(\mathsf{x}_i, \mathsf{x}_j) = \frac{2}{\pi}\sin^{-1}\left(\frac{2\mathsf{x}_i^{\mathrm{T}}\Sigma\mathsf{x}_j}{\sqrt{(1 + 2\mathsf{x}_i^{\mathrm{T}}\Sigma\mathsf{x}_i)(1 + 2\mathsf{x}_j^{\mathrm{T}}\Sigma\mathsf{x}_j)}}\right)$$

$$k(\mathsf{x}_i, \mathsf{x}_j) = \sigma^2 e^{||\mathsf{x}_i - \mathsf{x}_j||^2/\ell^2}$$

- 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

Learning

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

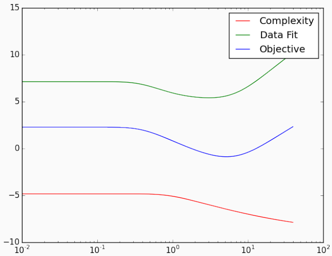
- Type-II Maximum likelihood [1] 3.5.0
- minimise logarithm of marginal likelihood

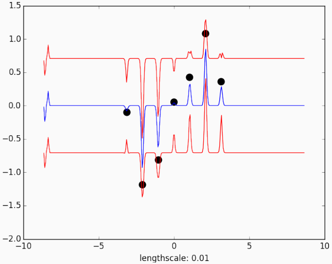
Learning

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

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

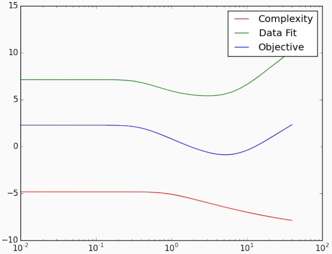
Learning

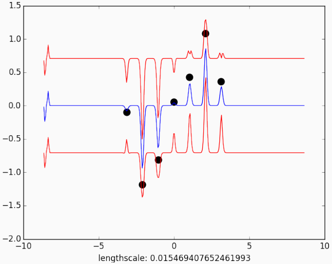




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

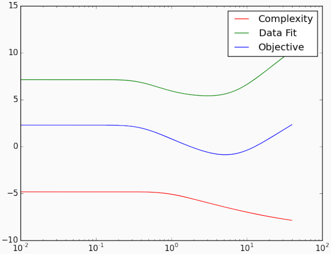
Learning

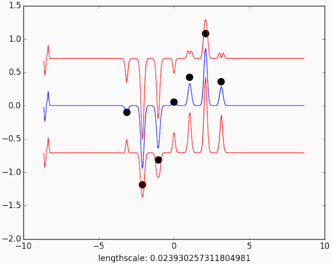




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

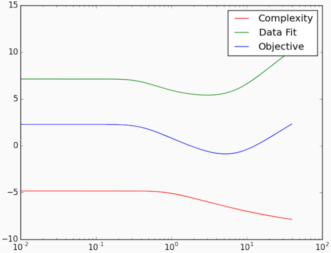
Learning

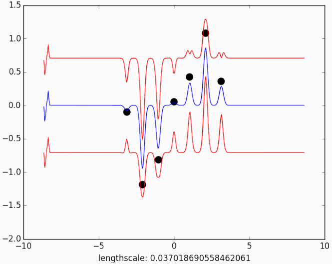




$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$

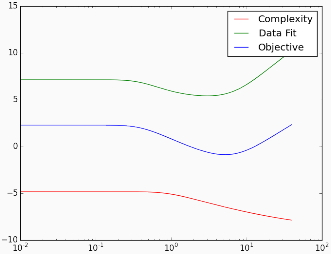
Learning

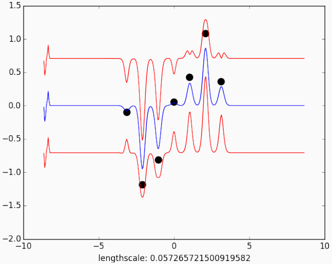




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

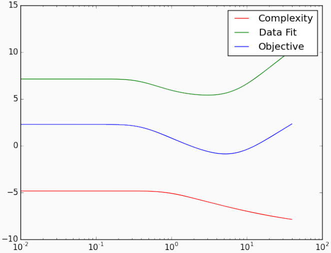
Learning

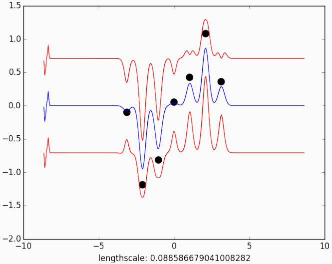




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

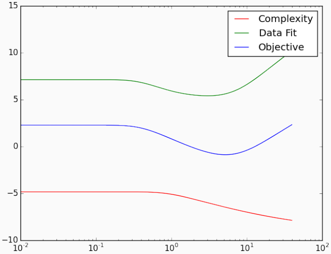
Learning

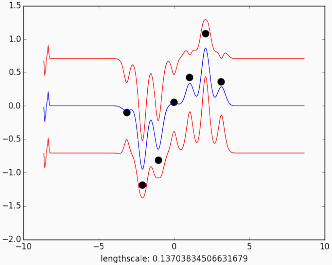




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

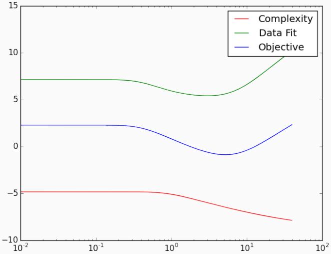
Learning

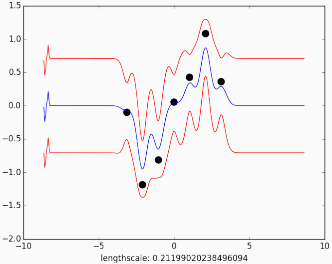




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

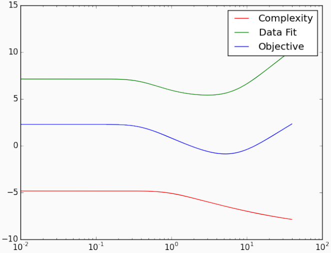
Learning

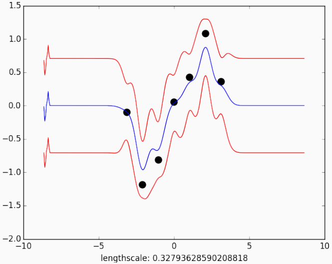




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

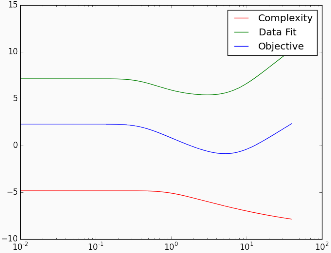
Learning

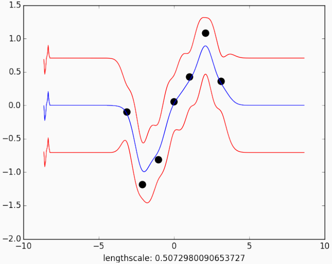




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

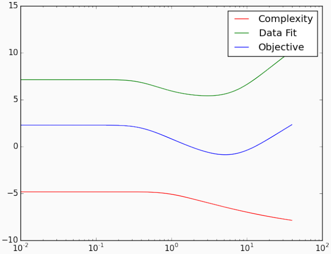
Learning

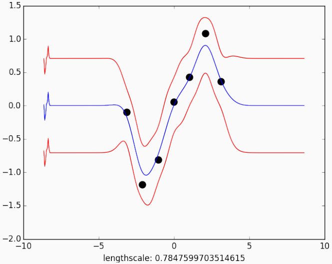




$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$ $\quad$ $\frac{1}{2}\log|\mathbf{K}|$

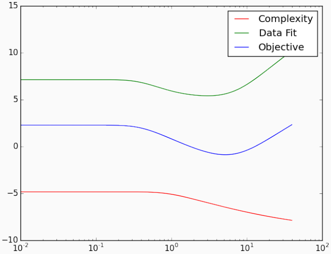
Learning

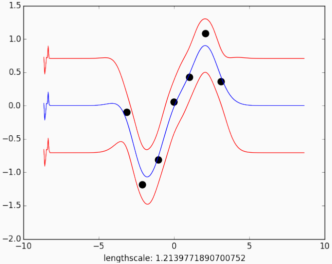




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

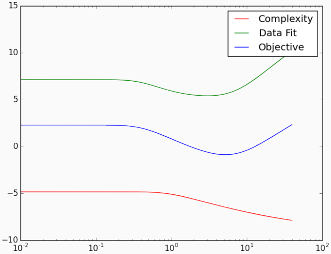
Learning

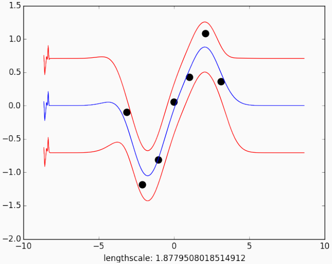




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

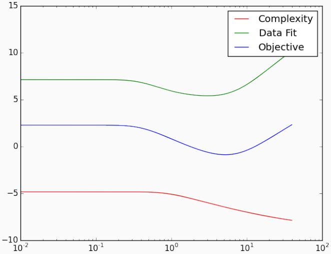
Learning

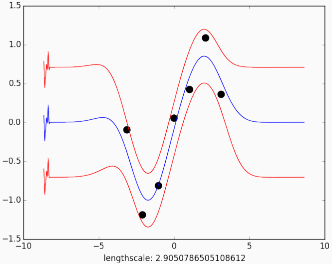




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

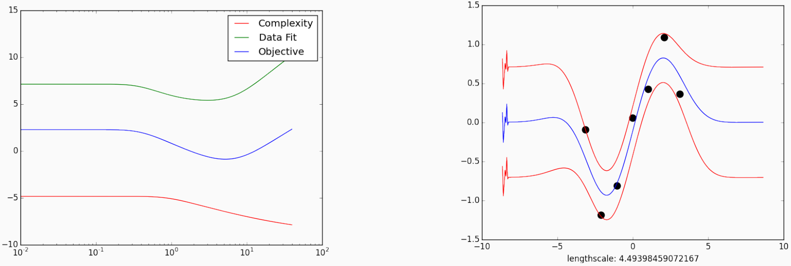
Learning





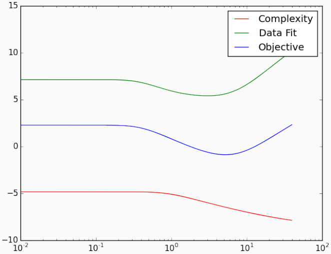
$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$

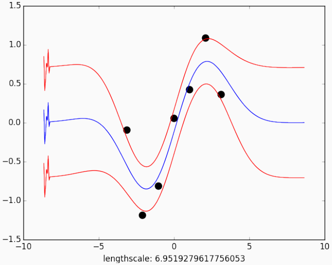
Learning



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

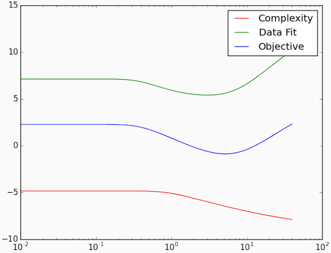
Learning

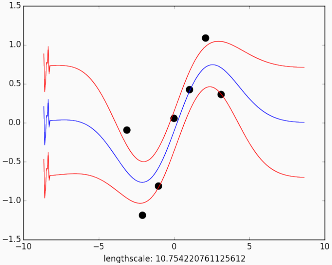




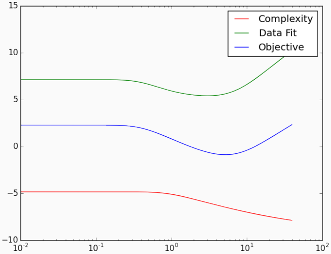
$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

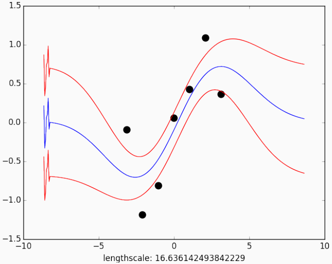
Learning



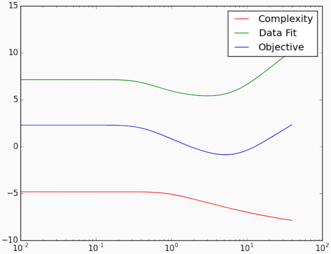


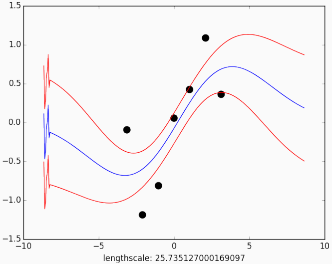
$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$





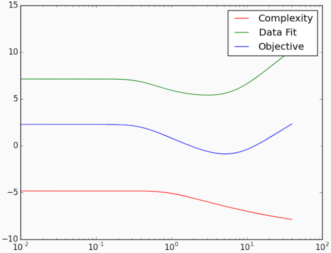
$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$ $\quad$ $\frac{1}{2}\log|\mathbf{K}|$

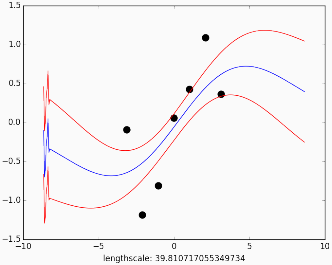




$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \frac{1}{2}\log|\mathbf{K}|$$

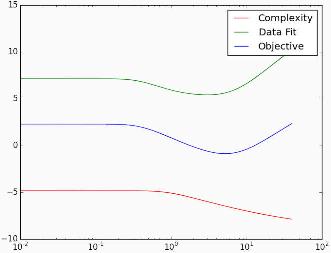
Learning

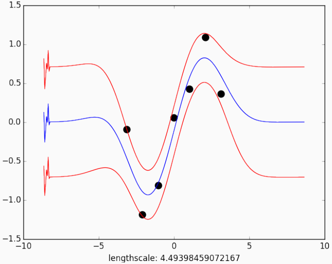




$$\tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \qquad \tfrac{1}{2}\log|\mathbf{K}|$$

Learning





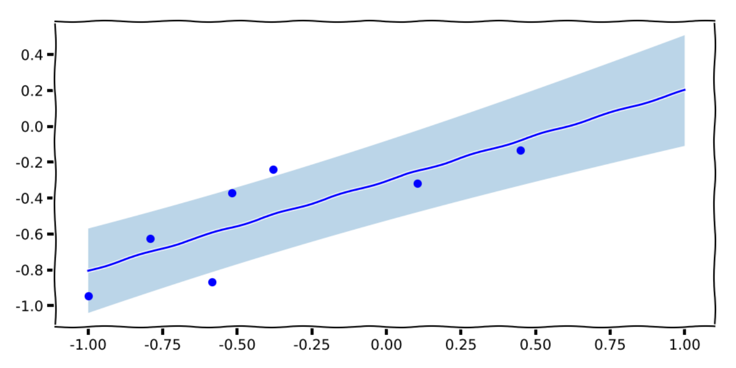
$$\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} \quad \frac{1}{2}\log|\mathbf{K}|$$

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

Supervised Learning

$$y_i = f(x_i)$$

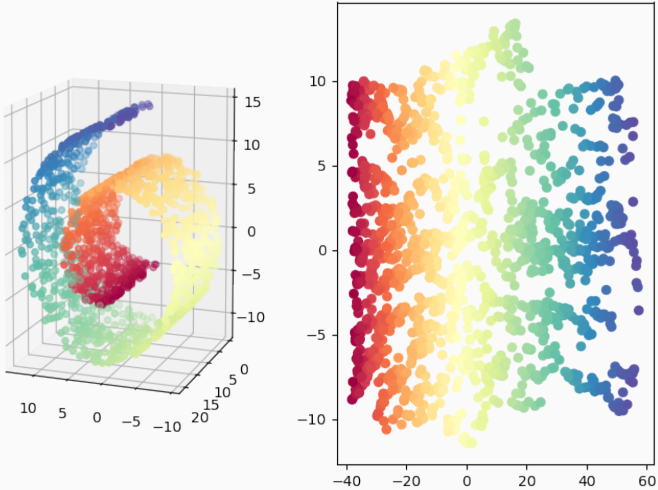
- learn relationship $f(\cdot)$ between pairs of data x_i and y_i

Unsupervised Learning

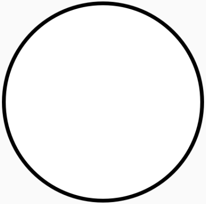
$$y_i = f(x_i)$$

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

Manifold



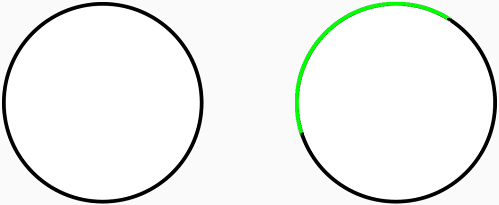
Manifolds



"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[1]

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

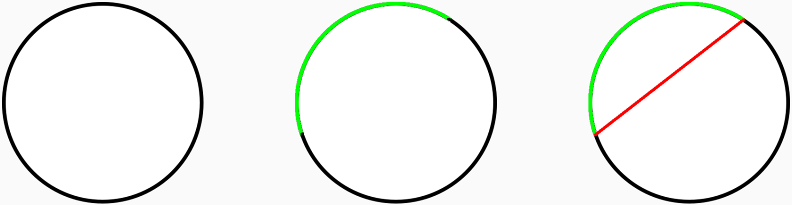
Manifolds



"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[1]

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

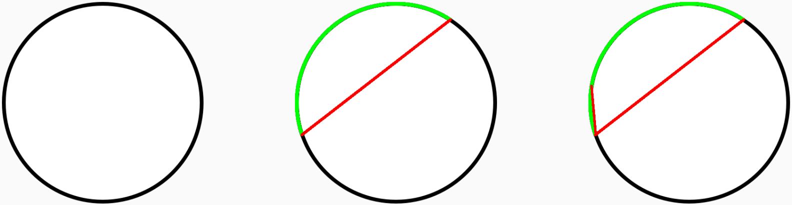
Manifolds



"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[1]

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

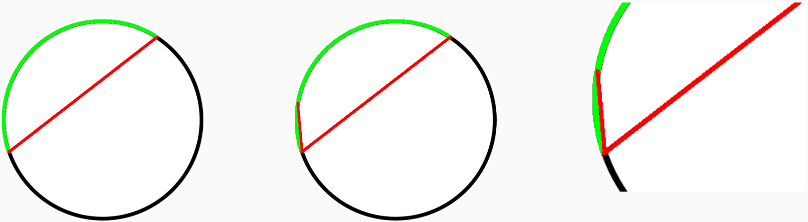
Manifolds



"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[1]

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

Manifolds

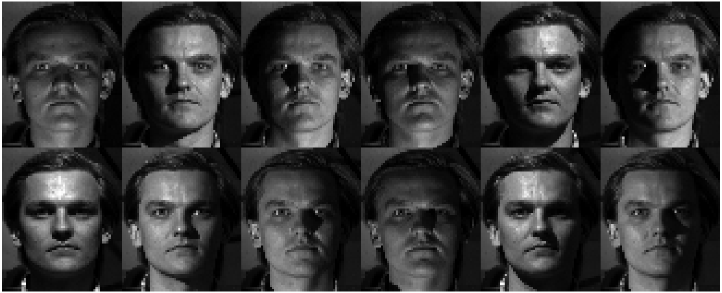


"In mathematics, a manifold is a topological space that near each point resembles Euclidean space"[1]

[1] http://en.wikipedia.org/wiki/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

Strength of Priors

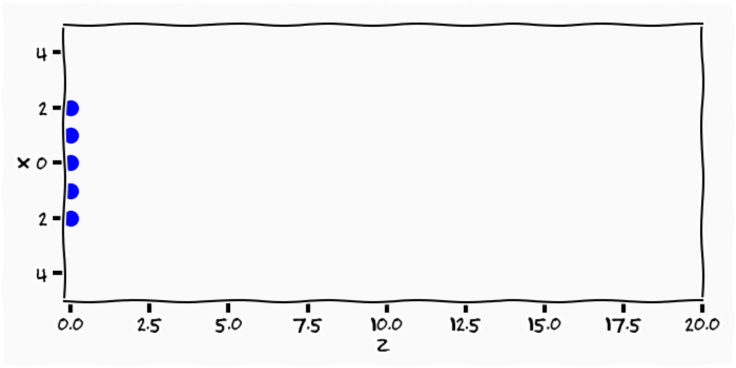
$$y = f(x)$$

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

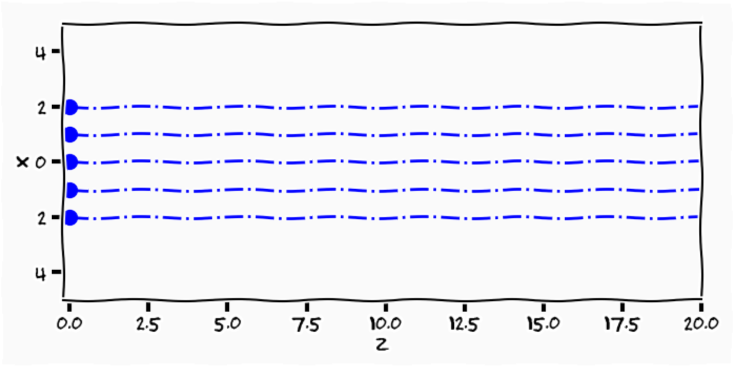
Unsupervised learning

$$y = f(x)$$

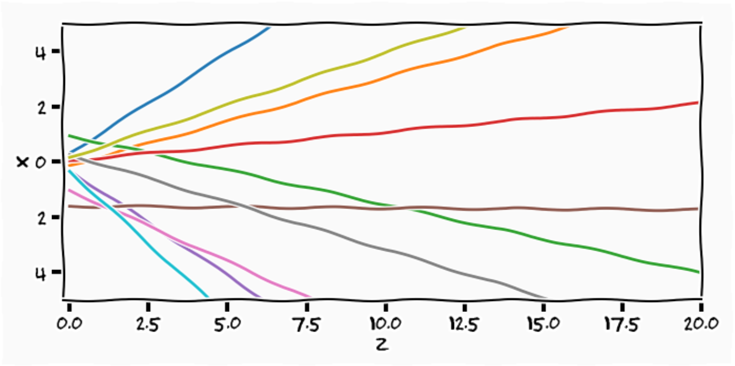
- In unsupervised learning we are given only output
- Input is *latent*
- Task: recover both f and x


4
2
x 0
2
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
z

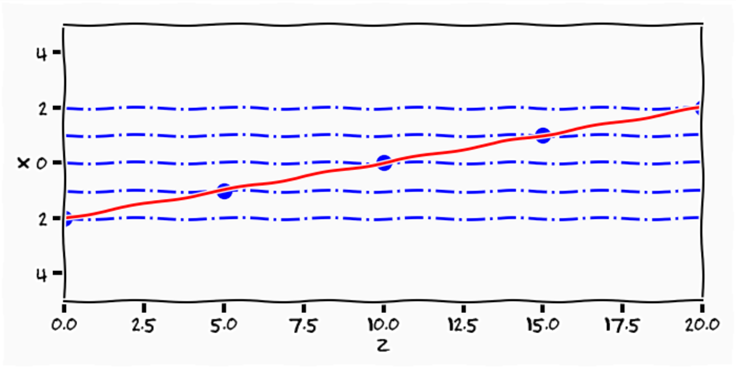
Unsupervised Learning



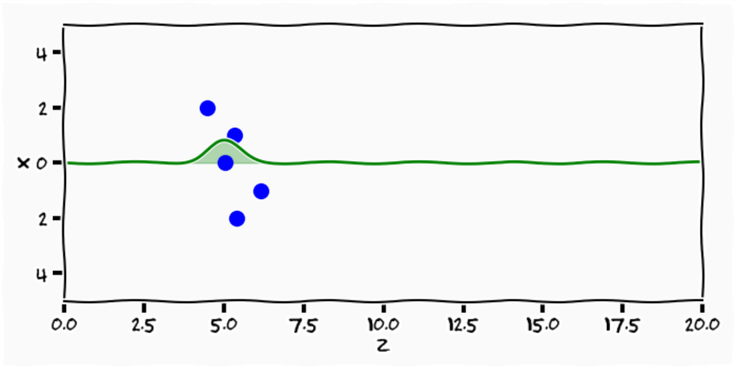
Unsupervised Learning



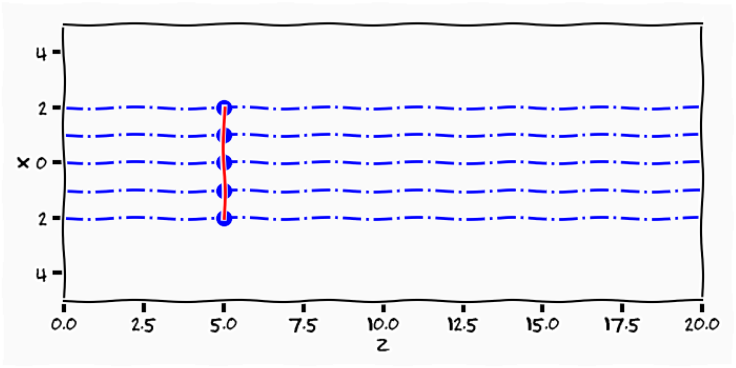
Unsupervised Learning



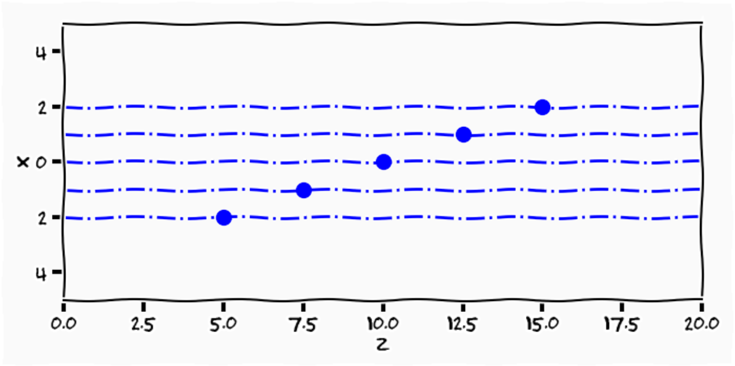
Unsupervised Learning



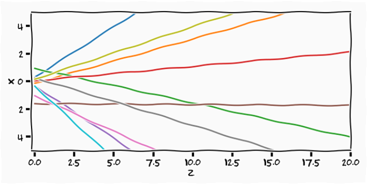
Unsupervised Learning

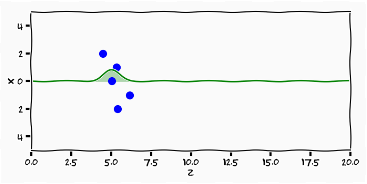


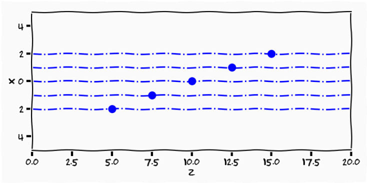
Unsupervised Learning



Unsupervised Learning







Linear Latent Variable Models [1] 12.2

- Linear Regression

$$p(\mathbf{W}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

- Linear Regression

$$p(\mathbf{W}|\mathbf{t},\mathbf{X}) \propto p(\mathbf{t}|\mathbf{W},\mathbf{X})p(\mathbf{W})$$

- Linear Unsupervised Learning

$$p(\mathbf{W},\mathbf{X}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{W},\mathbf{X})p(\mathbf{W})p(\mathbf{X})$$

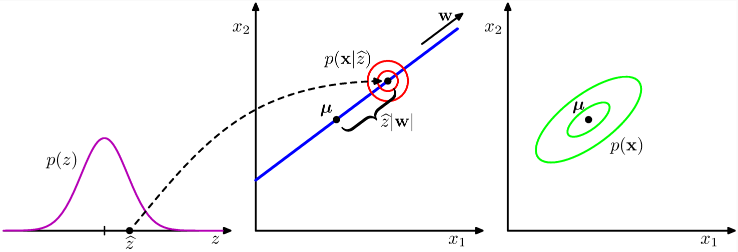
Linear Latent Variable Models [1] 12.2

- Linear Regression

$$p(\mathbf{W}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

- Linear Unsupervised Learning

$$p(\mathbf{W}, \mathbf{X}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})p(\mathbf{X})$$
$$p(\mathbf{W}, \mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{W}, \mathbf{z})p(\mathbf{W})p(\mathbf{z})$$



$$p(\mathbf{x}|\mathbf{W}, \mathbf{z}) = \mathcal{N}(\mathbf{Wz} + \mu, \sigma^2\mathbf{I})$$
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \boldsymbol{I})$$

Linear Latent Variable Model

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

$$p(\mathsf{x}) = \int p(\mathsf{x}|\mathsf{W}, \mathsf{z})p(\mathsf{W})p(\mathsf{z})\mathrm{d}\mathsf{W}\mathrm{d}\mathsf{x}$$

- Intractable to reach posterior distribution of both variables

Linear Latent Variable Model

1. Formulate joint distribution

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

Linear Latent Variable Model

1. Formulate joint distribution

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

2. Formulate marginal distribution over $\mathbf{x}$

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

Linear Latent Variable Model

1. Formulate joint distribution

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

2. Formulate marginal distribution over x

$$p(\mathsf{x}|\mathsf{W}) = \int p(\mathsf{x}|\mathsf{W}, \mathsf{z})p(\mathsf{z})\mathrm{d}\mathsf{x}$$

3. Find maximum-likelihood solution to W

$$\hat{\mathsf{W}} = \mathrm{argmax}_{\mathsf{W}} p(\mathsf{x}|\mathsf{W})$$

Linear Latent Variable Model

1. Formulate joint distribution

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

2. Formulate marginal distribution over x

$$p(\mathsf{x}|\mathsf{W}) = \int p(\mathsf{x}|\mathsf{W}, \mathsf{z})p(\mathsf{z})\mathrm{d}\mathsf{x}$$

3. Find maximum-likelihood solution to W

$$\hat{\mathsf{W}} = \mathrm{argmax}_{\mathsf{W}} p(\mathsf{x}|\mathsf{W})$$

4. Formulate posterior over the latent space

$$p(\mathsf{z}|\mathsf{x}, \hat{\mathsf{W}}) \propto p(\mathsf{x}|\hat{\mathsf{W}}, \mathsf{z})p(\mathsf{z})$$

Linear Latent Variable Model

$$\begin{aligned}
\mathbf{x} &= \mathbf{W}\mathbf{z} + \mu + \epsilon \\
\epsilon &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \\
p(\mathbf{x}|\mathbf{z}, \mathbf{W}) &= \mathcal{N}(\mathbf{W}\mathbf{z} + \mu, \sigma^2 \mathbf{I}) \\
p(\mathbf{z}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
p(\mathbf{x}, \mathbf{z}|\mathbf{W}) &= p(\mathbf{x}|\mathbf{z}, \mathbf{W})p(\mathbf{z})
\end{aligned}$$

Linear Latent Variable Model

$$p(\mathsf{x}, \mathsf{z}|\mathbf{W}) = \mathcal{N}\left(\begin{bmatrix} \mathbb{E}[\mathsf{x}] \\ \mathbb{E}[\mathsf{z}] \end{bmatrix},\right.$$
$$\left.\begin{bmatrix} \mathbb{E}[(\mathsf{x}-\mathbb{E}[\mathsf{x}])(\mathsf{x}-\mathbb{E}[\mathsf{x}])^{\mathrm{T}}] & \mathbb{E}[(\mathsf{x}-\mathbb{E}[\mathsf{x}])(\mathsf{z}-\mathbb{E}[\mathsf{z}])^{\mathrm{T}}] \\ \mathbb{E}[(\mathsf{z}-\mathbb{E}[\mathsf{z}])(\mathsf{x}-\mathbb{E}[\mathsf{x}])^{\mathrm{T}}] & \mathbb{E}[(\mathsf{z}-\mathbb{E}[\mathsf{z}])(\mathsf{z}-\mathbb{E}[\mathsf{z}])^{\mathrm{T}}] \end{bmatrix}\right)$$

- We can compute all the expectations above to figure out what the joint is
- Once we have the joint we can
  - pick out the marginal $p(\mathbf{x}|\mathbf{W})$
  - get the conditional $p(\mathbf{z}|\mathbf{x}, \mathbf{W})$

Linear Latent Variable Model

$$
\begin{aligned}
\mathbb{E}[\mathbf{z}] &= \mathbf{0} \\
\mathbb{E}[\mathbf{x}] &= \mathbb{E}[\mathbf{W}\mathbf{z} + \mu + \epsilon] = \mathbb{E}[\mathbf{W}\mathbf{z}] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon] \\
&= \mathbf{W}\mathbb{E}[\mathbf{z}] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon] = \mathbf{W}\mathbf{0} + \mu + \mathbf{0} \\
&= \mu
\end{aligned}
$$

Linear Latent Variable Model

$$
\begin{aligned}
&\mathbb{E}[(\mathsf{z} - \mathbb{E}[z])(\mathsf{x} - \mathbb{E}[x])^{\mathrm{T}}] = \mathbb{E}[(\mathsf{z} - 0)(\mathsf{x} - \mu)^{\mathrm{T}}] \\
&= \mathbb{E}[\mathsf{z}(\mathsf{Wz} + \mu + \epsilon - \mu)^{\mathrm{T}}] \\
&= \mathbb{E}[\mathsf{z}(\mathsf{Wz} + \epsilon)^{\mathrm{T}}] = \mathbb{E}[\mathsf{z}\left(\mathsf{Wz}\right)^{\mathrm{T}} + \mathsf{z}\epsilon^{\mathrm{T}}] \\
&= \mathbb{E}[\mathsf{z}\mathsf{z}^{\mathrm{T}}\mathsf{W}^{\mathrm{T}}] + \mathbb{E}[\mathsf{z}]\mathbb{E}[\epsilon] = \mathbb{E}[(\mathsf{z} - \mathbf{0})(\mathsf{z} - \mathbf{0})^{\mathrm{T}}]\mathsf{W}^{\mathrm{T}} + 0 \cdot 0 \\
&= \mathsf{I}\mathsf{W}^{\mathrm{T}} = \mathsf{W}^{\mathrm{T}}
\end{aligned}
$$

Linear Latent Variable Model

$$\mathbb{E}[(\mathbf{x} - \mathbb{E}[x])(\mathbf{x} - \mathbb{E}[x])^{\mathrm{T}}] = \mathbb{E}[(\mathbf{Wz} + \mu + \epsilon - \mu)(\mathbf{Wz} + \mu + \epsilon - \mu)^{\mathrm{T}}]$$

$$= \mathbb{E}[(\mathbf{Wz} + \epsilon)(\mathbf{Wz} + \epsilon)^{\mathrm{T}}] = \mathbb{E}[\mathbf{Wz}\,(\mathbf{Wz})^{\mathrm{T}} + \mathbf{Wz}\epsilon^{\mathrm{T}} + \epsilon\mathbf{Wz}^{\mathrm{T}} + \epsilon\epsilon^{\mathrm{T}}]$$

$$= \mathbb{E}[\mathbf{Wzz}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}] + \mathbb{E}[\mathbf{Wz}\epsilon^{\mathrm{T}}] + \mathbb{E}[\epsilon(\mathbf{Wz})^{\mathrm{T}}] + \mathbb{E}[\epsilon\epsilon^{\mathrm{T}}]$$
$$= \mathbf{W}\mathbb{E}[\mathbf{zz}^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}} + \mathbf{W}\mathbb{E}[\mathbf{z}]\mathbb{E}[\epsilon] + \mathbb{E}[\epsilon]\mathbb{E}[\mathbf{z}^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}} + \mathbb{E}[(\epsilon - 0)(\epsilon - 0)^{\mathrm{T}}]$$

$$= \mathbf{WIW}^{\mathrm{T}} + \mathbf{W}0 + 0\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$$
$$= \mathbf{WW}^{\mathrm{T}} + \sigma^2\mathbf{I}$$

Linear Latent Variable Model

Joint

$$p(\mathbf{x}, \mathbf{z}|\mathbf{W}) = \mathcal{N}\left(\begin{bmatrix} \mu \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I} & \mathbf{W} \\ \mathbf{W}^{\mathrm{T}} & \mathbf{I} \end{bmatrix}\right)$$

Marginal

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

Posterior

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}, \mathbf{W}) = \mathcal{N}(&\mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}\right)^{-1}(\mathbf{x} - \mu), \\ &\mathbf{I} - \mathbf{W}^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I})^{-1}\mathbf{W}) \end{aligned}$$

Maximum Likelihood [1] Ch. 12.2.1

$$\log p(\mathbf{x}|\mathbf{W})$$

- find stationary with respect to each variable gives Maximum likelihood solution to $\mathbf{W}$
- you can also do the same with $\boldsymbol{\mu}$ and σ^2

Maximum Likelihood [1] Ch. 12.2.1

$$\log p(\mathbf{x}|\mathbf{W}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W})$$

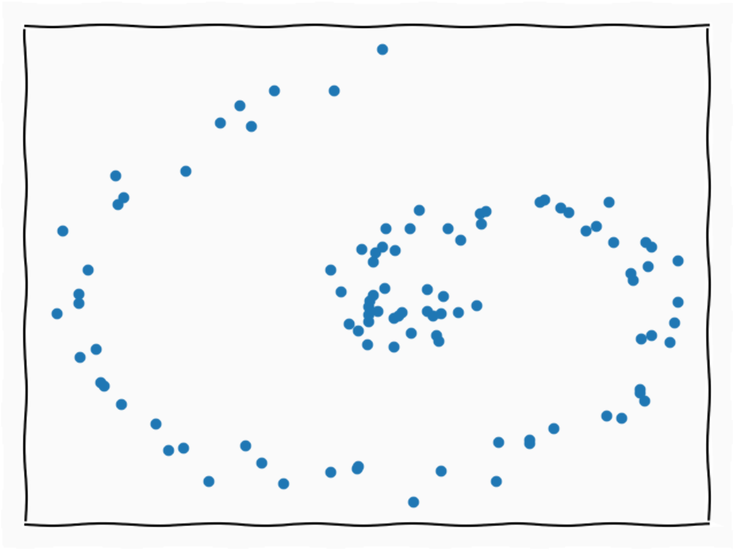
- find stationary with respect to each variable gives Maximum likelihood solution to $\mathbf{W}$
- you can also do the same with $\boldsymbol{\mu}$ and σ^2

Maximum Likelihood [1] Ch. 12.2.1

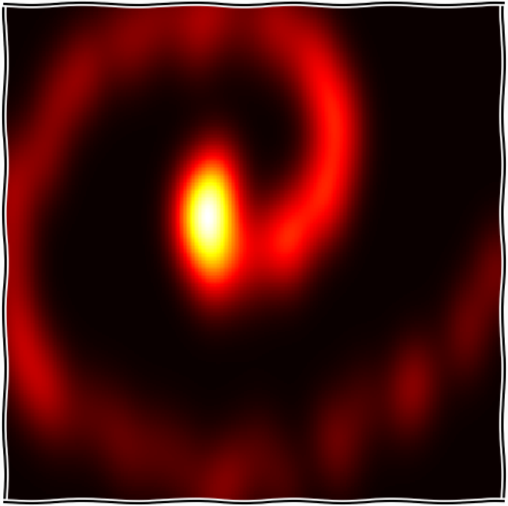
$$
\begin{aligned}
\log p(\mathbf{x}|\mathbf{W}) &= \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W}) \\
&= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}| \\
&\quad - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I})^{-1}(\mathbf{x}_n - \boldsymbol{\mu})
\end{aligned}
$$

- find stationary with respect to each variable gives Maximum likelihood solution to $\mathbf{W}$
- you can also do the same with $\boldsymbol{\mu}$ and σ^2

Example



Example II



Principal Component Analysis [2]

$$\mathbf{V}\Lambda\mathbf{V}^{\mathrm{T}} = \mathbf{x}^{\mathrm{T}}\mathbf{x}$$

$$\mathbf{z} = \sum_{i}^{d} \mathbf{x}\mathbf{V}_i$$

- The above is the solution if $\sigma^2 \rightarrow 0$

[2] [2]

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

Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

2. Formulate your likelihood

$$p(\mathbf{Y}|\theta = \theta_i)$$

Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations
$$p(\mathcal{Y}|\theta)$$
2. Formulate your likelihood
$$p(\mathbf{Y}|\theta = \theta_i)$$
3. Formluate your belief in the "setting" of the model
$$p(\theta)$$

Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

2. Formulate your likelihood

$$p(\mathbf{Y}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

4. Aquire data

Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

2. Formulate your likelihood

$$p(\mathbf{Y}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

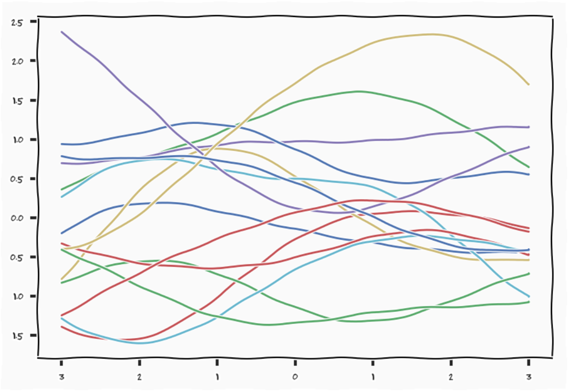
4. Aquire data
5. Derive your updated belief, derive knowledge from data

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{\int p(\mathbf{Y}|\theta)p(\theta)}$$

Unsupervised Learning

- Unsupervised learning is a misnomer, there is no such thing, you have to have beliefs in order to learn.
- Think about unsupervised learning as "less supervised" learning, you have to have stronger beliefs

More Interesting Priors



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

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.

Charles Spearman.
" General Intelligence," Objectively Determined and Measured.
The American Journal of Psychology, 15(2):201–292, 1904.