Special course on Gaussian processes: Session #2

Michael Riis Andersen

Aalto University

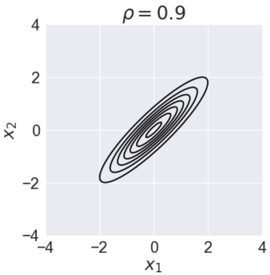
michael.riis@gmail.com

16/1-19

Last session

Last time, we talked about

- The multivariate Gaussian distribution
- The interpretation of the parameters
- Marginalization
- Conditional distributions
- How to sample from the distribution



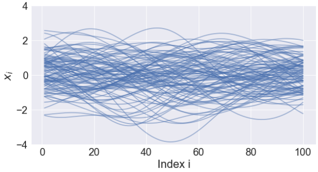
Conditioning one more time

- Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be a partitioning of $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2$, then

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

- The conditional distribution of $\boldsymbol{x}_1$ is given $\boldsymbol{x}_2$ by:

$$p(\boldsymbol{x}_1 | \boldsymbol{x}_2) = \mathcal{N}\left(\boldsymbol{x}_1 | \Sigma_{12}\Sigma_{22}^{-1}\left[\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right] + \boldsymbol{m}_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$$



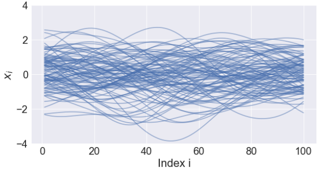
Conditioning one more time

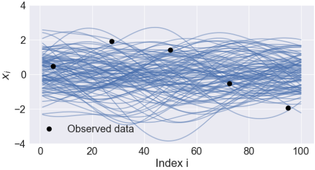
- Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be a partitioning of $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2$, then

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix} \Big| \begin{bmatrix}\boldsymbol{m}_1\\ \boldsymbol{m}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

- The conditional distribution of $\boldsymbol{x}_1$ is given $\boldsymbol{x}_2$ by:

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}\left(\boldsymbol{x}_1|\Sigma_{12}\Sigma_{22}^{-1}\left[\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right] + \boldsymbol{m}_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$





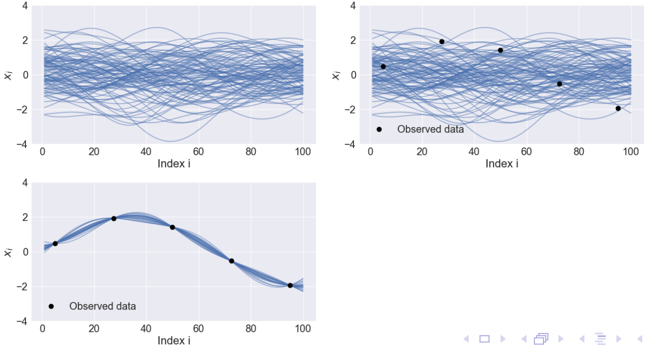
Conditioning one more time

- Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be a partitioning of $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$, then

$$p(\mathbf{x}) = p(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_1 \\ \mathbf{x}_2\end{bmatrix} \,\middle|\, \begin{bmatrix}\mathbf{m}_1 \\ \mathbf{m}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

- The conditional distribution of $\mathbf{x}_1$ is given $\mathbf{x}_2$ by:

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}\left(\mathbf{x}_1|\Sigma_{12}\Sigma_{22}^{-1}\left[\mathbf{x}_2 - \boldsymbol{\mu}_2\right] + \mathbf{m}_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$



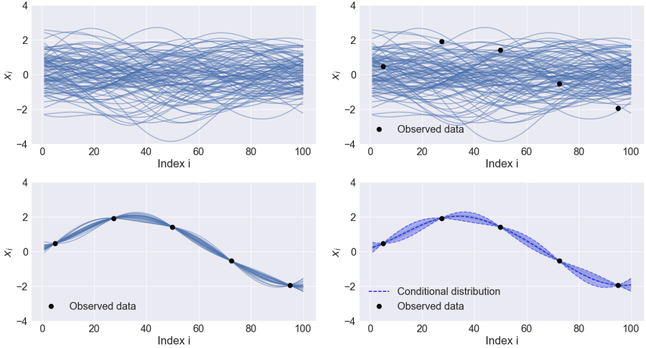
Conditioning one more time

- Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be a partitioning of $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2$, then

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_1\\\boldsymbol{x}_2\end{bmatrix} \Big| \begin{bmatrix}\boldsymbol{m}_1\\\boldsymbol{m}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

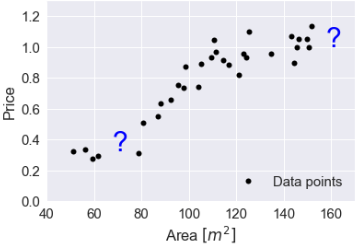
- The conditional distribution of $\boldsymbol{x}_1$ is given $\boldsymbol{x}_2$ by:

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}\left(\boldsymbol{x}_1|\Sigma_{12}\Sigma_{22}^{-1}\left[\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right] + \boldsymbol{m}_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$



Running example

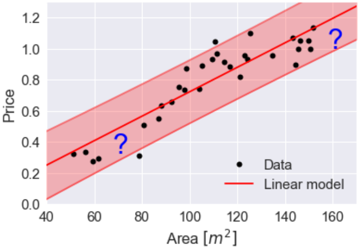
- Suppose we are given a data set of house prices in Helsinki

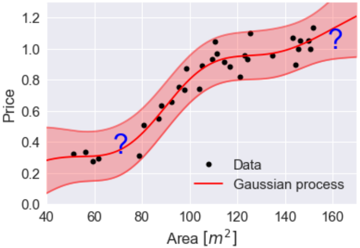


- Goal: Build a model using the data set and predict the average price for a house of $70m^2$ and $160m^2$

Road map for today

1. The Bayesian linear model
2. The linear model as special case of a Gaussian process
3. Gaussian processes: definition & properties
4. Questions & exercise time





General setup for linear regression

- We are given a data set: $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$
- House example: y_n = house price and x_n = house area
- Goal: Learn some function f such that

$$y_n = f(\boldsymbol{x}_n) + \epsilon_n$$

General setup for linear regression

- We are given a data set: $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$
- House example: y_n = house price and x_n = house area
- Goal: Learn some function f such that

$$y_n = f(\boldsymbol{x}_n) + \epsilon_n$$

- Assuming f is a linear model:

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + x_D x_D = \boldsymbol{w}^T \boldsymbol{x}$$

General setup for linear regression

- We are given a data set: $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$
- House example: y_n = house price and x_n = house area
- Goal: Learn some function f such that

$$y_n = f(\boldsymbol{x}_n) + \epsilon_n$$

- Assuming f is a linear model:

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + x_D x_D = \boldsymbol{w}^T \boldsymbol{x}$$

- Linear models are linear wrt. parameters, not the data:

$$f(\boldsymbol{x}) = w_1 \phi_1(x_1) + w_2 \phi_2(x_2) + \ldots + x_D \phi_D(x_D) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}),$$

where $\phi_i(\cdot)$ can be non-linear functions.

Which of the following models are linear models and why?

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2^2 + w_3 \sin(x_3) \qquad \text{(Model 1)}$$

$$f(\boldsymbol{x}) = w_1 x_1 + w_2^2 x_2 + w_3^3 x_3 \qquad \text{(Model 2)}$$

$$f(\boldsymbol{x}) = \left(\boldsymbol{w}^T \boldsymbol{x}\right)^2 \qquad \text{(Model 3)}$$

$$f(\boldsymbol{x}) = w_1 \exp(x_1) + w_2 \sqrt{x_2} + w_3 \qquad \text{(Model 4)}$$

$$f(\boldsymbol{x}) = w_1 x_1 + w_2^2 x_2^2 + w_3^3 x_3^3 \qquad \text{(Model 5)}$$

Slope and intercept

- The models so far have not included an intercept:

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \dots w_D x_D$$

- Most often we want to incorporate an intercept term

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \dots w_D x_D$$

- By assuming $x_0 = 1$, we can write

$$\begin{aligned} f(\boldsymbol{x}) &= w_0 \cdot 1 + w_1 x_1 + w_2 x_2 + \dots w_D x_D \\ &= w_0 \cdot x_0 + w_1 x_1 + w_2 x_2 + \dots w_D x_D \\ &= \boldsymbol{w}^T \boldsymbol{x} \end{aligned}$$

Bayesian linear regression

- The model

$$y_n = f(\boldsymbol{x}_n) + \epsilon = \boldsymbol{w}^T \boldsymbol{x}_n + \epsilon, \qquad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- Likelihood for one data point

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}\left(y_n \middle| f(\boldsymbol{x}_n), \sigma^2\right) = \mathcal{N}\left(y_n \middle| \boldsymbol{w}^T \boldsymbol{x}_n, \sigma^2\right)$$

- Likelihood for all data points

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{w}^T \boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{y} \middle| \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}\right)$$

- Next step: we introduce a prior distribution $p\left(\boldsymbol{w}\right)$ for the weights $\boldsymbol{w}$

Bayesian linear regression

- The prior $p(\boldsymbol{w})$ contains our prior knowledge about $\boldsymbol{w}$ **before** we see any data
- Bayes rule gives us the posterior distribution

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

Bayesian linear regression

- The prior $p(\boldsymbol{w})$ contains our prior knowledge about $\boldsymbol{w}$ **before** we see any data
- Bayes rule gives us the posterior distribution

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- Marginal likelihood

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w}$$

Bayesian linear regression

- The prior $p(\boldsymbol{w})$ contains our prior knowledge about $\boldsymbol{w}$ **before** we see any data

- Bayes rule gives us the posterior distribution

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- Marginal likelihood

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w}$$

- The posterior $p(\boldsymbol{w}|\boldsymbol{y})$ captures everything we know about $\boldsymbol{w}$ **after** seing the data

Bayesian linear regression: the posterior distribution

- We choose a Gaussian prior for $\boldsymbol{w}$

$$p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w}\middle|\mathbf{0}, \boldsymbol{\Sigma}_p\right)$$

- The posterior distribution becomes

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})} \\ &= \frac{\mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{w}\middle|\mathbf{0}, \boldsymbol{\Sigma}_p\right)}{p(\boldsymbol{y})} \\ &= \mathcal{N}\left(\boldsymbol{w}\middle|\mu, \boldsymbol{A}^{-1}\right) \end{aligned}$$

where

$$\mu = \frac{1}{\sigma^2}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad \boldsymbol{A} = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_p^{-1}$$

Bayesian linear regression: the predictive distribution

- We often want to compute the **predictive distribution** for y_* at new data point $\boldsymbol{x}_*$
- We obtain the predictive distribution by averaging over the posterior:

$$
\begin{aligned}
p(y_*|\boldsymbol{y}) &= \int p(y_*|\boldsymbol{x}_*)p(\boldsymbol{w}|\boldsymbol{y})\mathrm{d}\boldsymbol{w} \\
&= \int \mathcal{N}\left(y_*|\boldsymbol{w}^T\boldsymbol{x}_*, \sigma^2\right)\mathcal{N}\left(\boldsymbol{w}|\mu, \boldsymbol{A}^{-1}\right)\mathrm{d}\boldsymbol{w}
\end{aligned}
$$

Bayesian linear regression: the predictive distribution

- We often want to compute the **predictive distribution** for y_* at new data point $\boldsymbol{x}_*$
- We obtain the predictive distribution by averaging over the posterior:

$$
\begin{aligned}
p(y_*|\boldsymbol{y}) &= \int p(y_*|\boldsymbol{x}_*)p(\boldsymbol{w}|\boldsymbol{y})\mathrm{d}\boldsymbol{w} \\
&= \int \mathcal{N}\left(y_*|\boldsymbol{w}^T\boldsymbol{x}_*, \sigma^2\right)\mathcal{N}\left(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{A}^{-1}\right)\mathrm{d}\boldsymbol{w} \\
&= \mathcal{N}\left(y_*|\boldsymbol{\mu}^T\boldsymbol{x}_*, \sigma^2 + \boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{x}_*\right)
\end{aligned}
$$

Bayesian linear regression: the predictive distribution

- We often want to compute the **predictive distribution** for y_* at new data point $\boldsymbol{x}_*$

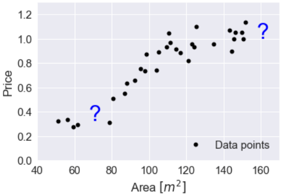
- We obtain the predictive distribution by averaging over the posterior:

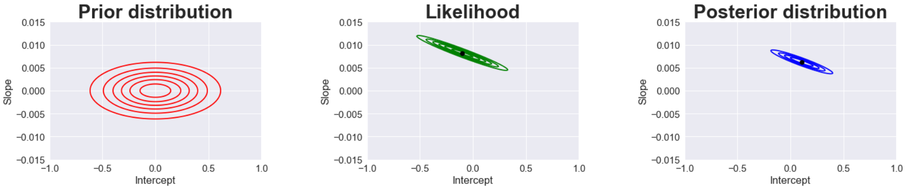
$$
\begin{aligned}
p(y_*|\boldsymbol{y}) &= \int p(y_*|\boldsymbol{x}_*)p(\boldsymbol{w}|\boldsymbol{y})\mathrm{d}\boldsymbol{w} \\
&= \int \mathcal{N}\left(y_*|\boldsymbol{w}^T\boldsymbol{x}_*, \sigma^2\right)\mathcal{N}\left(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{A}^{-1}\right)\mathrm{d}\boldsymbol{w} \\
&= \mathcal{N}\left(y_*|\boldsymbol{\mu}^T\boldsymbol{x}_*, \sigma^2 + \boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{x}_*\right)
\end{aligned}
$$

- The predictive distributions contains two sources of uncertainty:
  1. σ^2: measurement noise
  2. $\boldsymbol{A}^{-1}$: uncertainty of the weights $\boldsymbol{w}$

- $\boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{x}_*$: uncertainty of the weights $\boldsymbol{w}$ projected to the data space

- The posterior distribution is distribution over the parameter space

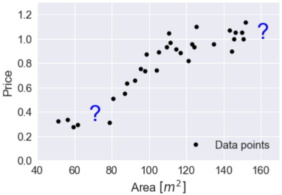


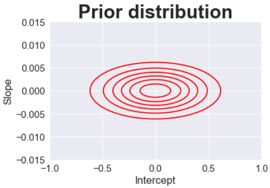


$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Sigma}_p)$ $\quad p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$ $\quad p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mu, \mathbf{A}^{-1})$

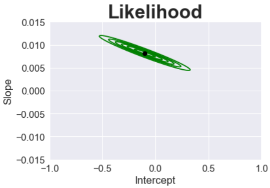
House price example: Posterior and predictive distributions

- The posterior distribution is distribution over the parameter space
- The posterior is compromise between prior and likelihood

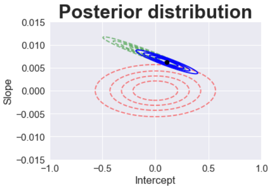




$$p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)$$



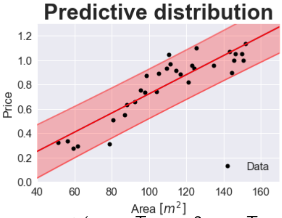
$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}\right)$$



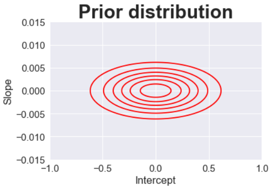
$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{w}\middle|\mu, \boldsymbol{A}^{-1}\right)$$

House price example: Posterior and predictive distributions

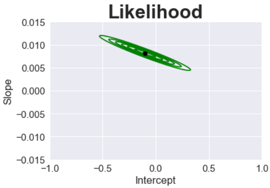
- The posterior distribution is distribution over the parameter space
- The posterior is compromise between prior and likelihood
- The predictive distribution is a distribution over the output space



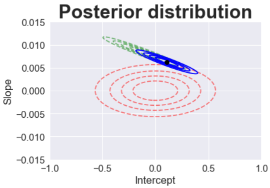
$$p(y^*|\mathbf{y}) = \mathcal{N}\left(y_*|\boldsymbol{\mu}^T\mathbf{x}_*, \sigma^2 + \mathbf{x}_*^T\mathbf{A}^{-1}\mathbf{x}_*\right)$$



$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}\middle|\mathbf{0}, \boldsymbol{\Sigma}_p\right)$$



$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}\left(\mathbf{y}\middle|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)$$



$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{w}\middle|\mu, \mathbf{A}^{-1}\right)$$

Discuss with your neighbor

Determine which of the following statements are true or false:

1. Changing the prior distribution influences the posterior distribution
2. Changing the prior distribution influences the likelihood
3. Changing the prior distribution influences the marginal likelihood
4. Changing the prior distribution influences the predictive distribution
5. The variance of the predictive distribution only depends on the measurement noise

Switching focus from parameters to functions (I)

- Our goal is to learn the function f

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

- Until now we have focused on the weights $\boldsymbol{w}$

$$p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{y} \big| \boldsymbol{w}) p(\boldsymbol{w})$$

- Let's introduce $\boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)] \in \mathbb{R}^N$ to the model

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w}) = p(\boldsymbol{y} \big| \boldsymbol{f}) p(\boldsymbol{f} \big| \boldsymbol{w}) p(\boldsymbol{w})$$

- Our model is still the same

$$p(\boldsymbol{y}, \boldsymbol{w}) = \int p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w}) \mathrm{d}\boldsymbol{f} = p(\boldsymbol{y} \big| \boldsymbol{w}) p(\boldsymbol{w})$$

Switching focus from parameters to functions (II)

- The augmented model

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w}) = p(\boldsymbol{y}\big|\boldsymbol{f})p(\boldsymbol{f}\big|\boldsymbol{w})p(\boldsymbol{w})$$

- What if we now marginalize over the weights

$$p(\boldsymbol{y}, \boldsymbol{f}) = \int p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w})\mathrm{d}\boldsymbol{w} = p(\boldsymbol{y}\big|\boldsymbol{f}) \int p(\boldsymbol{f}\big|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w}$$

- We can also decompose it likelihood and prior

$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}\big|\boldsymbol{f})p(\boldsymbol{f})$$

- where

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}, \boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}\big|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w}$$

Switching focus from parameters to functions (III)

- Let's study the prior distribution on $\boldsymbol{f}$

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}|\boldsymbol{w})\mathcal{N}\left(\boldsymbol{w}|\mathbf{0}, \boldsymbol{\Sigma}_p\right)\mathrm{d}\boldsymbol{w} = ?$$

- We could do the integral directly...

Switching focus from parameters to functions (III)

- Let's study the prior distribution on $\boldsymbol{f}$

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}|\boldsymbol{w})\mathcal{N}\left(\boldsymbol{w}\middle|\mathbf{0}, \boldsymbol{\Sigma}_p\right)\mathrm{d}\boldsymbol{w} = ?$$

- We could do the integral directly...

- But let's instead use the result from last week

$$\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{V}\right) \quad \Rightarrow \quad \boldsymbol{Az} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{Am} + \boldsymbol{b}, \boldsymbol{AVA}^T\right)$$

Switching focus from parameters to functions (III)

- Let's study the prior distribution on $\boldsymbol{f}$

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}|\boldsymbol{w})\mathcal{N}\left(\boldsymbol{w}\middle|\mathbf{0}, \boldsymbol{\Sigma}_p\right)\mathrm{d}\boldsymbol{w} = ?$$

- We could do the integral directly...
- But let's instead use the result from last week

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{V}) \quad \Rightarrow \quad \boldsymbol{Az} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{Am} + \boldsymbol{b}, \boldsymbol{AVA}^T\right)$$

- We know that $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{w}\middle|\mathbf{0}, \boldsymbol{\Sigma}_p\right)$ and $\boldsymbol{f} = \boldsymbol{Xw}$

$$\mathbb{E}[\boldsymbol{f}] = \qquad\qquad\qquad \mathbb{V}[\boldsymbol{f}] =$$

Switching focus from parameters to functions (III)

- Let's study the prior distribution on $\boldsymbol{f}$

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}|\boldsymbol{w})\mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)\mathrm{d}\boldsymbol{w} = ?$$

- We could do the integral directly...

- But let's instead use the result from last week

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{V}) \quad \Rightarrow \quad \boldsymbol{Az} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{Am} + \boldsymbol{b}, \boldsymbol{AVA}^T\right)$$

- We know that $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)$ and $\boldsymbol{f} = \boldsymbol{Xw}$

$$\mathbb{E}[\boldsymbol{f}] = \boldsymbol{X0} + \boldsymbol{0} = \boldsymbol{0} \qquad\qquad \mathbb{V}[\boldsymbol{f}] =$$

Switching focus from parameters to functions (III)

- Let's study the prior distribution on $\boldsymbol{f}$

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}|\boldsymbol{w})\mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)\mathrm{d}\boldsymbol{w} = ?$$

- We could do the integral directly...

- But let's instead use the result from last week

$$\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{V}\right) \quad \Rightarrow \quad \boldsymbol{Az} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{Am} + \boldsymbol{b}, \boldsymbol{AVA}^T\right)$$

- We know that $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)$ and $\boldsymbol{f} = \boldsymbol{Xw}$

$$\mathbb{E}\left[\boldsymbol{f}\right] = \boldsymbol{X0} + \boldsymbol{0} = \boldsymbol{0} \qquad\qquad \mathbb{V}\left[\boldsymbol{f}\right] = \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^T$$

Switching focus from parameters to functions (III)

- Let's study the prior distribution on $\boldsymbol{f}$

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f}|\boldsymbol{w})\mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)\mathrm{d}\boldsymbol{w} =?$$

- We could do the integral directly...

- But let's instead use the result from last week

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{V}) \quad \Rightarrow \quad \boldsymbol{Az} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{Am} + \boldsymbol{b}, \boldsymbol{AVA}^T\right)$$

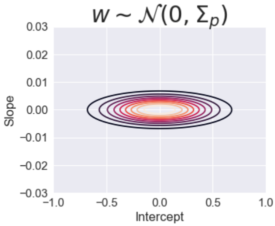
- We know that $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{w}\middle|\boldsymbol{0}, \boldsymbol{\Sigma}_p\right)$ and $\boldsymbol{f} = \boldsymbol{Xw}$

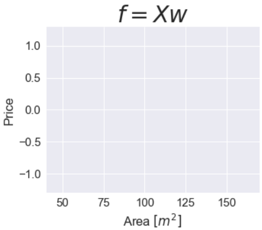
$$\mathbb{E}[\boldsymbol{f}] = \boldsymbol{X0} + \boldsymbol{0} = \boldsymbol{0} \qquad\qquad \mathbb{V}[\boldsymbol{f}] = \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^T$$

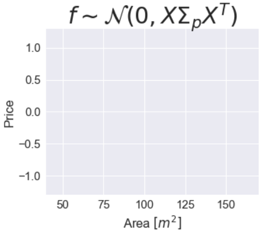
- In other words

$$p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\boldsymbol{0}, \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^T\right)$$

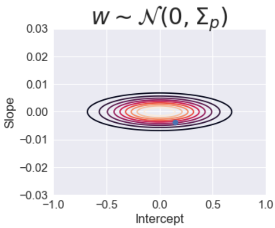
Weight view vs. function view

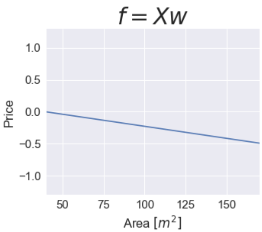


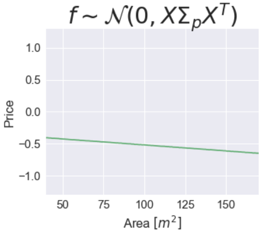




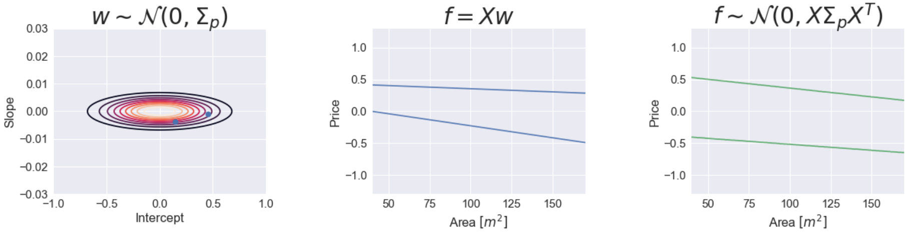
Weight view vs. function view



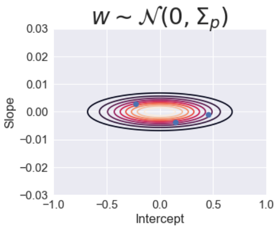


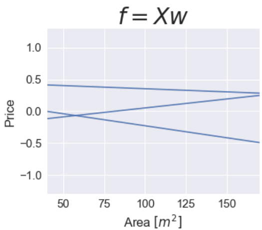


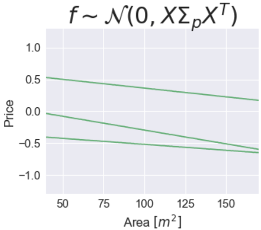
Weight view vs. function view



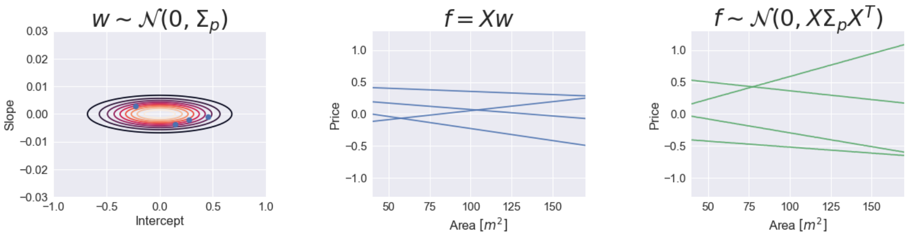
Weight view vs. function view



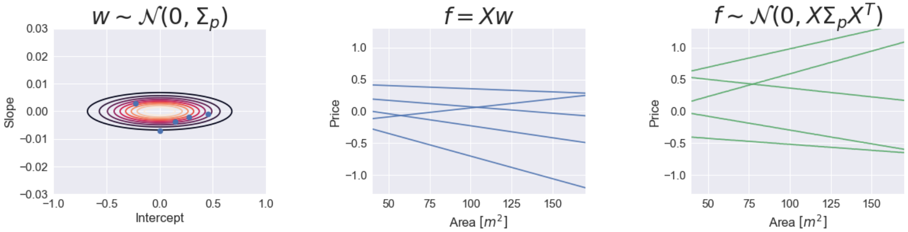




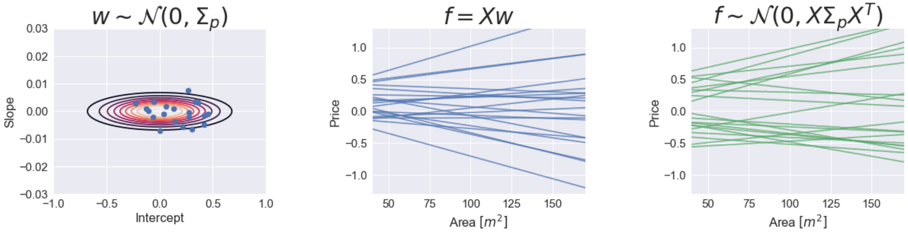
Weight view vs. function view



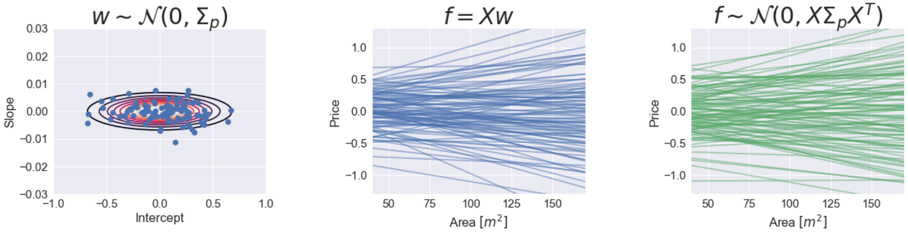
Weight view vs. function view



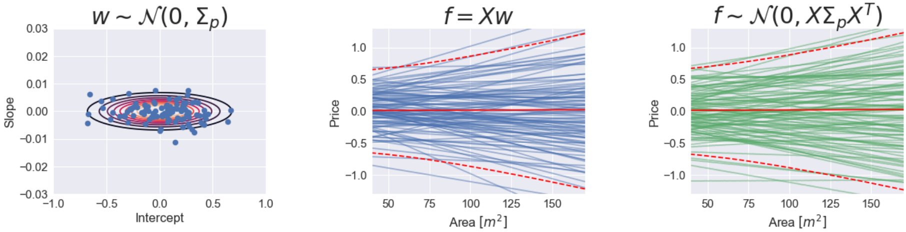
Weight view vs. function view



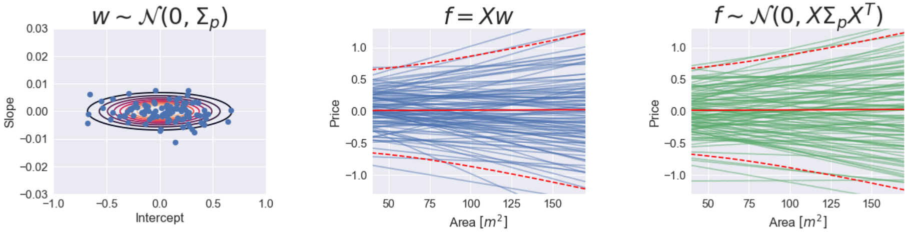
Weight view vs. function view



Weight view vs. function view



Weight view vs. function view



Same distribution for $\boldsymbol{f}$ in both cases but with two different representations

Weight view

- Prior on weights: $p(\boldsymbol{w})$
- $p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})$
- Posterior of weights: $p(\boldsymbol{w}|\boldsymbol{y})$

Function view

- Prior on function values: $p(\boldsymbol{f})$
- $p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})$
- Posterior of function values: $p(\boldsymbol{f}|\boldsymbol{y})$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) = \text{cov}\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right) = \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right)$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\boldsymbol{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$
\begin{aligned}
\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) = \text{cov}\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right) &= \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)}
\end{aligned}
$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$\begin{aligned}\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) &= \text{cov}\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right) = \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\ &= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)} \\ &= \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_j\right]\end{aligned}$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} \middle| \mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$\begin{aligned}
\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) = \text{cov}\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right) &= \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)} \\
&= \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i^T\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{x}_j\right]
\end{aligned}$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$
\begin{aligned}
\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) &= \text{cov}\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right) = \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)} \\
&= \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i^T\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \boldsymbol{x}_i^T\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^T\right]\boldsymbol{x}_j
\end{aligned}
$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} \middle| \mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$
\begin{aligned}
\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) &= \text{cov}\left(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)\right) = \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)} \\
&= \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i^T\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \boldsymbol{x}_i^T\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^T\right]\boldsymbol{x}_j \\
&= \boldsymbol{x}_i^T\boldsymbol{\Sigma}_p\boldsymbol{x}_j
\end{aligned}
$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

$$
\begin{aligned}
\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) = \text{cov}\left(f\left(\boldsymbol{x}_i\right), f\left(\boldsymbol{x}_j\right)\right) &= \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)} \\
&= \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i^T\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \boldsymbol{x}_i^T\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^T\right]\boldsymbol{x}_j \\
&= \boldsymbol{x}_i^T\Sigma_p\boldsymbol{x}_j \\
&= k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)
\end{aligned}
$$

A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}\middle|\boldsymbol{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\Sigma_p\boldsymbol{X}^T$

- Let's have a closer look on the covariance between f_i and f_j

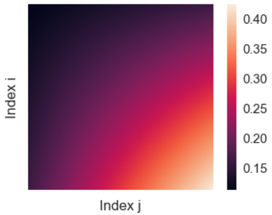
$$\begin{aligned}
\boldsymbol{K}_{ij} = \text{cov}\left(f_i, f_j\right) = \text{cov}\left(f\left(\boldsymbol{x}_i\right), f\left(\boldsymbol{x}_j\right)\right) &= \text{cov}\left(\boldsymbol{w}^T\boldsymbol{x}_i, \boldsymbol{w}^T\boldsymbol{x}_j\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{w}^T\boldsymbol{x}_i - 0\right)\left(\boldsymbol{w}^T\boldsymbol{x}_j - 0\right)\right] \qquad \text{(Why zero mean?)} \\
&= \mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i^T\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{x}_j\right] \\
&= \boldsymbol{x}_i^T\mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^T\right]\boldsymbol{x}_j \\
&= \boldsymbol{x}_i^T\Sigma_p\boldsymbol{x}_j \\
&= k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)
\end{aligned}$$

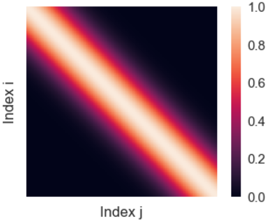
- What happens if we change the form of the **covariance function** $k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$?

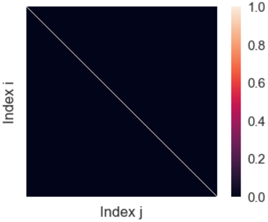
Covariance functions

Linear	Squared exponential	White noise
$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{\Sigma}_p \boldsymbol{x}_j$	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{300}\right)$	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \delta(\boldsymbol{x}_i - \boldsymbol{x}_j)$

$\boldsymbol{K}$



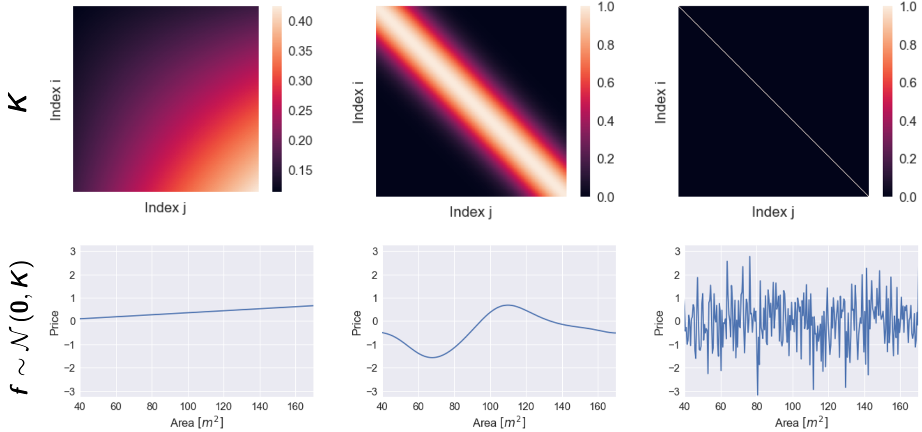




$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$

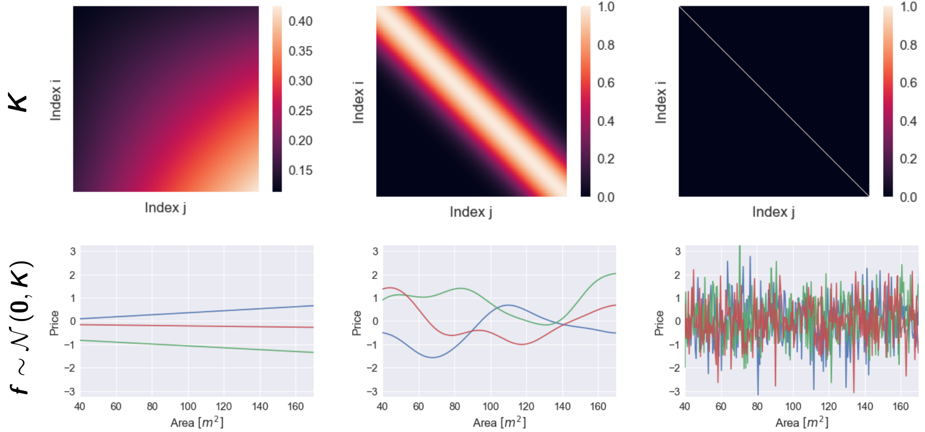
Covariance functions

Linear	Squared exponential	White noise
$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \boldsymbol{\Sigma}_p \mathbf{x}_j$	$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{300}\right)$	$k(\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_i - \mathbf{x}_j)$



Covariance functions

Linear	Squared exponential	White noise
$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i^T \boldsymbol{\Sigma}_p \mathbf{x}_j$	$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\frac{\lVert \mathbf{x}_i - \mathbf{x}_j \rVert}{300}\right)$	$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \delta\left(\mathbf{x}_i - \mathbf{x}_j\right)$



Covariance functions

Linear

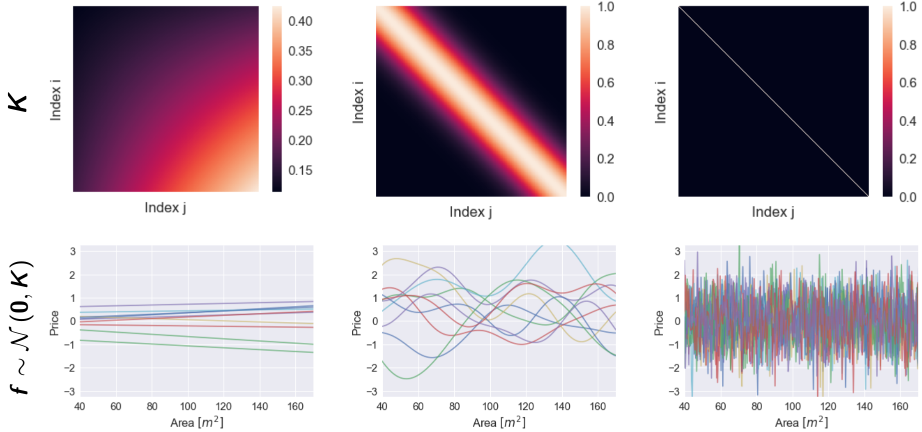
$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{\Sigma}_p \mathbf{x}_j$$

Squared exponential

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{300}\right)$$

White noise

$$k(\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_i - \mathbf{x}_j)$$



Covariance functions

Linear

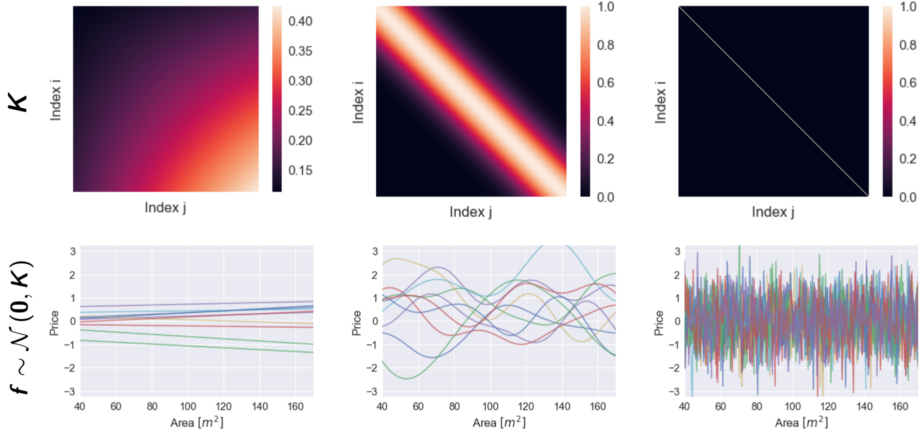
$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{\Sigma}_p \mathbf{x}_j$$

Squared exponential

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j|}{300}\right)$$

White noise

$$k(\mathbf{x}_i, \mathbf{x}_j) = \delta\left(\mathbf{x}_i - \mathbf{x}_j\right)$$



The form of the covariance function determines the characteristics of functions

Discuss with your neighbor

- Consider the following covariance function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = 1 \qquad \text{for all input pairs } (\mathbf{x}_i, \mathbf{x}_j) \tag{1}$$

1. What is the marginal distribution of $f(\mathbf{x}_i)$?
2. What is the covariance between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$?
3. What is the correlation between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$?
4. What kind of functions are represented by the kernel in eq. (1)?

The big picture: Summary so far

1. We started with a Bayesian linear model

$$p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})$$

2. We introduced $\boldsymbol{f}$ into the model and marginalized over the weights $\boldsymbol{w}$

$$p(\boldsymbol{y}, \boldsymbol{f}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})$$

3. This gave us a prior for linear functions in function space $p(\boldsymbol{f})$, where the covariance function for $\boldsymbol{f}$ was given by

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \Sigma_p \boldsymbol{x}$$

4. By changing the form of the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$, we can model much more interesting functions

Definitions

Definition of the multivariate Gaussian distribution

A random vector $\boldsymbol{x} = [x_1, x_2, \cdots, x_D]$ is said to have the **multivariate Gaussian distribution** if all linear combinations of $\boldsymbol{x}$ are Gaussian distributed:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_D x_D \sim \mathcal{N}(m, v)$$

for all $\boldsymbol{a} \in \mathbb{R}^D$

Definition of Gaussian process

A **Gaussian process** is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Characterization and notation

- A Gaussian process can be considered as a prior distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$ (the domain $\mathcal{X}$ is typically $\mathbb{R}^D$)
- A Gaussian process is completely characterized by its mean function $m(\boldsymbol{x})$ and its covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$.

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$$
$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))]$$

- This means that $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ are jointly Gaussian distributed with covariance $k(\boldsymbol{x}, \boldsymbol{x}')$
- Not all functions are valid covariance functions - more on that next session
- We'll use the notation

$$f \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$

Gaussian processes are consistent wrt. marginalization

- Assume the function f follows a Gaussian process distribution:

$$f \sim \mathcal{GP}\left(m\left(\boldsymbol{x}\right), k\left(\boldsymbol{x}, \boldsymbol{x}'\right)\right)$$

- The Gaussian process will induce a density for $\boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)]$:

$$p(\boldsymbol{f}) = p(f_1, f_2) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \middle| \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\right)$$

- The induced density function for $f_1 = f(\boldsymbol{x}_1)$ will always satisfy

$$p(f_1) = \mathcal{N}\left(f_1 \middle| m_1, K_{11}\right)$$

- In words: "Examination of a larger set of variables does not change the distribution of the smaller set"

- If $\mathcal{X} = \mathbb{R}^D$, the GP prior describes infinitely many random variable $\left\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^D\right\}$, but in practice we only have to deal with a finite subset corresponding to the data set at hand

Gaussian process intuition

- Gaussian process implements the assumption:

$$\boldsymbol{x} \approx \boldsymbol{x}' \quad \Rightarrow \quad f(\boldsymbol{x}) \approx f(\boldsymbol{x}')$$

- In other words: If the inputs are similar, the outputs should be similar as well.
- Using the squared exponential covariance function as example

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

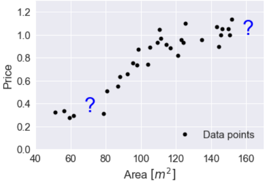
- Then covariance between $f(\boldsymbol{x})$ and $f(\boldsymbol{x})'$ is given by

$$\text{cov}\left[f(\boldsymbol{x}), f(\boldsymbol{x}')\right] = k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\right)$$

- Note: the covariance between outputs are given in terms of the inputs

Back to our house price example (I)

Goal: To predict to the price for a house with area $x_* = 70$ based on the training data $\{x_n, y_n\}_{n=1}^{N}$



- Model: $y_n = f(x_n)$, where f is an unknown function (no noise for now)
- We impose a GP prior on f: $\mathcal{GP}\left(m\left(x\right), k\left(x, x'\right)\right)$
- We choose $m(x) = 0$ and $k\left(x, x'\right)$ to be the covariance function to be the squared exponential (and linear + bias term)
- The joint density for the training data becomes

$$p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} \middle| \boldsymbol{0}, \boldsymbol{K}_{ff}\right)$$

where $\boldsymbol{f} = \left[f\left(x_1\right), f\left(x_2\right), \ldots, f\left(x_N\right)\right]$ and $\left(\boldsymbol{K}_{ff}\right)_{ij} = k\left(x_i, x_j\right)$

Back to our house price example (II)

- The joint density for the training data

$$p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} \middle| \boldsymbol{0}, \boldsymbol{K}_{ff}\right)$$

- But what about the predictions for the new point x_* and the value of $f(x_*)$?

- Let $f_* = f(x_*)$, then we can jointly model $\boldsymbol{f}$ and f_* (consistency property)

$$p(\boldsymbol{f}, f_*) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}\\ f_*\end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

  where $\boldsymbol{K}_{f_*f} = [k(x_*, x_1), k(x_*, x_2), \ldots, k(x_*, x_N)]$ and $K_{f_*f_*} = k(x_*, x_*)$

- Now we can use the rule for conditioning in Gaussian distributions to compute $p(f_*|\boldsymbol{f})$

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^{T}\right)$$

Back to our house price example (III)

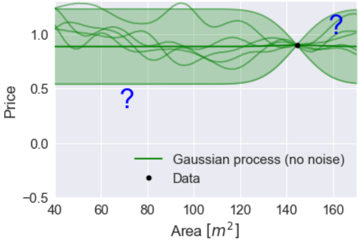
- The joint model for $\boldsymbol{f}$ and f_* is

$$p(\boldsymbol{f}, f_*) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}\\ f_*\end{bmatrix}\middle|\boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

  where $\boldsymbol{K}_{f_*f} = [k(x_*, x_1), k(x_*, x_2), \ldots, k(x_*, x_N)]$ and $K_{f_*f_*} = k(x_*, x_*)$

- Conditioning on $\boldsymbol{f}$ yields:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^T\right)$$



Back to our house price example (III)

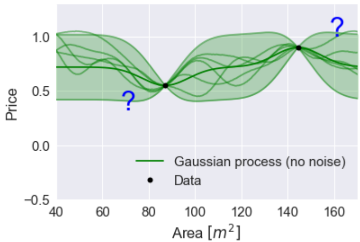
- The joint model for $\boldsymbol{f}$ and f_* is

$$p(\boldsymbol{f}, f_*) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f} \\ f_*\end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*} \\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

where $\boldsymbol{K}_{f_*f} = [k(x_*, x_1), k(x_*, x_2), \ldots, k(x_*, x_N)]$ and $K_{f_*f_*} = k(x_*, x_*)$

- Conditioning on $\boldsymbol{f}$ yields:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^T\right)$$



Back to our house price example (III)

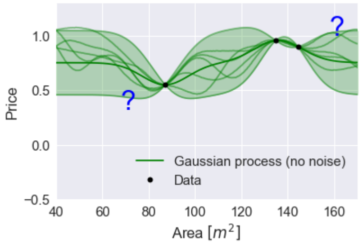
- The joint model for $\boldsymbol{f}$ and f_* is

$$p(\boldsymbol{f}, f_*) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}\\ f_*\end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

where $\boldsymbol{K}_{f_*f} = [k(x_*, x_1), k(x_*, x_2), \ldots, k(x_*, x_N)]$ and $K_{f_*f_*} = k(x_*, x_*)$

- Conditioning on $\boldsymbol{f}$ yields:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^{T}\right)$$



Back to our house price example (III)

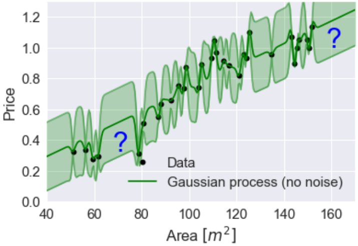
- The joint model for $\boldsymbol{f}$ and f_* is

$$p(\boldsymbol{f}, f_*) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}\\ f_*\end{bmatrix}\middle|\boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

where $\boldsymbol{K}_{f_*f} = [k(x_*, x_1), k(x_*, x_2), \ldots, k(x_*, x_N)]$ and $K_{f_*f_*} = k(x_*, x_*)$

- Conditioning on $\boldsymbol{f}$ yields:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^T\right)$$



Back to our house price example (IV)

- Consider now the noisy model: $y_n = f(x_n) + \epsilon_n$, where ϵ_n is Gaussian distributed

- Same likelihood as for the linear model:

$$p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{f}, \sigma^2\boldsymbol{I}\right)$$

- The joint model for the noisy case becomes

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{f}, f_*) &= p(\boldsymbol{y}\middle|\boldsymbol{f})p(\boldsymbol{f}, f_*) \\ &= \mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{f}, \sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}\\ f_*\end{bmatrix}\boldsymbol{f}\middle|\boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{f_*f}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right) \end{aligned}$$

- Marginalizing over $\boldsymbol{f}$ gives

$$\begin{aligned} p(\boldsymbol{y}, f_*) &= \int p(\boldsymbol{y}\middle|\boldsymbol{f})p(\boldsymbol{f}, f_*)\mathrm{d}\boldsymbol{f} \\ &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y}\\ f_*\end{bmatrix}\boldsymbol{f}\middle|\boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I} & \boldsymbol{K}_{f_*f}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right) \end{aligned}$$

Back to our house price example (V)

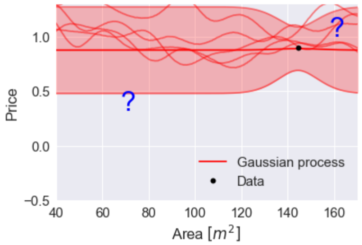
- The joint distribution

$$p(\boldsymbol{y}, f_*) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, f_*)\mathrm{d}\boldsymbol{f}$$

$$= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y} \\ f_*\end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I} & \boldsymbol{K}_{f_*f} \\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{f_*f}^T\right)$$



Back to our house price example (V)

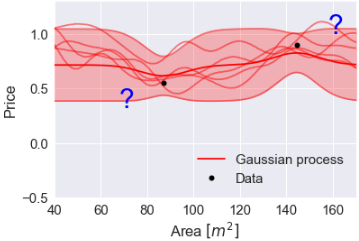
- The joint distribution

$$p(\boldsymbol{y}, f_*) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, f_*)\mathrm{d}\boldsymbol{f}$$

$$= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y}\\ f_*\end{bmatrix} \Big| \mathbf{0}, \begin{bmatrix}\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I} & \boldsymbol{K}_{f_* f}\\ \boldsymbol{K}_{f_* f} & K_{f_* f_*}\end{bmatrix}\right)$$

- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_* \Big| \boldsymbol{K}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, K_{f_* f_*} - \boldsymbol{K}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{f_* f}^T\right)$$



Back to our house price example (V)

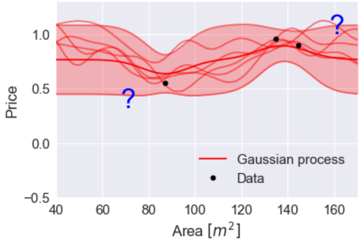
- The joint distribution

$$p(\boldsymbol{y}, f_*) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, f_*)\mathrm{d}\boldsymbol{f}$$

$$= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y}\\ f_*\end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I} & \boldsymbol{K}_{f_*f}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_*f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{f_*f}^T\right)$$



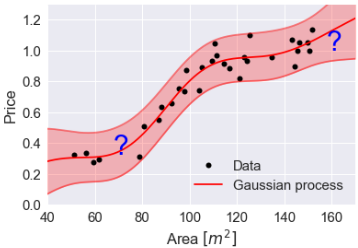
Back to our house price example (V)

- The joint distribution

$$p(\boldsymbol{y}, f_*) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, f_*)\mathrm{d}\boldsymbol{f}$$
$$= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y}\\ f_*\end{bmatrix}\middle|\boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I} & \boldsymbol{K}_{f_*f}\\ \boldsymbol{K}_{f_*f} & K_{f_*f_*}\end{bmatrix}\right)$$

- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_*f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{f_*f}^T\right)$$



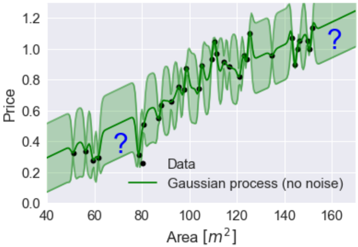
Back to our house price example (V)

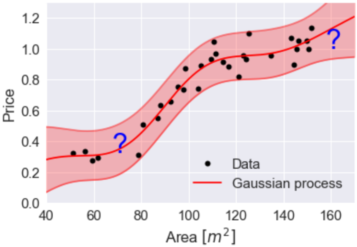
- The joint distribution

$$p(\boldsymbol{y}, f_*) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}, f_*)\mathrm{d}\boldsymbol{f}$$
$$= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y}\\ f_*\end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I} & \boldsymbol{K}_{f_* f}\\ \boldsymbol{K}_{f_* f} & K_{f_* f_*}\end{bmatrix}\right)$$

- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, K_{f_* f_*} - \boldsymbol{K}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{f_* f}^T\right)$$





Questions

Posterior distribution in the noiseless case:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_* f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{y}, K_{f_* f_*} - \boldsymbol{K}_{f_* f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_* f}^T\right)$$

Posterior distribution for the noisy case:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_*\middle|\boldsymbol{K}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, K_{f_* f_*} - \boldsymbol{K}_{f_* f}\left(\boldsymbol{K}_{ff} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{f_* f}^T\right)$$

Is the following statements true or false?:

1. Gaussian processes can fit high non-linear functions, but the predictive means are given by a linear combination of the observed variables $\boldsymbol{y}$.
2. The variance of the posterior distribution is indepedent of the observed variables $\boldsymbol{y}$.

End of todays lecture

Next time:

- Kernels and covariance functions
- Model selection and hyperparameters
- Read ch. 4.2 and ch. 5.1-5.4 in Gaussian processes for Machine Learning by Carl Rasmussen (http://www.gaussianprocess.org/gpml)

Rest of the time today:

- Time to work on assignment #1 (deadline 23rd of January)
- Should be handed in through the my courses system
- In notebook format or in PDF with the same content