CS-E4895 Gaussian Processes
Lecture 2: Bayesian regression

Ti John

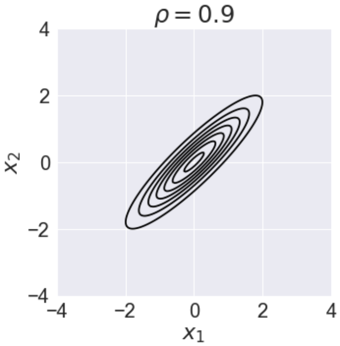
Aalto University

Tuesday 28.2.2023

Last session

Last time, we talked about

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



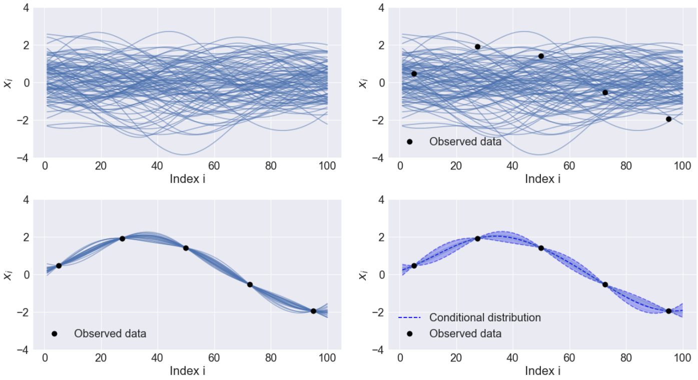
Conditioning one more time

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

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) \tag{1}$$

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

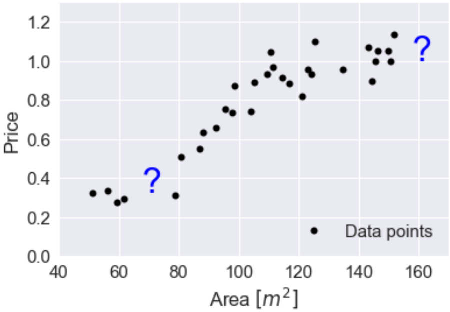
$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}\left(\boldsymbol{x}_1|\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}[\boldsymbol{x}_2 - \boldsymbol{m}_2] + \boldsymbol{m}_1, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right) \tag{2}$$



Gaussian processes for regression

Running example

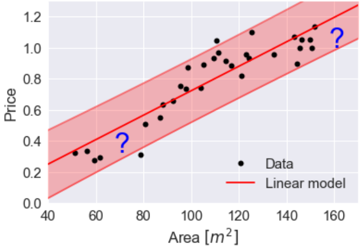
- Suppose we are given a data set of house prices in Helsinki

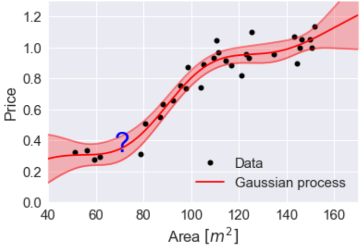


- Goal: Build a model using the data set and predict the average price for a house of 70 m^2 and 160 m^2

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





General setup for linear regression

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

$$y_n = f(\boldsymbol{x}_n) + \epsilon_n \tag{3}$$

- Assuming f is a linear model:

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_D x_D = \sum_i w_i x_i = \boldsymbol{w}^\top \boldsymbol{x} \tag{4}$$

- Linear models are linear w.r.t. parameters, not the data:

$$f(\boldsymbol{x}) = w_1 \phi_1(x_1) + w_2 \phi_2(x_2) + \ldots + w_{D'} \phi_{D'}(x_{D'}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}), \tag{5}$$

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

Quiz

Which of the following models are linear models and why?

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

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

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

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

Quiz

Which of the following models are linear models and why?

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

Yes, w_1, w_2, and w_3 all appear linearly

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

No, because of w_2^2 and w_3^3

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

No, because the weights will not appear linearly

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

Yes, w_1, w_2, and w_3 all appear linearly

Slope and intercept

- The models so far have not included an intercept or bias term
- Most often we want to incorporate an intercept/bias term

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D \tag{6}$$

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

$$\begin{aligned} f(\boldsymbol{x}) &= w_0 \cdot 1 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D \\ &= w_0 \cdot x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D \\ &= \boldsymbol{w}^\top \boldsymbol{x} \end{aligned} \tag{7}$$

Bayesian linear regression: Model and likelihood

- The model

$$y_n = f(\boldsymbol{x}_n) + \epsilon = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon, \qquad \epsilon \sim \mathcal{N}\left(0, \sigma^2_{\mathsf{obs}}\right) \tag{8}$$

- Likelihood for one data point

$$p(y_n | \boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}\left(y_n \middle| f(\boldsymbol{x}_n), \sigma^2_{\mathsf{obs}}\right) = \mathcal{N}\left(y_n \middle| \boldsymbol{w}^\top \boldsymbol{x}_n, \sigma^2_{\mathsf{obs}}\right) \tag{9}$$

- Likelihood for all data points

$$p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n | \boldsymbol{w}^\top \boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{y} \middle| \boldsymbol{X}\boldsymbol{w}, \sigma^2_{\mathsf{obs}} \boldsymbol{I}\right) \tag{10}$$

- Since the data is assumed constant, the likelihood is a function of parameters $\boldsymbol{w}$
- Next step: we introduce a prior distribution $p(\boldsymbol{w})$ for the weights $\boldsymbol{w}$

Bayesian linear regression: prior, posterior, evidence

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

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \tag{11}$$

$$p(\boldsymbol{w} \big| \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w})}{p(\boldsymbol{y})} \tag{12}$$

- Marginal likelihood (or *evidence*)

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = \mathbb{E}_{p(\boldsymbol{w})}\big[p(\boldsymbol{y}|\boldsymbol{w})\big]$$

- The posterior $p(\boldsymbol{w} \big| \boldsymbol{y})$ captures everything we know about $\boldsymbol{w}$ **after** seing the data
- By convention we use $p(\boldsymbol{w} \big| \boldsymbol{y})$ instead of the rigorous form $p(\boldsymbol{w} \big| \boldsymbol{y}, \boldsymbol{X})$

Bayesian linear regression: the posterior distribution

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

$$p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w} \middle| \boldsymbol{0}, \boldsymbol{\Sigma}_p\right) \tag{13}$$

- The parameter posterior distribution becomes

$$\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w})}{p(\boldsymbol{y})} && (14)\\
&= \frac{\mathcal{N}\left(\boldsymbol{y} \middle| \boldsymbol{X}\boldsymbol{w}, \sigma^2_{\text{obs}}\boldsymbol{I}\right) \mathcal{N}\left(\boldsymbol{w} \middle| \boldsymbol{0}, \boldsymbol{\Sigma}_p\right)}{p(\boldsymbol{y})} && (15)\\
&= \mathcal{N}\left(\boldsymbol{w} \middle| \boldsymbol{\mu}, \boldsymbol{A}^{-1}\right) && (16)
\end{aligned}$$

where

$$\boldsymbol{\mu} = \frac{1}{\sigma^2_{\text{obs}}} \boldsymbol{A}^{-1}\boldsymbol{X}^\top \boldsymbol{y} \qquad\qquad \boldsymbol{A} = \frac{1}{\sigma^2_{\text{obs}}} \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{\Sigma}_p^{-1} \tag{17}$$

- See Rasmussen book section 2.1.1 for derivation (book eq 2.7).

Bayesian linear regression: the predictive distribution

- We often want to compute the predictive distribution (or predictive posterior) for the noisy observation y_* at new data point $\boldsymbol{x}_*$, given as $p(y_*|\boldsymbol{y})$
- We obtain the predictive distribution by averaging/marginalizing over the posterior:

$$p(y_*|\boldsymbol{y}) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})\, p(\boldsymbol{w}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{w} \tag{18}$$

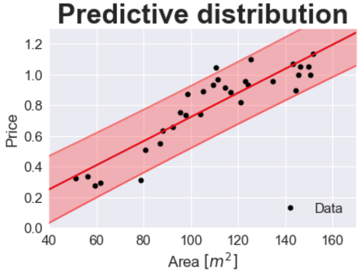
$$= \int \mathcal{N}\left(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \sigma^2_{\text{obs}}\right) \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{A}^{-1}\right) \mathrm{d}\boldsymbol{w} \tag{19}$$

$$= \mathcal{N}\left(y_*|\boldsymbol{\mu}^\top \boldsymbol{x}_*, \sigma^2_{\text{obs}} + \boldsymbol{x}_*^\top \boldsymbol{A}^{-1} \boldsymbol{x}_*\right) \tag{20}$$

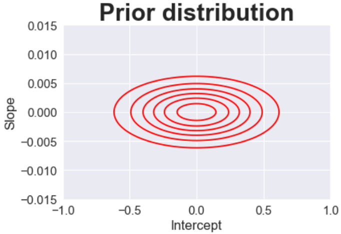
- The predictive distribution contains two sources of uncertainty:
  1. σ^2_{obs}: measurement noise
  2. $\boldsymbol{A}^{-1}$: uncertainty of the weights $\boldsymbol{w}$
- $\boldsymbol{x}_*^\top \boldsymbol{A}^{-1} \boldsymbol{x}_*$: uncertainty of the weights $\boldsymbol{w}$ projected to the data space

House price example: Posterior and predictive distributions

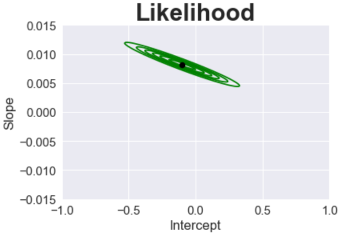
- The posterior distribution is a distribution over the parameter space
- The posterior is compromise between prior and likelihood
- The predictive distribution is a distribution over the output space



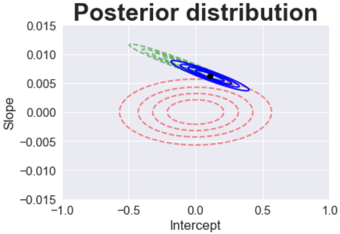
$$p(y^*|\boldsymbol{y}) = \mathcal{N}\left(y_*|\boldsymbol{\mu}^\top \boldsymbol{x}_*, \sigma^2_{\text{obs}} + \boldsymbol{x}_*^\top \boldsymbol{A}^{-1}\boldsymbol{x}_*\right)$$



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



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



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

Quiz

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

Quiz

Determine which of the following statements are true or false:

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

Switching focus from parameters to functions (I)

- Our goal is to learn the function f

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} \tag{21}$$

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

$$p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{y} \big| \boldsymbol{w})\, p(\boldsymbol{w}) \tag{22}$$

- Let's introduce $\boldsymbol{f} = \big[f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)\big] \in \mathbb{R}^N$ to the model
  The vector of predicted function values is $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w}$

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w}) = p(\boldsymbol{y} \big| \boldsymbol{f})\, p(\boldsymbol{f} \big| \boldsymbol{w})\, p(\boldsymbol{w}) \tag{23}$$

- Our model is still the same

$$p(\boldsymbol{y}, \boldsymbol{w}) = \int p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w})\, \mathrm{d}\boldsymbol{f} = p(\boldsymbol{y} \big| \boldsymbol{w})\, p(\boldsymbol{w}) \tag{24}$$

Switching focus from parameters to functions (II)

- The augmented model

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w}) = p(\boldsymbol{y} \big| \boldsymbol{f})\, p(\boldsymbol{f} \big| \boldsymbol{w})\, p(\boldsymbol{w}) \tag{25}$$

- What if we now marginalize over the weights?

$$p(\boldsymbol{y}, \boldsymbol{f}) = \int p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = p(\boldsymbol{y} \big| \boldsymbol{f}) \underbrace{\int p(\boldsymbol{f} \big| \boldsymbol{w})\, p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w}}_{p(\boldsymbol{f})} \tag{26}$$

- We can decompose as likelihood and prior

$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y} \big| \boldsymbol{f})\, p(\boldsymbol{f}) \tag{27}$$

where

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f}, \boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f} \big| \boldsymbol{w})\, p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} \tag{28}$$

Switching focus from parameters to functions (III)

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

$$p(\boldsymbol{f}) = \int p(\boldsymbol{f} | \boldsymbol{w})\, p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = \int p(\boldsymbol{f} | \boldsymbol{w})\, \mathcal{N}\left(\boldsymbol{w} | \mathbf{0}, \boldsymbol{\Sigma}_p\right) \mathrm{d}\boldsymbol{w} = ? \tag{29}$$

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

$$\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{V}\right) \quad \Rightarrow \quad \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{m} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^\top\right) \tag{30}$$

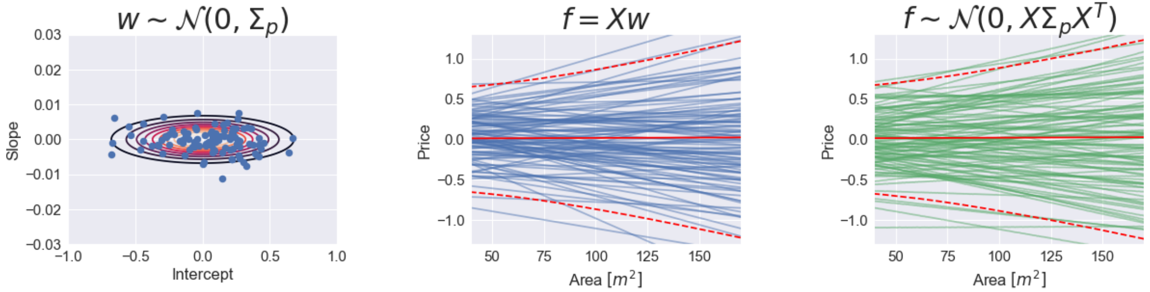
- We know that $\boldsymbol{w} \sim \mathcal{N}\left(\boldsymbol{w} | \mathbf{0}, \boldsymbol{\Sigma}_p\right)$ and $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w}$

$$\mathbb{E}\left[\boldsymbol{f}\right] = \boldsymbol{X}\mathbf{0} + \mathbf{0} = \mathbf{0} \qquad\qquad \mathbb{V}\left[\boldsymbol{f}\right] = \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^\top \tag{31}$$

- In other words

$$p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} | \mathbf{0}, \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^\top\right) \tag{32}$$

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: $p(\boldsymbol{f}|\boldsymbol{y})$

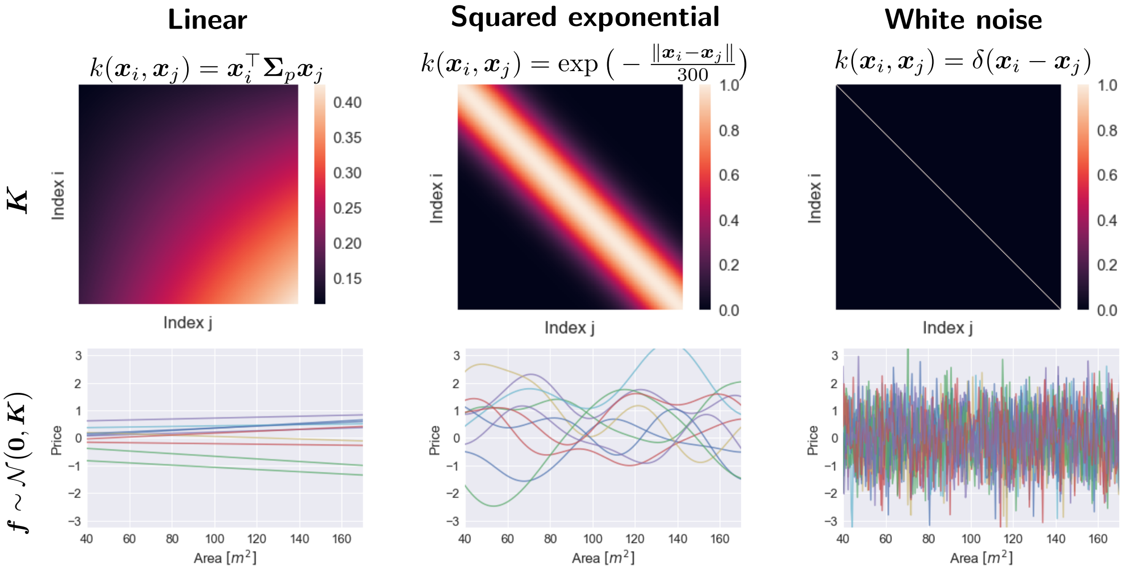
A closer look at the covariance matrix

- Prior on linear functions: $p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} \middle| \mathbf{0}, \boldsymbol{K}\right)$, where $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{\Sigma}_p\boldsymbol{X}^\top$
- Let's have a closer look at the covariance between f_i and f_j

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

- The covariance function is called a kernel function
- What happens if we change the **covariance function** $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$?
- It would change $f(\cdot)$!

Covariance functions



The form of the covariance function determines the characteristics of functions

Quiz

Consider the following covariance function:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = 4 \qquad \text{for all input pairs } (\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{33}$$

1. What is the marginal distribution of $f(\boldsymbol{x}_i)$?
2. What is the covariance between $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}_j)$?
3. What is the correlation between $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}_j)$?
4. What kind of functions are represented by the kernel in eq. (33)?

Quiz

Consider the following covariance function:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = 4 \qquad \text{for all input pairs } (\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{33}$$

1. What is the marginal distribution of $f(\boldsymbol{x}_i)$?
   $p(f_i) = \mathcal{N}(0, 2^2)$ $\qquad$ ($\sigma^2 = 4$ or $\sigma = 2$)
2. What is the covariance between $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}_j)$?
   $\mathrm{cov}(f_i, f_j) = 4$
3. What is the correlation between $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}_j)$?
   $\mathrm{corr}(f_i, f_j) = 1$
4. What kind of functions are represented by the kernel in eq. (33)?
   constant functions

The big picture: Summary so far

1. We started with a Bayesian linear model

$$p(\boldsymbol{y}, \boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w}) \tag{34}$$

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

$$p(\boldsymbol{y}, \boldsymbol{f}) = \int p(\boldsymbol{y}|\boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{w})\, p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} = p(\boldsymbol{y}|\boldsymbol{f})\, p(\boldsymbol{f}) \tag{35}$$

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

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{\Sigma}_p \boldsymbol{x} \tag{36}$$

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

Definitions

Definition: multivariate Gaussian distribution

A random vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_D]$ 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: Gaussian process

A **Gaussian process** is a collection of random variables indexed over space, any finite subset of which have a joint Gaussian distribution.

Characterization and notation

- A Gaussian process can be considered as a prior distribution over functions $f : \mathcal{X} \to \mathbb{R}$ (the domain or index space $\mathcal{X}$ is typically $\mathbb{R}^D$)

$$f(\boldsymbol{x}) \sim \mathcal{GP}\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big) \tag{37}$$

- A Gaussian process is completely characterized by its mean function $m(\boldsymbol{x})$ and its covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$, which define

$$\mathbb{E}\big[f(\boldsymbol{x})\big] = m(\boldsymbol{x}) \tag{38}$$

$$\operatorname{cov}\big[f(\boldsymbol{x}), f(\boldsymbol{x}')\big] = k(\boldsymbol{x}, \boldsymbol{x}') \qquad = \mathbb{E}\Big[\big(f(\boldsymbol{x}) - m(\boldsymbol{x})\big)\big(f(\boldsymbol{x}') - m(\boldsymbol{x}')\big)\Big] \tag{39}$$

  This means that $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ are jointly Gaussian distributed with covariance $k(\boldsymbol{x}, \boldsymbol{x}')$
- The probability of any subset of function values $\boldsymbol{f} = \big[f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)\big]$ at any inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ is

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f} | \boldsymbol{m}, \boldsymbol{K}) \tag{40}$$

  where $\boldsymbol{m} = \big[m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_N)\big]$ and $[\boldsymbol{K}]_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$

Gaussian processes are consistent wrt. marginalization

- Assume the function f follows a Gaussian process distribution:

$$f \sim \mathcal{GP}\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big) \tag{41}$$

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

$$p(\boldsymbol{f}) = p(f_1, f_2) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \,\middle|\, \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \right) \tag{42}$$

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

$$p(f_1) = \mathcal{N}\big(f_1 \big| m_1, K_{11}\big) \tag{43}$$

- 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 variables $\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^D\}$, but in practice we only have to deal with a finite subset corresponding to the data set at hand and where we want to evaluate or 'test' the function

Gaussian process intuition

- A Gaussian process implements the assumption:

$$\boldsymbol{x} \approx \boldsymbol{x}' \quad \Rightarrow \quad f(\boldsymbol{x}) \approx f(\boldsymbol{x}') \tag{44}$$

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

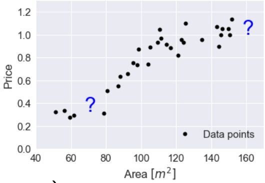
$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\Big) \tag{45}$$

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

$$\text{cov}\big[f(\boldsymbol{x}), f(\boldsymbol{x}')\big] = k(\boldsymbol{x}, \boldsymbol{x}') = \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\Big) \tag{46}$$

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

Back to our house price example (I)



Goal: To predict to the price for a house with area $x_* = 70\,\mathrm{m}^2$ 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: $\mathcal{GP}\big(m(x), k(x, x')\big)$
  - The prior is defined for all $x \in \mathbb{R}$
  - We choose to evaluate the model at 70 observed points and evaluation points
- We choose $m(x) = 0$ and the covariance function $k(x, x')$ to be the squared exponential (and linear + bias term)
- The joint density for the training data becomes

$$p(\boldsymbol{f}) = \mathcal{N}\big(\boldsymbol{f} \big| \boldsymbol{0}, \boldsymbol{K}_{ff}\big) \tag{47}$$

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

Back to our house price example (II)

- The joint density for the training data

$$p(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f} \middle| \mathbf{0}, \boldsymbol{K}_{ff}\right) \tag{48}$$

- 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)

$$p(\boldsymbol{f}, f_*) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{f} \\ f_* \end{bmatrix} \middle| \mathbf{0}, \begin{bmatrix} \boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*} \\ \boldsymbol{K}_{f_*f} & K_{f_*f_*} \end{bmatrix}\right) \tag{49}$$

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

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

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{f}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^\top\right) \tag{50}$$

Back to our house price example (III)

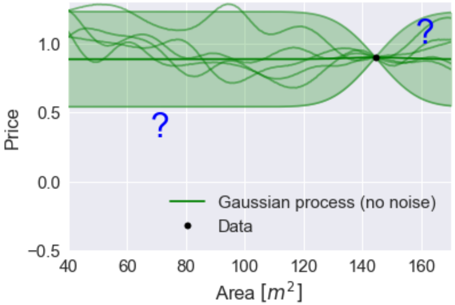
- The joint model for $\boldsymbol{f}$ and f_* is

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

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

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

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



Back to our house price example (III)

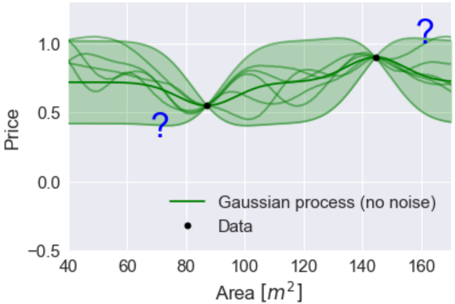
- The joint model for $\boldsymbol{f}$ and f_* is

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

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

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

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



Back to our house price example (III)

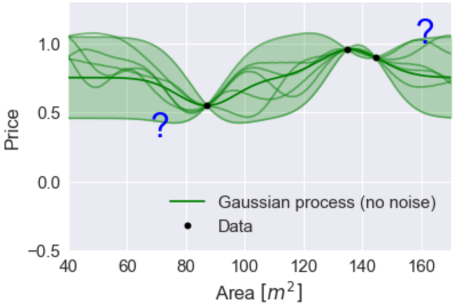
- The joint model for $\boldsymbol{f}$ and f_* is

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

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

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

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



Back to our house price example (III)

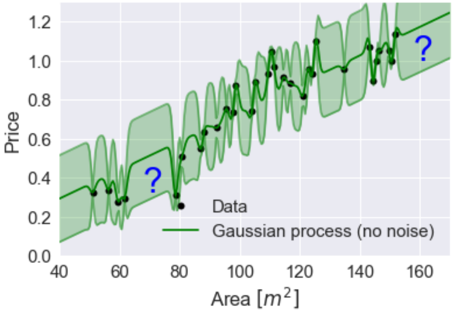
- The joint model for $\boldsymbol{f}$ and f_* is

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

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

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

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



Back to our house price example (IV)

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

- Gaussian likelihood:

$$p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{f}, \sigma^2_{\text{obs}}\boldsymbol{I}\right) \tag{51}$$

- The joint model for the noisy case becomes

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{f}, f_*) &= p(\boldsymbol{y}\big|\boldsymbol{f})\, p(\boldsymbol{f}, f_*) \\ &= \mathcal{N}\left(\boldsymbol{y}\big|\boldsymbol{f}, \sigma^2_{\text{obs}}\boldsymbol{I}\right) \mathcal{N}\left(\begin{bmatrix} \boldsymbol{f} \\ f_* \end{bmatrix} \Big| \boldsymbol{0}, \begin{bmatrix} \boldsymbol{K}_{ff} & \boldsymbol{K}_{ff_*} \\ \boldsymbol{K}_{f_*f} & K_{f_*f_*} \end{bmatrix} \right) \end{aligned} \tag{52}$$

- Marginalizing over $\boldsymbol{f}$ gives

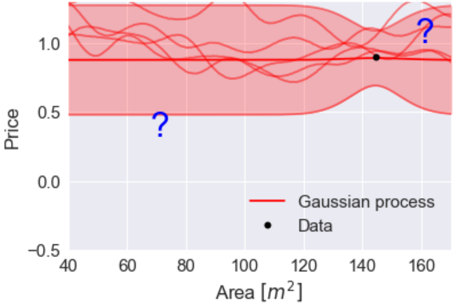
$$\begin{aligned} p(\boldsymbol{y}, f_*) &= \int p(\boldsymbol{y}|\boldsymbol{f})\, p(\boldsymbol{f}, f_*)\, \mathrm{d}\boldsymbol{f} \\ &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{y} \\ f_* \end{bmatrix} \Big| \boldsymbol{0}, \begin{bmatrix} \boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I} & \boldsymbol{K}_{ff_*} \\ \boldsymbol{K}_{f_*f} & K_{f_*f_*} \end{bmatrix} \right) \end{aligned} \tag{53}$$

Back to our house price example (V)

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

- Once again, we can use the rule for conditioning

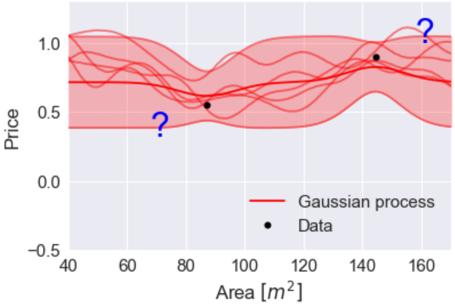
$$p(f_*|\boldsymbol{y}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_* f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_* f_*} - \boldsymbol{K}_{f_* f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{K}^{\top}_{f_* f}\right) \quad (54)$$



Back to our house price example (V)

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

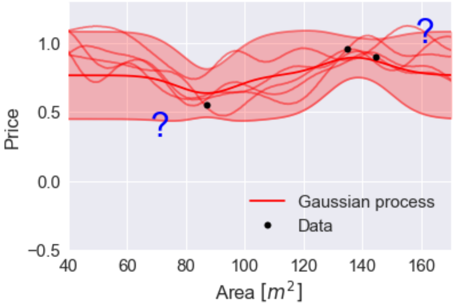
- Once again, we can use the rule for conditioning
$$p(f_*|\boldsymbol{y}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{K}^{\top}_{f_*f}\right) \quad (54)$$



Back to our house price example (V)

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

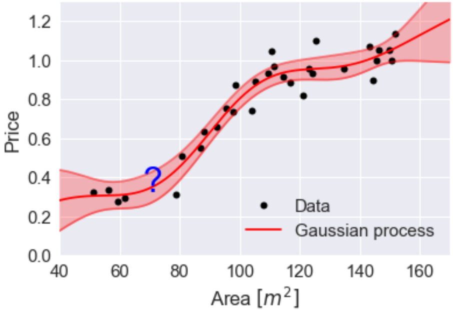
- Once again, we can use the rule for conditioning
$$p(f_*|\boldsymbol{y}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{K}^{\top}_{f_*f}\right) \quad (54)$$



Back to our house price example (V)

- The joint distribution $$p(\boldsymbol{y}, f_*) = \int p(\boldsymbol{y}|\boldsymbol{f})\, p(\boldsymbol{f}, f_*)\, \mathrm{d}\boldsymbol{f}$$ $$= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{y} \\ f_* \end{bmatrix} \middle| \boldsymbol{0}, \begin{bmatrix} \boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I} & \boldsymbol{K}_{ff_*} \\ \boldsymbol{K}_{f_*f} & K_{f_*f_*} \end{bmatrix} \right)$$
- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{y}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{K}^\top_{f_*f}\right) \quad (54)$$

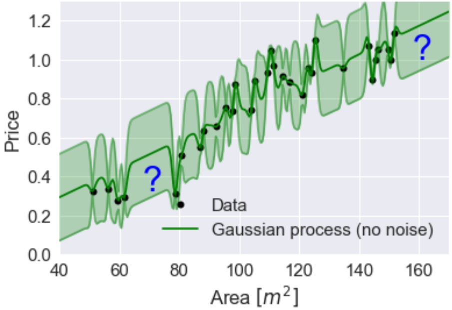


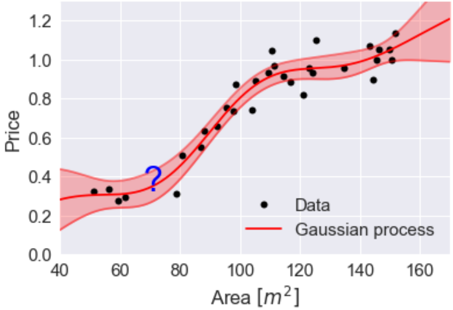
Back to our house price example (V)

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

- Once again, we can use the rule for conditioning

$$p(f_*|\boldsymbol{y}) = \mathcal{N}\left(f_* \middle| \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{K}^\top_{f_*f}\right) \quad (54)$$





Quiz

Posterior distribution in the noiseless case:

$$p(f_*|\boldsymbol{f}) = \mathcal{N}\big(f_*\big|\boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{f}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}\boldsymbol{K}_{ff}^{-1}\boldsymbol{K}_{f_*f}^{\top}\big)$$

Posterior distribution for the noisy case ($y = f + \epsilon$):

$$p(f_*|\boldsymbol{y}) = \mathcal{N}\big(f_*\big|\boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma_{\text{obs}}^2\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_*f_*} - \boldsymbol{K}_{f_*f}(\boldsymbol{K}_{ff} + \sigma_{\text{obs}}^2\boldsymbol{I})^{-1}\boldsymbol{K}_{f_*f}^{\top}\big)$$

Are the following statements true or false?:

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

Quiz

Posterior distribution in the noiseless case:

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

Posterior distribution for the noisy case ($y = f + \epsilon$):

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

Are the following statements true or false?:

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

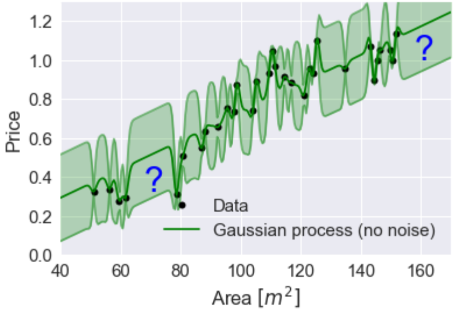
What did we do?

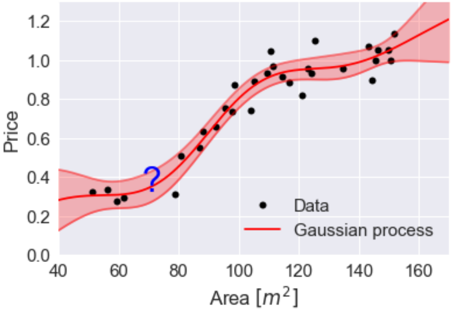
- The predictive function posterior is conveniently a single equation (. . . for regression)

$$p(f_*|\boldsymbol{y}) = \mathcal{N}\big(f_* \big| \boldsymbol{K}_{f_* f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{y}, K_{f_* f_*} - \boldsymbol{K}_{f_* f}(\boldsymbol{K}_{ff} + \sigma^2_{\text{obs}}\boldsymbol{I})^{-1}\boldsymbol{K}^\top_{f_* f}\big)$$

- We ended up not optimizing any parameters, how is this possible?
- Problem: how to define the hyperparameters
  - The noise variance σ^2_{obs}
  - The kernel bandwidth or shape

$\Rightarrow$ Next lecture





End of today's lecture

Room change

- Exercise sessions: Thursdays 10:15 – 12:00, R001/U142 U4

Next lecture:

- Kernels and covariance functions
- Model selection and hyperparameters
- Read ch. 4.2 and ch. 5.1-5.4 in Gaussian process book (`http://gaussianprocess.org/gpml/`)

Assignment:

- Time to work on assignment #1 (released on Wednesday, deadline next Wednesday 8th of March)
- Should be handed in (see Mycourses)
- In Jupyter notebook format