Probabilistic modeling – linear regression & Gaussian processes

Fredrik Lindsten Thomas B. Schön Andreas Svensson Niklas Wahlström

February 23, 2017

Contents

Preface

This is a text on probabilistic modeling for the master level course 'Statistical Machine Learning' given at the Department of Information Technology, Uppsala University during the spring term 2017 and it is a complement to the course books James et al. (2013) and Hastie et al. (2009). It consists of three chapters and one appendix. The three chapters cover an introduction to probabilistic modeling, probabilistic (Bayesian) linear regression, and Gaussian processes. The appendix introduces the multivariate Gaussian distribution and presents key results needed in the chapters. Consequently, the appendix has an important role in this document and should therefore be studied carefully.

Chapter 1

Probabilistic models and learning

In probabilistic modeling we treat all our knowledge in terms of probabilities. We want to answer questions like: *what is chance of a certain temperature $y_\star$ tomorrow at noon, given recorded weather data* $\mathbf{y}$ *from the last week?* We are not primarily interested of a point estimate of $y_\star$, but rather a (mathematical) description of the chance for all possible outcomes of $y_\star$. More formally, the question would be: *what is the probability density $p(y_\star|\mathbf{y})$ for a prediction $y_\star$ conditional on the data* $\mathbf{y}$*?* To answer such questions, we need to reason about *uncertainties* in the data $\mathbf{y}$, predictions $y_\star$ and possible model parameters θ.

In this chapter we will briefly review the fundamentals of random variables. The focus is on the two key rules that underly most of what is done when it comes to probabilistic modeling: marginalization (sometimes also referred to as the sum rule) and conditioning (sometimes also referred to as the product rule). We then move on and introduce the idea of probabilistic modeling in general terms. In Chapter 2 we focus on linear regression and introduce a probabilistic linear regression model. Finally, in Chapter 3 we consider a nonparametric probabilistic regression model using *Gaussian processes*. Appendix A presents the multivariate Gaussian probability distribution and discusses some of its properties which are extensively used throughout this note.

1.1 Random variables

A random variable Z is a variable that can take any value z on a certain set Z and its value depends on the outcome of a random event. For example, if Z describes the outcome of rolling a dice, the possible outcomes are $\mathsf{Z} = \{1, 2, 3, 4, 5, 6\}$ and the probability of the outcome 3 of a die roll is typically modeled to be $1/6$.

In this document, however, we will work with random variables where Z is continuous, for example $\mathsf{Z} = \mathbb{R}$ (Z is a scalar) or $\mathsf{Z} = \mathbb{R}^N$ (Z is an N-vector). Since there are infinitely many possible outcomes z, we cannot speak of the probability of the outcome z (it is almost always zero!). Instead, we use the *probability density function*, denoted by $p(z)$.

Remark 1. In this document we will use the symbol $p(\cdot)$ as a general probability density function, and we will let its argument indicate what the underlying random variable is. For instance, when writing $p(z)$ it is implicit that this is the probability density function for Z, $p(y)$ is the probability density function for Y, etc. Furthermore, we will use the word "distribution" somewhat sloppily, also when referring to a probability density function.

The probability density function $p : \mathsf{Z} \mapsto \mathbb{R}^+$ describes the probability of Z to be within a certain set $\mathsf{C} \subseteq \mathsf{Z}$

$$\Pr[Z \in \mathsf{C}] = \int_{z \in \mathsf{C}} p(z) \mathrm{d}z. \tag{1.1}$$

For example, if Z is a random variable with the probability density function $p(z)$ describing the predicted temperature tomorrow, the chance for this temperature to be between $15°$ and $20°$ is $\Pr[15 < Z < 20] = \int_{15}^{20} p(z)\mathrm{d}z$.

A common probability distribution is the *Gaussian* (or *Normal*) distribution, whose density is defined as

$$p(z) = \mathcal{N}\left(z \,\middle|\, \mu,\, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right), \tag{1.2}$$

where we have made use of $\exp$ to denote the exponential function. In Appendix A we discuss the Gaussian distribution, and also its multivariate extension, in more detail. The multivariate Gaussian distribution is key

to much of the material presented in Chapters 2 and 3, so for readers not familiar with this distribution it is recommended to read Appendix A before moving on with next chapter.

Now, consider two random variables Z_1 and Z_2 (both of which could be vectors). If we are interested in computing the probability that $Z_1 \in \mathsf{C}_1$ *and* $Z_2 \in \mathsf{C}_2$ we need their *joint* probability density function $p(z_1, z_2)$. Using this joint distribution we can compute the probability analogously to the previous case according to

$$\Pr[Z_1 \in \mathsf{C}_1,\ Z_2 \in \mathsf{C}_2] = \int_{z_1 \in \mathsf{C}_1,\ z_2 \in \mathsf{C}_2} p(z_1, z_2) \mathrm{d}z_1 \mathrm{d}z_2. \tag{1.3}$$

From the joint probability density function we can deduce both its two marginal densities $p(z_1)$ and $p(z_2)$ using *marginalization*, as well as the so called conditional probability density function $p(z_2 \mid z_1)$ using *conditioning*. These two concepts will be explained below.

1.1.1 Marginalization

Consider a multivariate random variable Z which is composed of two components Z_1 and Z_2, which could be either scalars or vectors, as $Z = \begin{pmatrix} Z_1^\mathsf{T} & Z_2^\mathsf{T} \end{pmatrix}^\mathsf{T}$. If we know the (joint) probability density function $p(z) = p(z_1, z_2)$, but are interested only in the *marginal* distribution for z_1, we can obtain the density $p(z_1)$ by *marginalization*

$$p(z_1) = \int_{\mathsf{Z}_2} p(z_1, z_2) \mathrm{d}z_2 \tag{1.4}$$

where Z_2 is the space on which Z_2 is defined. The other marginal $p(z_2)$ is obtained analogously by integrating over z_1 instead. In Figure 1.1 a joint two-dimensional density $p(z_1, z_2)$ is illustrated along with their marginal densities $p(z_1)$ and $p(z_2)$.

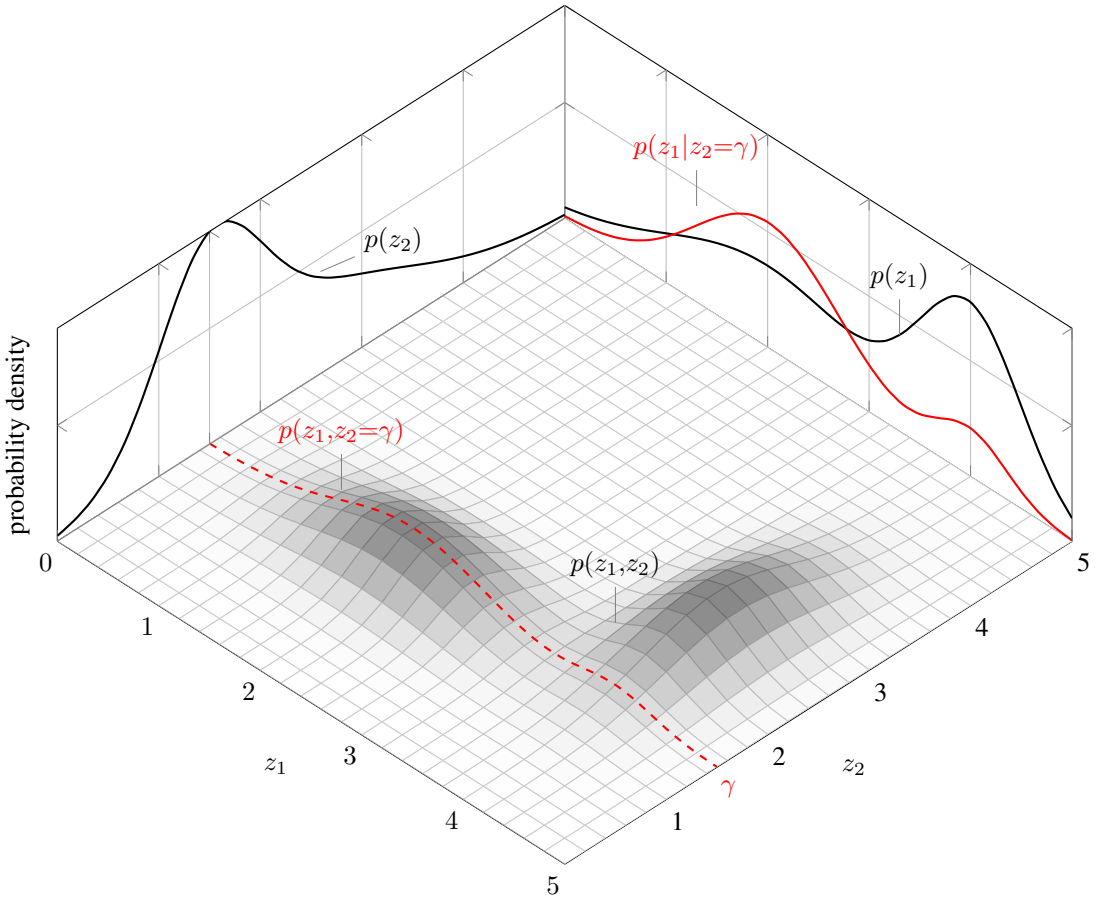


Figure 1.1: Illustration of a two-dimensional *joint* probability distribution $p(z_1, z_2)$ (the surface) and its two *marginal* distributions $p(z_1)$ and $p(z_2)$ (the black lines). We also illustrate the *conditional* distribution $p(z_1|z_2 = \gamma)$ (the red line), which is the distribution of the random variable Z_1 conditioned on the observation $z_2 = \gamma$ ($\gamma = 1.5$ in the plot).

1.1.2 Conditioning

Consider again the multivariate random variable Z which can be partitioned in two parts $Z = \begin{pmatrix} Z_1^\mathsf{T} & Z_2^\mathsf{T} \end{pmatrix}^\mathsf{T}$. We can now define the *conditional* distribution of Z_1, conditioned on having observed a value $Z_2 = z_2$, as

$$p(z_1 \,|\, z_2) = \frac{p(z_1, z_2)}{p(z_2)}. \tag{1.5}$$

If we instead have observed a value of $Z_1 = z_1$ and want to use that to find the conditional distribution of Z_2 given $Z_1 = z_1$, it can be done analogously. In Figure 1.1 a joint two-dimensional probability density function $p(z_1, z_2)$ is illustrated along with a conditional probability density function $p(z_1 \,|\, z_2)$.

1.2 Data $\mathbf{y}$ and the data distribution $p(y \,|\, \theta)$

Most machine learning starts with some data[1] $\mathbf{y} = \{y_1, \ldots, y_N\}$, and then answer questions about what a future not yet seen data point $y_\star$ is likely to be. In many cases, the key for solving the problems is to define a *data distribution* $p(y \,|\, \theta)$, which links y to some parameters θ:

Data distribution

The data distribution $p(y \,|\, \theta)$ is the probability density function for data y conditional on a particular model parameter θ. I.e., the data distribution describes how likely an observation y is, given a parameter value θ.

The data distribution describes the model structure, since it links parameters and data. As an example, the data distribution for the linear regression model is (as we will see in the next chapter)

$$p(y \,|\, \beta) = \mathcal{N}\left(y \,|\, \beta^\mathsf{T} X,\, \sigma^2\right).$$

In general we use θ for unknown parameters, and for the special case of linear regression we use β.

A common assumption which we will make throughout this note is that the data points y_i are conditionally independent. That means that we assume that when we consider the data distribution for all our training data $\mathbf{y}$, we can factorize this as

$$p(\mathbf{y} \,|\, \theta) = p(y_1, y_2, \ldots, y_N \,|\, \theta) = \prod_{i=1}^{N} p(y_i \,|\, \theta). \tag{1.6}$$

1.3 Learning parameters from data

The data distribution $p(y \,|\, \theta)$ essentially describes our model, but it involves some unknown parameters θ. How, then, do we learn these model parameters from observed training data $\mathbf{y}$? One strategy to learn the unknown θ is the maximum likelihood approach:

The maximum likelihood approach

Define the likelihood function $L(\theta) \triangleq p(\mathbf{y} \,|\, \theta)$, and take $\widehat{\theta}_{\text{ML}} = \arg\max_\theta L(\theta)$. In words, this amounts to finding the parameter value $\widehat{\theta}$ which maximizes the chances of having observed the data $\mathbf{y}$.

With the maximum likelihood approach, we obtain a single parameter value $\widehat{\theta}$, i.e. a point estimate, as an answer. This is mainly the approach that we have taken in the course so far, and is also the approach taken by the books Hastie et al. (2009); James et al. (2013). There exists, however, an alternative strategy for learning θ, namely the probabilistic approach. The idea in the probabilistic approach is to consider also the parameters θ as random variables, enabling us to reason probabilistically about our belief regarding θ. We therefore need to introduce another key player: the prior distribution $p(\theta)$.

Prior distribution

The prior $p(\theta)$ is the density function for our unknown parameters θ, *before* we have considered any data.

[1] For regression and classification problems an input x_i is also preset, but we omit that for now.

The prior encodes our *a priori* belief about plausible parameter values. However, as we observe data we naturally want to update this belief according to the evidences provided by the data. Specifically, we want to compute the *posterior* distribution $p(\theta \mid \mathbf{y})$, that is, the distribution of θ conditionally on the observed data. This can be done by making use of Bayes' theorem,

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{p(\mathbf{y})}, \tag{1.7}$$

which allows us to express the posterior distribution in terms of the prior distribution $p(\theta)$ and the data distribution $p(\mathbf{y} \mid \theta)$. The factor $p(\mathbf{y})$, which is the marginal distribution of the data $\mathbf{y}$, does not depend on θ and can thus be seen as a normalization constant in the expression above. With this, we arrive at the probabilistic (or Bayesian) approach:

The probabilistic approach

Compute the posterior distribution $p(\theta \mid \mathbf{y})$, i.e, condition our belief about θ on the data $\mathbf{y}$.

The probabilistic approach provides us not with a number (as did the maximum likelihood approach), but instead a distribution over θ. The extra effort associated with computing this whole distribution instead of a single point estimate pays off in the sense that the posterior distribution $p(\theta \mid \mathbf{y})$ can be seen as a more complete description of our belief about θ. Specifically, it can be used to reason systematically about uncertainties present in our problem.

1.4 Further reading

There are by now quite a few textbooks written on the topic of statistical Machine Learning that makes extensive use of probabilistic models. We recommend Bishop (2006); MacKay (2003); Barber (2012); Murphy (2012), which all take a rather probabilistic view of modeling. As discussed, the course books Hastie et al. (2009); James et al. (2013) focuses more on the maximum likelihood-type of methods. There are also relevant and useful textbooks that are more oriented towards the field of statistics, such as for example for work by Gelman et al. (2013).

The use of probabilistic modeling in designing machines that can learn from experience is explained in the paper by Ghahramani (2015). The learning problems resulting from probabilistic modeling are often lacking a closed-form analytical solution. However, we can still deal with these problems by approximate methods. The work by Efron and Hastie (2016) provides a great overview of how machine learning and statistical fields has evolved due to the introduction of computational approximations in the 1950's.

There is a lot written about the Gaussian distribution, and some concrete starting points are Chapter 2 in Bishop (2006) or Chapter 5 in Gut (1995).

Chapter 2

Probabilistic linear regression

This chapter consider linear regression. Even though being a rather simple model, it is an important building block in more advanced models used in machine learning. For example, the models underlying deep learning can be interpreted as a sequential use of nonlinear transformations of linear regression models. In this chapter we first give the maximum likelihood treatment of linear regression, and then turn to the probabilistic setting.

2.1 Problem formulation

Linear regression models the relationship between p quantitative and/or qualitative inputs $X_1, X_2, \ldots, X_p$ and a quantitative output Y as a linear combination of the input variables, parameterized by some unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$. We also include an additive stochastic noise modeled as a *random variable* ε. The model can thus be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon = \beta^\mathsf{T} X + \varepsilon, \tag{2.1}$$

where

$$X = \begin{pmatrix} 1 & X_1 & X_2 & \ldots & X_p \end{pmatrix}^\mathsf{T} \qquad \text{and} \qquad \beta = \begin{pmatrix} \beta_0, \beta_1, \ldots, \beta_p \end{pmatrix}^\mathsf{T} \tag{2.2}$$

More specifically we will assume that the distribution for ε is known to be a Gaussian random variable with zero mean value and variance σ^2, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Here we use the notation $\sim$ to indicate that ε "is distributed according to" $\mathcal{N}(0, \sigma^2)$. We also make the assumption that the noise ε is independent between different measurements. The input X is seen as known. For the maximum likelihood setting we will model the unknown parameter β as unknown deterministic variable, and later for the probabilistic setting as an unknown random variable. In both cases, the output Y will be a random variable since it is the sum of $\beta^\mathsf{T} X$ and the random variable ε. This gives us a data distribution which inherits the Gaussian distribution from the noise term as

$$p(y \,|\, \beta) = p_\varepsilon(y - \beta^\mathsf{T} X) = \mathcal{N}\left(y \,\middle|\, \beta^\mathsf{T} X,\, \sigma^2\right), \tag{2.3}$$

where $p_\varepsilon(\cdot)$ denotes the probability density function for the random variable ε and $\mathcal{N}\left(y \,\middle|\, \beta^\mathsf{T} X,\, \sigma^2\right)$ is shorthand notation for the Gaussian probability density function of the random variable Y with mean value $\beta^\mathsf{T} X$ and variance σ^2.

When we have access to a training dataset $\mathcal{T} = \{x_i, y_i\}_{i=1}^N$ consisting of N input-output data pairs it can sometimes be convenient to make use of an even more compact notation based on matrices,

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{E}, \tag{2.4}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Np} \end{pmatrix} = \begin{pmatrix} x_1^\mathsf{T} \\ x_2^\mathsf{T} \\ \vdots \\ x_N^\mathsf{T} \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}. \tag{2.5}$$

2.2 Maximum likelihood for a linear Gaussian model

With the data distribution (2.3) in place, we can start the problem of learning θ from $\mathbf{y}$ in the training data $\mathcal{T}$, for which we will take the maximum likelihood approach in this section. Remember that we can factorize the data distribution (1.6)

$$p(\mathbf{y} \mid \beta) = p(y_1, y_2, \ldots, y_N \mid \beta) = \prod_{i=1}^{N} p(y_i \mid \beta) = \prod_{i=1}^{N} \mathcal{N}\left(y_i \mid \beta^{\mathsf{T}} x_i,\, \sigma^2\right), \tag{2.6}$$

where the second equality is due to the fact that the measurements are assumed to be conditionally independent and the third equality amounts to making use of (2.3). We now remind ourselves of the the so-called *likelihood function* $L(\beta)$, the value of the data distribution evaluated at the training data $\mathbf{y}$,

$$L(\beta) = \prod_{i=1}^{N} \mathcal{N}\left(y_i \mid \beta^{\mathsf{T}} x_i,\, \sigma^2\right). \tag{2.7}$$

Note that $L(\beta)$ is a function of the model parameter β with *the training data kept fixed*. The likelihood function is thus a deterministic function of the unknown deterministic variable β obtained by considering $p(\mathbf{y} \mid \beta)$ for a fixed $\mathbf{y}$. The idea in maximum likelihood is to select the value for β that *maximizes the likelihood function*, resulting in

$$\widehat{\beta}_{\text{ML}} = \arg\max_{\beta}\ L(\beta). \tag{2.8}$$

Hence, the maximum likelihood estimate $\widehat{\beta}_{\text{ML}}$ is defined as the parameter value that makes the observed outputs as likely as possible. An equivalent formulation of (2.8) is obtained by instead maximizing the logarithm of the likelihood function

$$\ell(\beta) = \log(L(\beta)) = \log\left(\prod_{i=1}^{N} \mathcal{N}\left(y_i \mid \beta^{\mathsf{T}} x_i,\, \sigma^2\right)\right). \tag{2.9}$$

The reason for this equivalence is that the logarithmic function is monotonically increasing, implying that a value for β that maximizes $\ell(\beta)$ will also maximize $L(\beta)$. The motivation for maximizing the logarithm of the likelihood function $\ell(\beta)$ rather than the likelihood function $L(\beta)$ itself is twofold. First, it simplifies the mathematical analysis as we will see shortly. Secondly, it helps numerically, since the product of—a potentially large number N of—probabilities in (2.6) is challenging to represent in computers, whereas the sum of the logarithms of these probabilities is much easier. The sum come about due to one of the basic rules of logarithms stating that $\log(ab) = \log(a) + \log(b)$. Repeated use of this rule allows us to conclue that

$$\begin{aligned}
\ell(\beta) &= \log\left(\prod_{i=1}^{N} \mathcal{N}\left(y_i \mid \beta^{\mathsf{T}} x_i,\, \sigma^2\right)\right) = \sum_{i=1}^{N} \log\left(\mathcal{N}\left(y_i \mid \beta^{\mathsf{T}} x_i,\, \sigma^2\right)\right) \\
&= \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta^{\mathsf{T}} x_i)^2}{2\sigma^2}\right)\right) = N \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{N} \frac{(y_i - \beta^{\mathsf{T}} x_i)^2}{2\sigma^2}.
\end{aligned} \tag{2.10}$$

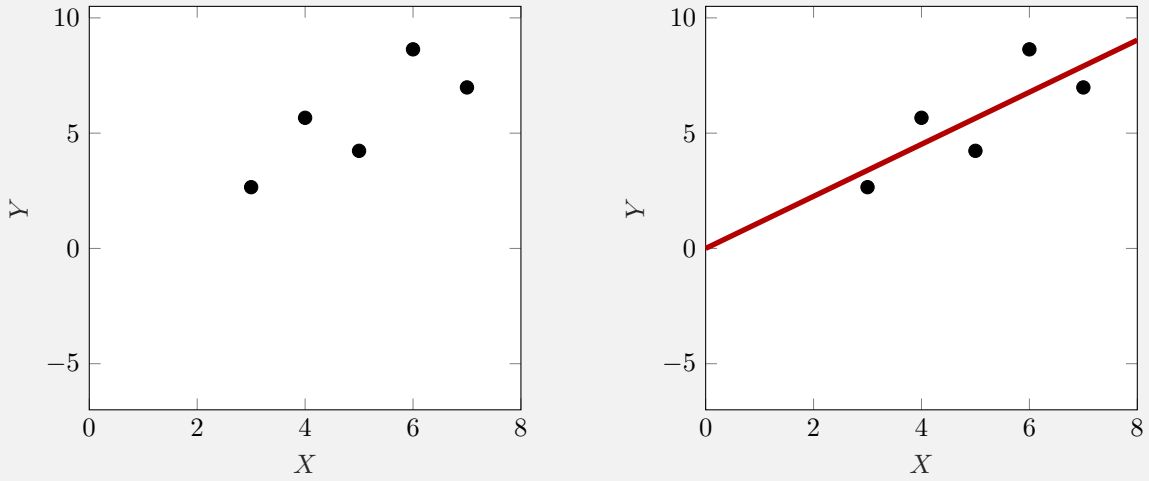
The resulting maximum likelihood problem is given by

$$\widehat{\beta}_{\text{ML}} = \arg\max_{\beta}\ \ell(\beta) = \arg\min_{\beta}\ \sum_{i=1}^{N} (y_i - \beta^{\mathsf{T}} x_i)^2, \tag{2.11}$$

where we have removed the terms that do not contain the unknown parameters β from (2.10) since these terms do not affect the solution of the optimization problem. Note that this is exactly the least squares problem. We have thus showed that when it comes to linear regression with Gaussian noise the maximum likelihood problem is equivalent to least squares and hence the solutions are of course the same, and we refer to the textbook Hastie et al. (2009) for the details on how to compute the least square solution. (However, if other assumption about the noise ε are made, the solution is not equivalent to least square.)

Example 2.1: Maximum likelihood linear regression in a toy example

Consider some training data $\mathcal{T} = \{(3, 2.7), (4, 5.7), (5, 5.7), (6, 4.2), (7, 7.0)\}$, the black dots in the left panel below. We decide to use a maximum likelihood linear regression model with no intercept term, i.e., $Y = \beta_1 X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. Least squares immediately gives $\widehat{\beta}_1 = 1.13$, and $\widehat{\beta}_1 X$ is plotted as a red line in the right panel.



2.3 A probabilistic model for linear regression

We will now turn our attention to a probabilistic model for the linear regression relationship, in the sense that the unknown parameters are modeled as unobserved random variables, as discussed in Chapter 1. We thus require a *prior* distribution $p(\beta)$, representing our prior beliefs about the values of the parameters. The choice of prior distribution is primarily affected by two considerations: *(i)* whether or not we actually have any *a priori* insight into plausible values for the parameters, and *(ii)* in order to obtain computational tractability.

For the probabilistic linear regression model, a simple choice is to assume that the parameters are Gaussian distributed with some mean μ_0 and covariance Σ_0,

$$p(\beta) = \mathcal{N}\left(\beta \,|\, \mu_0,\, \Sigma_0\right). \tag{2.12}$$

If we have little *a priori* insight into plausible values for the parameters we can for instance choose $\mu_0 = 0$ and $\Sigma_0 = \sigma_0^2 I_p$, where σ_0 is some large number.

Once we have observed a set of training data points $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$ we seek to update our belief about β by computing the conditional, or *posterior*, distribution $p(\beta \,|\, \mathbf{y})$, where $\mathbf{y} = (y_1, \ldots, y_N)$ as before.

Remark 2. It is worth pointing out that the posterior distribution $p(\beta \,|\, \mathbf{y})$ depends on the training inputs $(x_1, \ldots, x_N)$ as well as the outputs, even though this dependence is not made explicit in the notation. The reason for why we do not "condition on" the inputs in the notation is that the inputs are viewed as *known deterministic variables*, whereas the outputs are viewed as *observed random variables*.

The posterior distribution is, in accordance with Chapter 1, $p(\beta \,|\, \mathbf{y}) = p(\beta, \mathbf{y})\frac{1}{p(\mathbf{y})}$, and $p(\beta, y) = p(\mathbf{y} \,|\, \beta)p(\beta)$. Since $p(\mathbf{y})$ is independent of β, we can write $p(\beta \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \beta)p(\beta)$, where the proportionality ($\propto$ sign) is with respect to β. The data distribution $p(\mathbf{y} \,|\, \beta)$ is given by (2.6), i.e.,

$$p(\mathbf{y} \,|\, \beta) = \prod_{i=1}^{N} \mathcal{N}\left(y_i \,|\, \beta^{\mathsf{T}} x_i,\, \sigma^2\right). \tag{2.13}$$

Equivalently, we can use the compact matrix notation (2.4) to write the data distribution as

$$p(\mathbf{y} \,|\, \beta) = \mathcal{N}(\mathbf{y} \,|\, \mathbf{X}\beta, \sigma^2 I_N). \tag{2.14}$$

Using the Gaussian prior distribution (2.12) for β, we can now make use of Corollary 1 (with $x_a = \beta, x_b = \mathbf{y}, M = \mathbf{X}, b = 0, \Sigma_{b\,|\,a} = \sigma^2 I_N$) to find the posterior distribution, yielding

$$p(\beta \,|\, \mathbf{y}) = \mathcal{N}(\beta \,|\, \mu_N, \Sigma_N), \tag{2.15a}$$

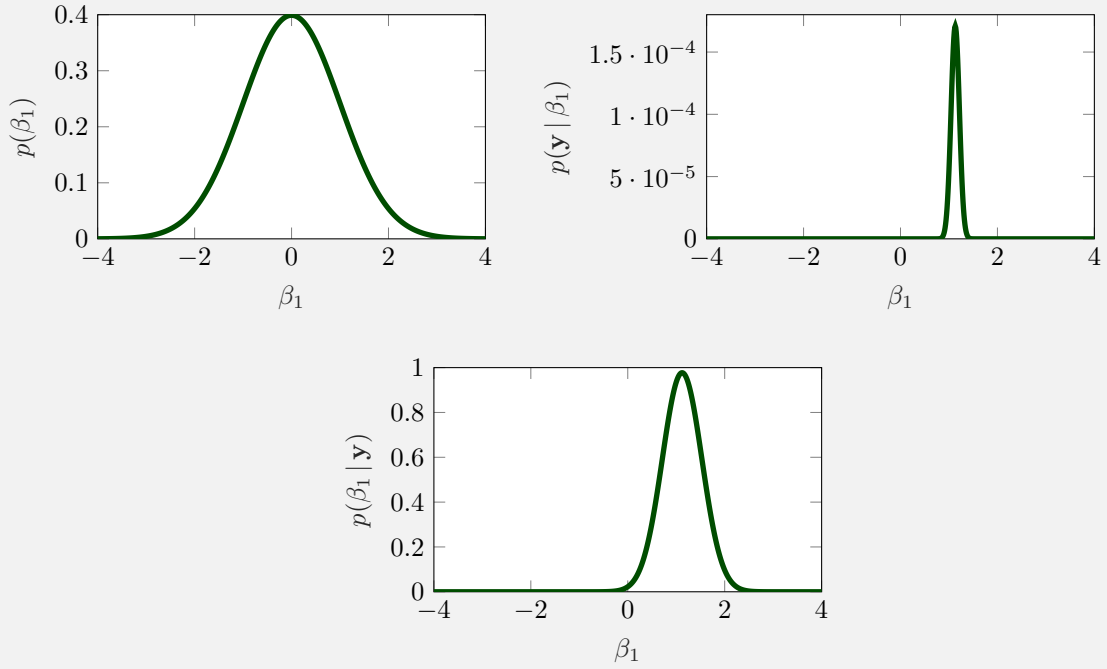
$$\mu_N = \Sigma_N \left(\Sigma_0^{-1}\mu_0 + \sigma^{-2}\mathbf{X}^\mathsf{T}\mathbf{y}\right), \tag{2.15b}$$

$$\Sigma_N = \left(\Sigma_0^{-1} + \sigma^{-2}\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}. \tag{2.15c}$$

Example 2.2: Probabilistic linear regression in a toy example

Consider the same data as in Example 2.1. Now we decide to do probabilistic linear regression with the model $Y = \beta_1 X$. For this, we also need to make a prior assumption about β_1, which we decide to be $p(\beta_1) = \mathcal{N}(\beta_1 \,|\, 0, 1)$, shown to the left below. The posterior, which we seek, is found using Bayes theorem (1.7), or simpler $p(\beta_1 \,|\, \mathbf{y}) \propto p(\beta_1)p(\mathbf{y} \,|\, \beta_1)$. We have therefore plotted $p(\mathbf{y} \,|\, \beta_1)$ to the right below (note that this is *not* a probability density function in β_1!).

Inserting our prior and the data $\mathbf{y}$ into the expressions (2.15) we can see that the posterior distribution $p(\beta_1 \,|\, \mathbf{y})$ is a Gaussian distribution with variance $\Sigma_N = 0.17$ and mean $\mu_N = 1.12$, i.e., $p(\beta_1 \,|\, \mathbf{y}) = \mathcal{N}(\beta_1 \,|\, 1.12, 0.17)$, which we have plotted at the bottom.



2.4 Prediction model

In maximum likelihood linear regression models, where we compute some point estimate $\widehat{\beta}$ of the unknown parameters, the prediction model is simply given by $\widehat{f}(X) = \widehat{\beta}^\mathsf{T} X$. For a probabilistic model, on the other hand, we do not have a single point estimate representing our belief about the unknown parameters. Indeed, training the model amounts to computing the posterior distribution $p(\beta \,|\, \mathbf{y})$ and this whole probability distribution is used to represent our belief about β. When it comes to making predictions, having access to the whole posterior distribution is useful since it allows us to systematically transfer any uncertainty about the parameters into a measure of uncertainty regarding the prediction. This is in contrast with a prediction model based on a single point estimate, which can lead to inaccurate and over-confident predictions since there is always uncertainty about the actual value of β.

More specifically, in a probabilistic modeling setting we can express our prediction model, not using a single point estimate of the parameters, but by taking all possible parameter values into account. Each possible parameter

value is weighted according to its posterior probability. Assume that we want to predict the output $y_\star$ for some test input $X = x_\star$. Using the trained model—i.e., the posterior distribution—we can write the full conditional probability distribution of $y_\star$ as

$$p(y_\star \,|\, \mathbf{y}) = \int p(y_\star \,|\, \beta) p(\beta \,|\, \mathbf{y}) \mathrm{d}\beta, \tag{2.16}$$

which amounts to marginalizing out the unknown parameters β. This conditional distribution captures our complete belief about the value of $y_\star$, given the information that is available through the training data $\mathbf{y}$. It can for instance be used to compute the mean value,

$$\widehat{y}_\star = \mathbb{E}[y_\star \,|\, \mathbf{y}] = \int y_\star p(y_\star \,|\, \mathbf{y}) \mathrm{d}y_\star,$$

which is the predicted value for $y_\star$ that is known to minimize the mean-squared error. Using (2.16) it is also possible to compute the standard deviation, say, of $y_\star$, or the probability that $y_\star$ exceeds some critical value. This type of additional information about the predictions produced by the model can be very valuable in many applications.

For the linear regression model with Gaussian noise ε considered above, the conditional distribution (2.16) can be expressed on closed-form. Using the expression for the posterior distribution (2.15) we have

$$p(y_\star \,|\, \mathbf{y}) = \int \mathcal{N}(y_\star \,|\, \beta^\mathsf{T} x_\star, \sigma^2) \mathcal{N}(\beta \,|\, \mu_N, \Sigma_N) \mathrm{d}\beta. \tag{2.17}$$

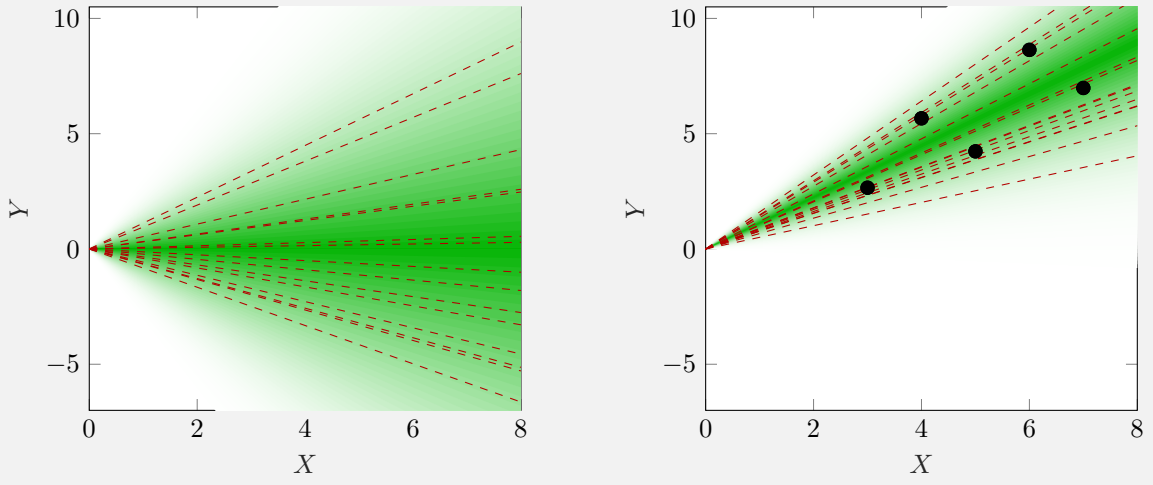
The solution to this integral is provided by Corollary 1 and the result is given by

$$p(y_\star \,|\, \mathbf{y}) = \mathcal{N}(y_\star \,|\, \mu_N^\mathsf{T} x_\star, x_\star^\mathsf{T} \Sigma_N x_\star + \sigma^2). \tag{2.18}$$

Let us briefly reflect upon this result. Our model states that $y_\star = \beta^\mathsf{T} x_\star + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Since ε is independent of both β and $\mathbf{y}$, and the mean of ε is 0, our best guess of the output $y_\star$ (for a given input $x_\star$) is provided by the expected value $\mathbb{E}[y_\star \,|\, \mathbf{y}] = \mathbb{E}[\beta^\mathsf{T} \,|\, \mathbf{y}] x_\star + \mathbb{E}[\varepsilon] = \mu_N^\mathsf{T} x_\star$. The uncertainty in the prediction is encoded by its variance and stems from two sources. The first source is the uncertainty about the parameters β, which is quantified explicitly by the term $x_\star^\mathsf{T} \Sigma_N x_\star$. The second source of uncertainty stems from the uncertainty in the current output itself, given by σ^2. This is the irreducible error or the model and it is thus independent of the training data.

Example 2.3: Predictions for probabilistic linear regression in a toy example

Continuing Example 2.2, we now take a closer look at the predictive distribution $p(y_\star \,|\, \mathbf{y})$. By inserting our previous results into (2.18), we obtain the shaded green regions in the plots below: to the left is the predictive distribution under the prior (i.e., before observing any data), and to the right is the predictive distribution for the posterior (i.e., conditioned on $\mathbf{y}$, black dots). We have also drawn some samples from the distributions, shown as the dotted red lines.



2.5 Relationship to regularized maximum likelihood and ridge regression

A relationship between the posterior probability density function $p(\beta \mid \mathbf{y})$ and the (regularized) maximum likelihood parameter estimate can be found by considering the so called *maximum a posteriori (MAP)* point estimate of β. The MAP estimate is the value of β for which the posterior probability density function reaches its maximum,

$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} p(\beta \mid \mathbf{y}) = \arg\max_{\beta} p(\mathbf{y} \mid \beta)p(\beta) = \arg\max_{\beta} \left(\log p(\mathbf{y} \mid \beta) + \log p(\beta)\right) \tag{2.19}$$

where the second equality follows from the fact that $p(\beta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \beta)p(\beta)}{p(\mathbf{y})}$ and that $p(\mathbf{y})$ does not depend on β. Comparing this to the maximum likelihood point estimate (2.8),

$$\widehat{\beta}_{\text{ML}} = \arg\max_{\beta} p(\mathbf{y} \mid \beta) \tag{2.20}$$

we see that the only difference lies in the addition of the logarithm of the prior probability density function in the former optimization problem. Since the prior distribution does not depend on the data $\mathbf{y}$ the second term in (2.19) can be interpreted as a regularization term—it will "pull" the maximum likelihood estimate towards regions in the parameter space where the prior probability is high.

This connection between MAP estimates and regularized maximum likelihood estimates is general. However, it becomes even more clear when considering the specific case of a Gaussian probabilistic linear regression model as considered above. Assume that the prior for the unknown parameters is given by,

$$p(\beta) = \mathcal{N}(\beta \mid 0, \alpha I_p). \tag{2.21}$$

We can then write the MAP estimate of β as

$$\begin{aligned}
\widehat{\beta}_{\text{MAP}} &= \arg\max_{\beta} \prod_{i=1}^{N} \mathcal{N}(y_i \mid \beta^{\mathsf{T}} x_i, \sigma^2)\mathcal{N}(\beta \mid 0, \alpha I_p) \\
&= \arg\max_{\beta} -\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - x_i^{\mathsf{T}}\beta)^2 - \alpha\sum_{i=1}^{p}\beta_i^2 \\
&= \arg\min_{\beta} \sum_{i=1}^{N}(y_i - x_i^{\mathsf{T}}\beta)^2 + \sigma^2\alpha\|\beta\|_2^2.
\end{aligned} \tag{2.22}$$

where we obtained the second line by taking the logarithm, using the definition of the Gaussian probability density function (A.4) and neglecting all terms that do not depend on the optimization variable β. The third line was obtained via multiplication by $-\sigma^2$ and recalling that $\arg\max_\beta V(\beta) = \arg\min_\beta -V(\beta)$.

The cost function in (2.22) is exactly the *ridge regression* cost function, with regularization parameter $\lambda = \sigma^2\alpha$. That is, computing a MAP estimate for a Gaussian probabilistic linear regression model with prior distribution given by (2.21) is equivalent to computing the ridge regression estimator!

In fact, we could have made this connection to ridge regression directly by noting that the posterior distribution $p(\beta \mid \mathbf{y})$ is Gaussian, with mean and covariance according to (2.15). Since the maximum value of a Gaussian probability density function is attained at its mean, it must hold that the solution to the MAP problem, when using the prior (2.21), is given by

$$\widehat{\beta}_{\text{MAP}} = \mu_N = \left(\alpha I_p + \sigma^{-2}\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\sigma^{-2}\mathbf{X}^{\mathsf{T}}\mathbf{y} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \sigma^2\alpha I_p\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}, \tag{2.23}$$

where we have used (2.15b) with $\mu_0 = 0$ and $\Sigma_0 = \alpha I_p$. Again, we see that this is exactly the solution obtained using ridge regression with $\lambda = \sigma^2\alpha$.

As pointed out above, the regularization term in the MAP problem (2.19) is due to the prior distribution. Therefore, different types of regularization can be obtained by considering different choices for the prior distribution. The explicit derivations above show that in ridge regression can be interpreted from the probabilistic perspective as placing a Gaussian prior over β. If we instead want to recover LASSO, we note that the prior probability density function needs to be such that

$$\log p(\beta) = \alpha\|\beta\|_1 + \text{const.}$$

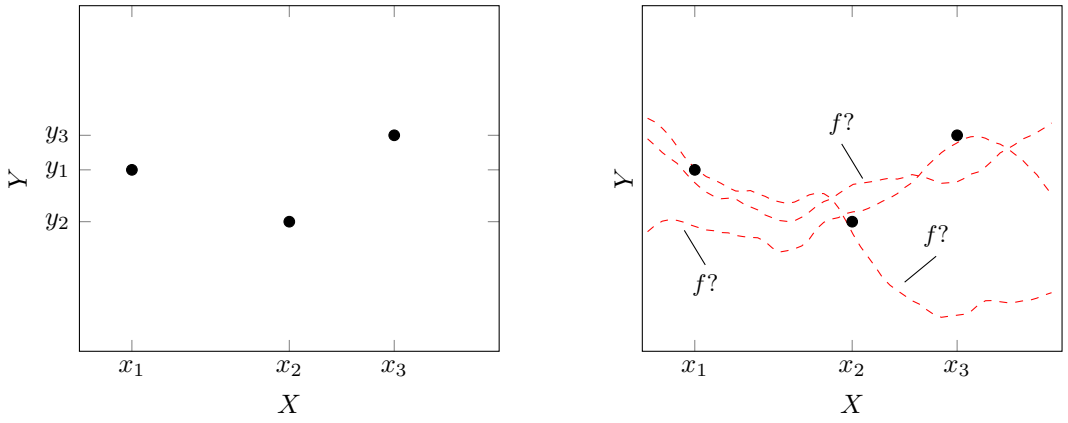
for some value of α. The probability distribution which has this property is referred to as the *Laplace distribution*, meaning that the LASSO estimator can be interpreted as a MAP estimator using a Laplace prior.

Chapter 3

Gaussian Processes

The Gaussian process (GP) is a nonparametric and probabilistic model also for nonlinear relationships. Here we will use it for the purpose of regression. The *nonparametric* nature means that the GP does not rely on any parametric model assumption—instead the GP is flexible with the capability to adapt the model complexity as more data arrives. This means that the training data is *not* summarized by a few parameters (as for linear regression) but is part of the model (as for k-NN). The *probabilistic* nature of the GP provides a structured way of representing and reasoning about the uncertainty that is present both in the model itself and the measured data.

3.1 Constructing the Gaussian process



(a) The data $\{x_i, y_i\}_{i=1}^3$, which we want to have a model for.

(b) We assume there exists *some* function f, which describes the data as $y_i = f(x_i) + \varepsilon_i$.

Figure 3.1: Some data are shown in the left panel, which would not be well explained by a linear model. Instead, we assume there exists some function f (right panel), about which we are going to reason by making use of the Gaussian process.

Assume that we want to fit a model to some training data $\mathcal{T} = \{x_i, y_i\}_{i=1}^3$, as we show in Figure 3.1a. We could make use of linear regression, but even from just these three data points it looks like a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ might be inadequate. Using nonlinear transformations of the input X (polynomials, say) is a possibility, but it can be hard to know what transformations to consider in practice. Instead, we try a different approach in specifying a model. Instead of assuming that we have a linear function, let us just say there exists *some* (possibly non-linear) function f, which describes the data points as $y_i = f(x_i) + \varepsilon_i$, as illustrated by Figure 3.1b.

For two different input values x and x', the unknown function f takes some output values $f(x)$ and $f(x')$, respectively. Let us now *reason probabilistically about this unknown f, by assuming that $f(x)$ and $f(x')$ are*

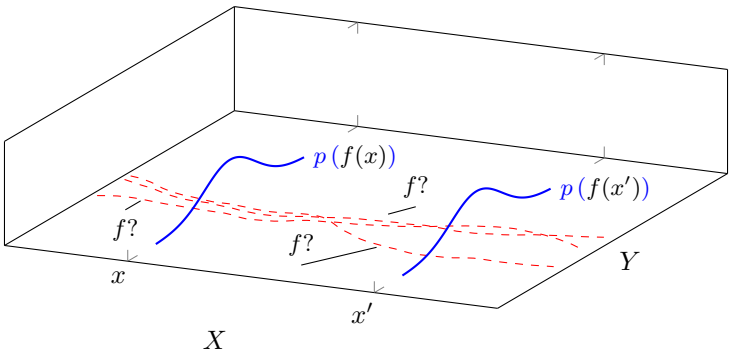


Figure 3.2: The function f is unknown to us, we have given it a pictorial representation by three dashed red lines. The Gaussian process assumption is to model f as random itself, and *assume* that the value of f for any two arbitrary inputs x and x' ($f(x)$ and $f(x')$ respectively) has a joint Gaussian distribution, here represented with the solid blue lines. The distribution over $f(x)$ and $f(x')$ is, however, a *joint* distribution (cf. Figure A.2), even though we have only plotted its two marginal distributions.

jointly Gaussian distributed:

$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \sim \mathcal{N}(\mu, \mathbf{K}), \tag{3.1}$$

We illustrate this by Figure 3.2. Of course, there is nothing limiting us to making this assumption about only two input values x and x', but we may extend it to any *arbitrary* set of input values $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$. This assumption implies that f is what we refer to as a Gaussian process:

Definition 1 (Gaussian process (GP))**.** *A Gaussian process is a (potentially infinite) collection of random variables such that any finite subset of it has a joint multivariate Gaussian distribution.*

In other words, f is unknown to us, and by considering an arbitrary (but finite) set of inputs $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, we reason about this ignorance by assuming that the function values, or outputs, $\{f(x^{(1)}), f(x^{(2)}), \ldots, f(x^{(n)})\}$ are distributed according to a multivariate Gaussian distribution. Since we are free to choose the inputs $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ arbitrarily, and the Gaussian assumption holds for any collection of inputs, this implicitly gives us a distribution for *all possible inputs*. In other words, we obtain a probabilistic model for the function f itself. Note that we now reason probabilistically about the function f in a way similar to how we probabilistically reasoned about the parameters β in the probabilistic linear regression.

So far, we have only talked about assuming *some* multivariate Gaussian distribution over $f(x)$ and $f(x')$, but not specified its mean μ or covariance matrix $\mathbf{K}$. One choice would be $\mu = 0$ and a covariance matrix $\mathbf{K}$ with only diagonal elements. That would be a *white* Gaussian process, implying that there is no correlation between $f(x)$ and $f(x')$, and such an assumption would be of very little help when reasoning about f in a regression setting. Instead, we need a way to construct a mean vector and a covariance matrix which adhere to the various properties that we might require from f, such as smoothness and trends. For instance, if we evaluate f at two points x and x' which are very close in the input space, then we would expect that $f(x)$ and $f(x')$ are strongly correlated (if the function f is assumed to be continuous, which is often the case). At the same time, we need this construction to generalize in a natural way to an arbitrary selection (and number) of inputs for it to be applicable to the definition of the Gaussian process above.

This can be accomplished by defining the mean vector and the covariance matrix by using a so called *mean function* $m(x)$ and a *covariance function* (or kernel) $k(x, x')$, and defining the joint distribution of $f(x)$ and $f(x')$ as:

$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \sim \mathcal{N}\left(\underbrace{\begin{pmatrix} m(x) \\ m(x') \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix}}_{\mathbf{K}} \right). \tag{3.2}$$

The covariance function $k(x, x')$ can be interpreted as a measure of the correlation level between the two inputs x and x'. The choice of covariance function is important, and we will later come back to different alternatives. It is often sensible to let it be a function of the distance between x and x', $r = \|x - x'\|$, and one popular choice which

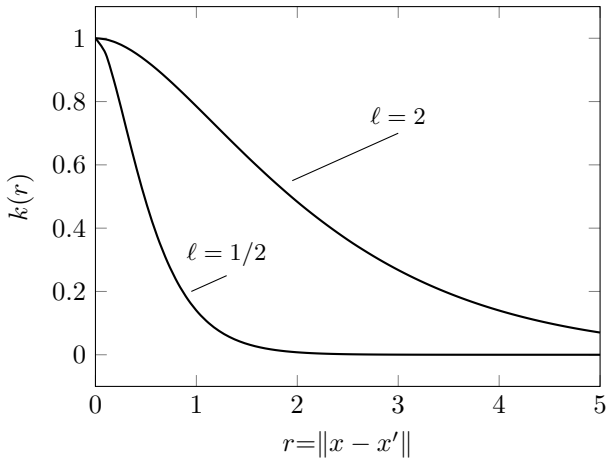


Figure 3.3: The Matérn 3 covariance function (3.3) for two different length scales ℓ.

we will use as an example is the Matérn 3 covariance function

$$k(x, x') = k(\underbrace{\|x - x'\|}_{=r}) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \tag{3.3}$$

where σ_f^2 is a scaling parameter and ℓ is referred to as the length scale, see Figure 3.3. A main characteristic of this, and many other covariance functions, is that it decays as r increases: It encodes the assumption that $f(1)$ tells more about $f(1.1)$ than $f(3)$, for instance. It is, however, possible to construct covariance functions with other properties as well, as we will come back to in Section 3.3. The mean function $m(x)$ can be used to encode any *a priori* knowledge about the shape of f. For instance, if we have reason to believe that f has a linear trend, then $m(x) = ax$ for some parameter a could be used to describe this knowledge. However, the mean function is often not needed and the choice $m(x) = 0$ works well in many cases.

We have now introduced the Gaussian process as a way to reason about the unknown function f. Technically, we assume that f is a realization of a Gaussian process, for which we will use the shorthand

$$f \sim \mathcal{GP}(m, k). \tag{3.4}$$

In other words, we assign a prior "distribution" for the function f, given by the Gaussian process. In fact, the red dashed lines in Figure 3.1b and 3.2 were samples drawn from this prior distribution. The power of the Gaussian process assumption will become clear when we do what we usually do with probability distributions—conditioning on data, or equivalently, computing the posterior. When we condition the Gaussian process on the observed data, we will force the red dashed lines to pass through the data points.

3.2 Gaussian process regression—computing the posterior

With the Gaussian process, we reason about the unknown f by modeling its output values $\{f(x^{(1)}), f(x^{(2)}), \ldots, f(x^{(n)})\}$ (for the inputs $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$) as jointly Gaussian distributed. *Now, what if $x^{(i)}$, and accordingly $f(x^{(i)})$, is a point in our set with observed training data?*

Before answering the question, let us replace the arbitrary set of inputs $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ with $\{x_1, x_2, \ldots, x_N, x_\star\}$, where $\{x_1, \ldots, x_N\}$ are the inputs in our training data set, and $x_\star$ is some arbitrary test input. We now have

$$\left(\begin{array}{c} f(x_1) \\ \vdots \\ f(x_N) \\ \hdashline f(x_\star) \end{array}\right) \sim \mathcal{N}\left(\left(\begin{array}{c} m(x_1) \\ \vdots \\ m(x_N) \\ \hdashline m(x_\star) \end{array}\right), \left(\begin{array}{ccc:c} k(x_1,x_1) & \cdots & k(x_1,x_N) & k(x_1,x_\star) \\ \vdots & \ddots & \vdots & \vdots \\ k(x_N,x_1) & \cdots & k(x_N,x_N) & k(x_N,x_\star) \\ \hdashline k(x_\star,x_1) & \cdots & k(x_\star,x_N) & k(x_\star,x_\star) \end{array}\right)\right), \tag{3.5}$$

or in a more compact notation

$$\begin{pmatrix} \mathbf{f} \\ f(x_\star) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{X}) \\ m(x_\star) \end{pmatrix}, \begin{pmatrix} k(\mathbf{X},\mathbf{X}) & k(\mathbf{X},x_\star) \\ k(x_\star,\mathbf{X}) & k(x_\star,x_\star) \end{pmatrix}\right), \tag{3.6}$$

where we let $k(x_\star, \mathbf{X})$ denote the matrix $\begin{pmatrix} k(x_\star,x_1) & \cdots & k(x_\star,x_N) \end{pmatrix}$, etc.

With the notation in place, we are ready to answer the above question: What can be said about $f(x_\star)$ if we have observed $\mathbf{f}$? Since these variables are jointly Gaussian according to (3.6), the answer follows directly from Theorem 2,

$$f(x_\star) \,|\, \mathbf{f} \sim \mathcal{N}\Big(m(x_\star) + k(x_\star,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}(\mathbf{f} - m(\mathbf{X})),\; k(x_\star,x_\star) - k(x_\star,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x_\star)\Big). \tag{3.7}$$

This result seems rather technical, but the illustration of it in Figures 3.4 and 3.5 is perhaps more intuitive: In Figure 3.4 we show the conditional distribution for $f(x)$ conditioned on the observations $\mathbf{f}$, for three different values of x_*. In Figure 3.5 we have taken so many values of x_* that it appears to the eye as a continuous line, and illustrated the Gaussian density by changing the color intensity. This provides an illustration of the posterior distribution for the entire function f.

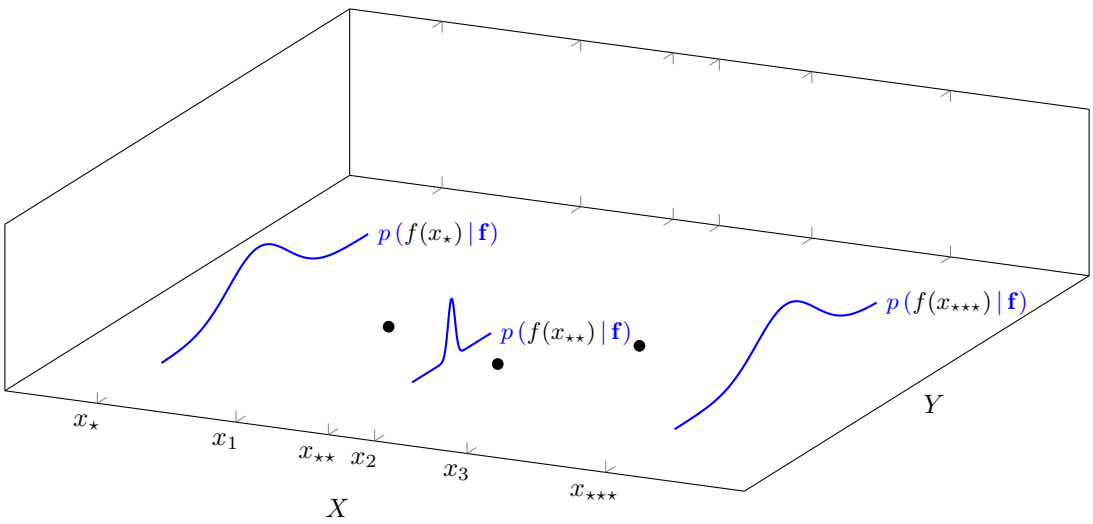


Figure 3.4: The distribution of $f(x_\star)$, $f(x_{\star\star})$ and $f(x_{\star\star\star})$ for the three inputs $x_\star$, $x_{\star\star}$ and $x_{\star\star\star}$, conditional on the observed values $\mathbf{f}$, i.e., $f(x_1)$, $f(x_2)$ and $f(x_3)$.

In the regression problem defined at the beginning of Section 3.1 we modeled the observations as $y_i = f(x_i) + \varepsilon_i$, where ε_i is some noise. In the expressions above, however, we have assumed that we instead observed $f(x_i)$ directly, i.e. without the noise term. Not including the noise term in the model would imply that we expect exactly the same measurement whenever the input is the same. In many real-world problems, that is not the case, and there are indeed certain errors not captured by the model which can only be described as noise. Fortunately, the incorporation of noise in the Gaussian process model is straightforward: if the assumptions (prior to observing

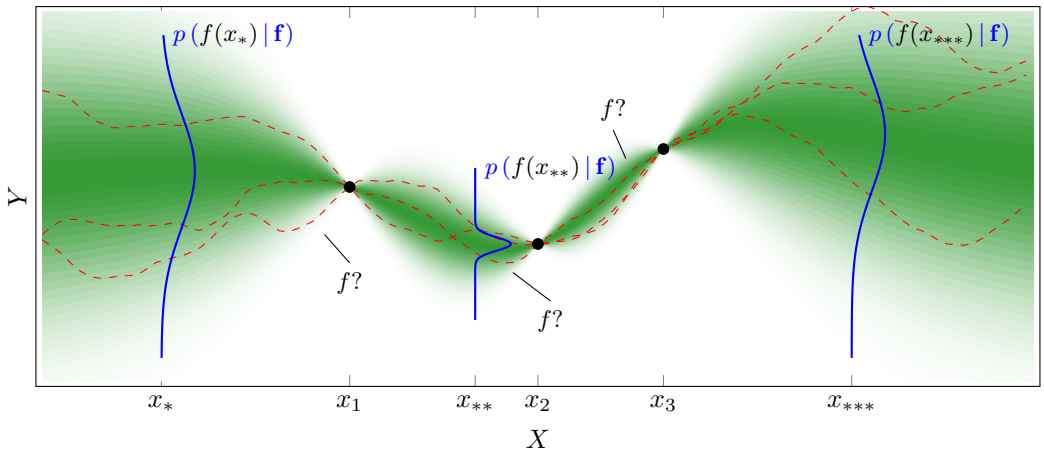


Figure 3.5: The same situation as in Figure 3.4, but we have now evaluated $p(f(x_\star)\,|\,\mathbf{f})$ for every pixel on the screen or every dot in the printer and used the color density to illustrate the Gaussian density. In addition, we have also plotted three samples (dotted red) from the distribution, which all passes through the data points now (cf. Figure 3.1a). The distributions from Figure 3.4 are also overlaid for reference.

the data) are $\mathbf{f} \sim \mathcal{N}\left(m(X), k(\mathbf{X},\mathbf{X})\right)$ and $\varepsilon \sim \mathcal{N}(0,\sigma^2)$, then $\mathbf{y} \sim \mathcal{N}\left(m(X), k(\mathbf{X},\mathbf{X}) + \sigma^2 I_N\right)$. We can thus write (3.6) and (3.7) including the noise ε as

$$\begin{pmatrix} \mathbf{y} \\ f(x_\star) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{X}) \\ m(x_\star) \end{pmatrix}, \begin{pmatrix} k(\mathbf{X},\mathbf{X}) + \sigma^2 I_N & k(\mathbf{X},x_\star) \\ k(x_\star,\mathbf{X}) & k(x_\star,x_\star) \end{pmatrix}\right), \tag{3.8}$$

and

$$f(x_\star)\,|\,\mathbf{y} \sim \mathcal{N}\big(m(x_\star) + \mathbf{s}^\mathsf{T}(\mathbf{y} - m(\mathbf{X})),\; k(x_\star,x_\star) - \mathbf{s}^\mathsf{T} k(\mathbf{X},x_\star)\big), \tag{3.9}$$

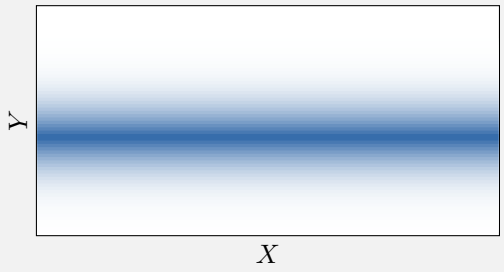
where, for notational brevity, we have introduced the vector $\mathbf{s}$ as

$$\mathbf{s}^\mathsf{T} = k(x_\star,\mathbf{X})(k(\mathbf{X},\mathbf{X}) + \sigma^2 I_N)^{-1}. \tag{3.10}$$

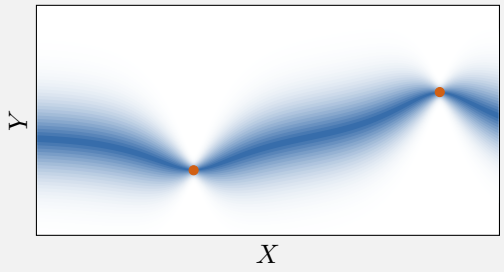
This equation, (3.9), is the real workhorse in Gaussian process regression. We illustrate the use of it in practice with Example 3.1.

Example 3.1: The Gaussian process as a regression model

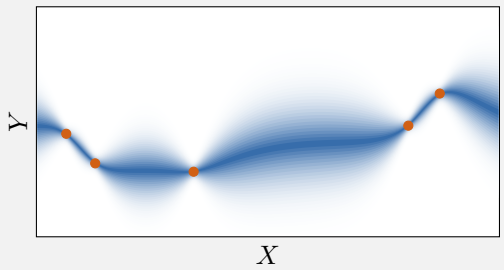
We start with a Gaussian process prior over the unknown function f, illustrated with a shaded blue plot (the darker blue, the higher probability density). The prior is completely determined by the mean and covariance functions, here takes as zero and the Matérn 3, respectively.



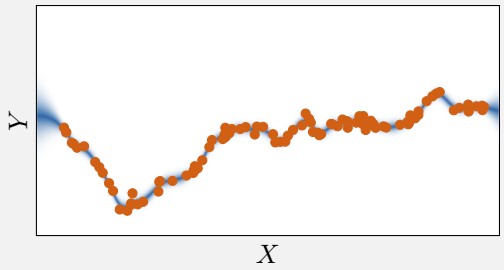
After having observed two data points $\{x_1, y_1\}$ and $\{x_2, y_2\}$ (orange dots), we condition the Gaussian process distribution over f on the observed data. We now have a distribution which looks like



Note that the posterior distribution is obtained by evaluating (3.7) for each point on the x-axis (on some fine grid). After 3 additional observed data points, we compute the distribution for f conditioned on all observations so far. Note that the uncertainty is much smaller in regions where data is observed, and larger where we have not observed any data yet.



Finally, the distribution for f conditioned on 100 observations.



3.3 Design choices: covariance functions

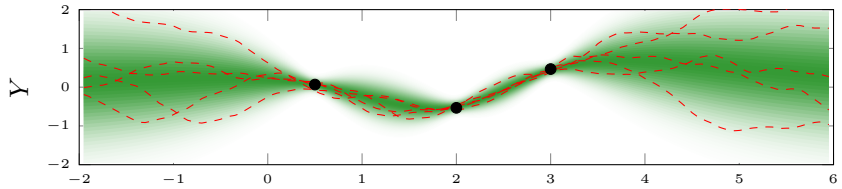
The choice of covariance function is important, as it encodes assumptions made about f. Some common covariance functions are listed in Table 3.1, and exemplified in Figure 3.6. New covariance functions can be constructed by adding or multiplying the covariance functions in the table.

Name	**Covariance function** $k(x, x')$	**Description**
Squared exponential (SE)	$\sigma_f^2 \exp\left(-\frac{1}{2l^2} r^2\right)$	Generates infinitely differentiable (i.e., extremely smooth) functions. Also called exponentiated quadratic.
Linear (LI)	$\sigma_b^2 + \sigma_v^2 (x - c)(x' - c)$	The offset c determines the x-coordinate that all lines go through. In the context of GPs it is mainly useful in combination with other covariance functions.
Exponential (Exp)	$\sigma_f^2 \exp\left(-\frac{r}{\ell}\right)$	Generates continuous but non-differentiable functions.
Matérn 3 (M3)	$\sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right)$	Generates one-time differentiable functions.
Matérn 5 (M5)	$\sigma_f^2 \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right)$	Generates two-times differentiable functions.
Periodic (Per)	$\sigma_f^2 \exp\left(-\frac{2}{\ell^2} \sin^2\left(\frac{\pi r}{p}\right)\right)$	Produce functions that are periodic with a period p. Hence, the distance between exact repetitions of the function is given by p.
$(r = \|x - x'\|)$		

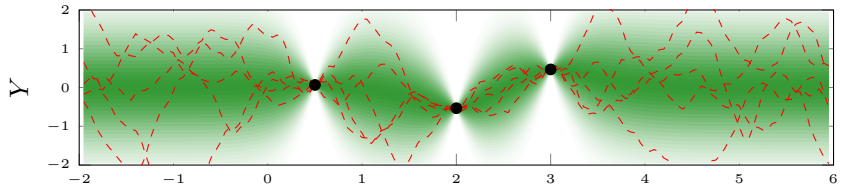
Table 3.1: Some commonly used covariance functions. (The words "continuous" and "differentiable" above should be interpreted in a mean-square sense, as f is a stochastic process.)

3.4 Further reading

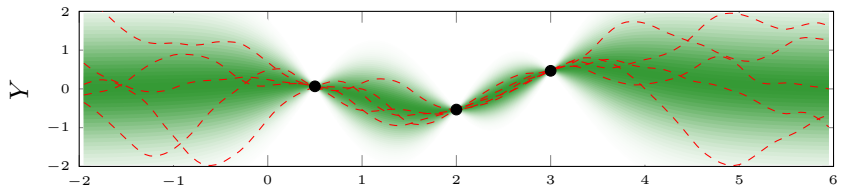
On the historical side it is interesting to mention that the Gaussian process was popularized under the name of *Kriging* within the field of geostatistics. The name stems from the South African Engineer Daniel Krige who made use of the Gaussian process to estimate the distribution of gold based on findings from a few boreholes. This is documented in his Master's thesis (Krige, 1951). Today the Gaussian process is used for countless application and a solid text-book introduction is provided by Rasmussen and Williams (2006).



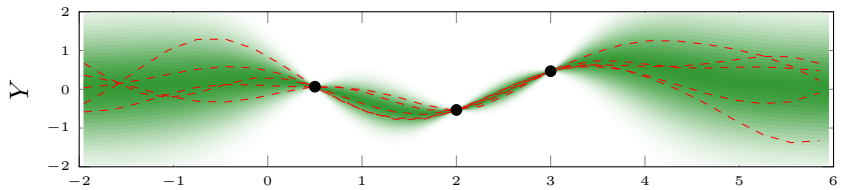
(a) Matérn 3, $\ell = 2$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



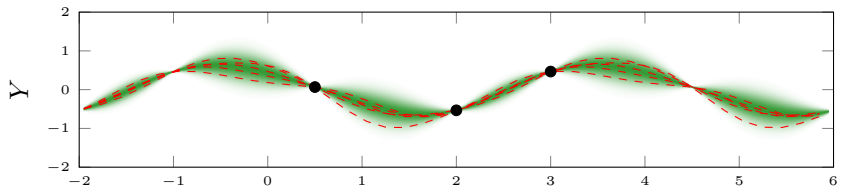
(b) Matérn 3, $\ell = 1/2$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



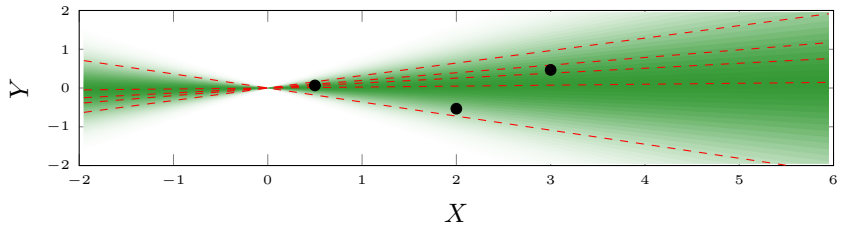
(c) Matérn 5, $\ell = 1$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



(d) Squared exponential, $\ell = 2$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



(e) Periodic, $\ell = 2$, $p = 4$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



(f) Linear, $c = 0$, $\sigma_b^2 = 0$, $\sigma_v^2 = 0$, $\sigma^2 = 1$

Figure 3.6: The posterior when using some covariance functions, and also some samples from them.

Appendix A

Multivariate Gaussian distribution

The multivariate Gaussian distribution is the most important and the most commonly used probability distribution for continuous random variables. We will from now on refer to the multivariate Gaussian simply as the Gaussian and let the context decide if it is the scalar or the multivariate case that is relevant.

An appealing and highly useful property of the Gaussian is that it is preserved under many different transformations. As a first example of this we will in Section A.1 see that an affine transformation of a Gaussian is still a Gaussian. Other commonly used transformations that preserve Gaussianity is marginalization and conditioning which are both studied in detail in Section A.2. Finally, we will see that marginalization and conditioning in the presence of an affine transformation will also preserve the Gaussian nature.

A.1 Definition and geometry

The multivariate Gaussian is an extension of the univariate (scalar) Gaussian distribution to vector-valued random variables. To see this we will in Example 1.1 investigate what happens when we study the joint distribution of two independent scalar Gaussian random variables. Let us just first recall that the scalar Gaussian probability density function $p(x)$ for a scalar $X \sim \mathcal{N}(\mu, \sigma^2)$ is defined as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}(x-\mu)\sigma^{-2}(x-\mu)\right), \tag{A.1}$$

where we commonly refer to $Z = 1/\sqrt{2\pi\sigma^2}$ as the normalization constant.

Example 1.1: Joint distribution of two independent scalar Gaussian random variables

Let us assume that we have two independent scalar Gaussian random variables $X_a \sim \mathcal{N}(\mu_a, \sigma_a)$ and $X_b \sim \mathcal{N}(\mu_b, \sigma_b)$, meaning that if we know something about x_a this does not tell us anything about X_b and the other way around. Let us now form the vector $X = \begin{pmatrix} X_a & X_b \end{pmatrix}^\mathsf{T}$ and find the joint distribution for X_a and X_b, i.e. $p(x)$. The fact that the variables X_a and X_b are independent implies that $p(x) = p(x_a)p(x_b)$, since the joint distribution of two or more independent random variables is given by the product of the distributions of the individual variables. Hence,

$$\begin{aligned} p(x) &= \frac{1}{Z_a}\exp\left(-\frac{(x_a-\mu_a)^2}{2\sigma_a^2}\right)\frac{1}{Z_b}\exp\left(-\frac{(x_b-\mu_b)^2}{2\sigma_b^2}\right) = \frac{1}{Z_aZ_b}\exp\left(-\frac{(x_a-\mu_a)^2}{2\sigma_a^2}-\frac{(x_b-\mu_b)^2}{2\sigma_b^2}\right) \\ &= \frac{1}{Z_aZ_b}\exp\left(-\frac{1}{2}\begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}^\mathsf{T}\begin{pmatrix} 1/\sigma_a^2 & 0 \\ 0 & 1/\sigma_b^2 \end{pmatrix}\begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}\right) \\ &= \frac{1}{Z}\exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right) \end{aligned} \tag{A.2}$$

where Z_a, Z_b and $Z = Z_aZ_b$ denotes the normalization constants in the corresponding Gaussian distributions and

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}. \tag{A.3}$$

Hence, the joint distribution (A.2) of the two independent Gaussian random variables X_a and X_b has the same form as the scalar Gaussian distribution (A.1), save for the fact that the mean value is now a vector μ and the variance is instead a matrix Σ that we refer to as a covariance matrix. This is in fact a first instance of the multivariate Gaussian.

Recall that covariance is a measure of the *joint variability* of two random variables. Our random variables X_a and X_b in this example are independent, meaning that they are completely uncorrelated. Hence, even if we have some information about one of these variables that information is not revealing any information about the other variable. The diagonal covariance matrix (A.3) is encoding exactly this information. In general, the covariance matrix Σ of a Gaussian random vector with independent components is diagonal.

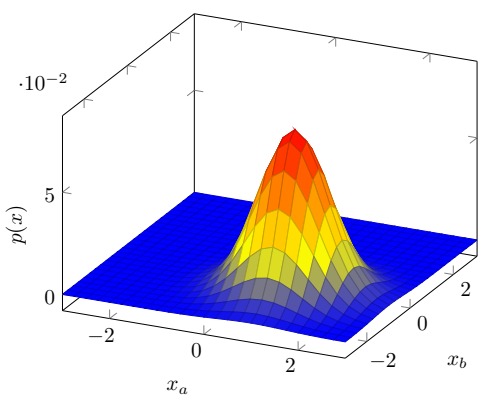
Definition 2 (Multivariate Gaussian)**.** *A random variable $X \in R^p$ with $\mathbb{E}(X) = \mu$ and $\mathrm{Cov}(X) = \Sigma$ such that $\det\Sigma > 0$ is a multivariate Gaussian if and only if the density is*

$$p(x) = \mathcal{N}(x \,|\, \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right), \qquad x \in R^p. \tag{A.4}$$

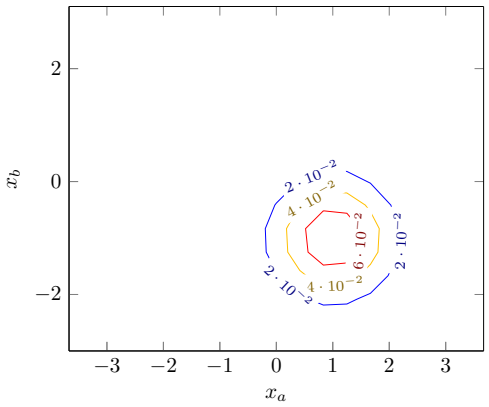
The Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is uniquely determined by its mean vector μ and covariance matrix Σ. For intuition it is helpful to think of the Gaussian distribution as consisting of a normalization constant $Z = 1/(2\pi)^{p/2}\sqrt{\det\Sigma}$ times the exponential of a quadratic form $q(x) = (x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)$, i.e.

$$\text{Gaussian} \propto e^{\text{quadratic form}}. \tag{A.5}$$

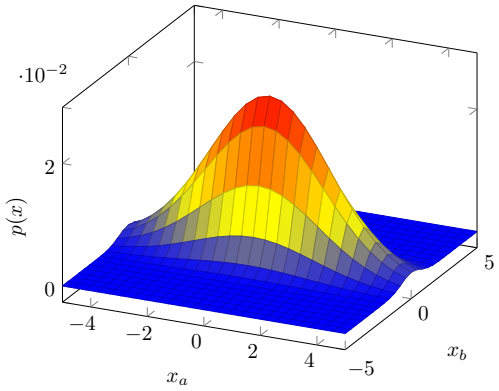
In Figure A.1 we provide a plot of the multivariate Gaussian that was examined in Example 1.1 for particular values of μ and Σ.



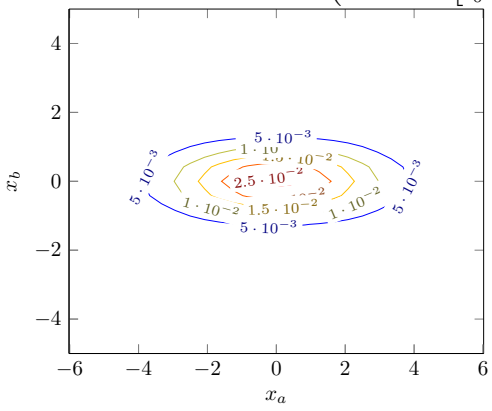
(a) 3D plot of $p(x) = \mathcal{N}\left(x\middle|\ \begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}1^2 & 0\\0 & 1^2\end{bmatrix}\right)$



(b) Contour plot of $p(x) = \mathcal{N}\left(x\middle|\ \begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}1^2 & 0\\0 & 1^2\end{bmatrix}\right)$



(c) 3D plot of $p(x) = \mathcal{N}\left(x\middle|\ \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}3^2 & 0\\0 & 1^2\end{bmatrix}\right)$



(d) Contour plot of $p(x) = \mathcal{N}\left(x\middle|\ \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}3^2 & 0\\0 & 1^2\end{bmatrix}\right)$

Figure A.1: A 3D plot and a countourplot of two different two-dimensional Guassians distributions as presented in Example 1.1. In Figure A.1a and A.1b $\mu_a = 1$, $\mu_b = -1$, $\sigma_a = 1$, $\sigma_b = 1$ and in Figure A.1c and A.1d $\mu_a = 0$, $\mu_b = 0$, $\sigma_a = 3$, $\sigma_b = 1$.

In general, the level sets of a quadratic form (when Σ is a positive semi-definite matrix) are ellipsoids described by the equation $q(x) = (x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu) = \text{const}$.

A very useful fact when it comes to Gaussian random vectors is that any affine transformation

$$Y = AX + b, \qquad A \in R^{p\times p}, b \in R^p, \tag{A.6}$$

of a Gaussian random variable $X \sim \mathcal{N}(\mu, \Sigma)$ results in random variable Y that is *also* Gaussian. The mean value and covariance matrix of the result of the affine transform are given by

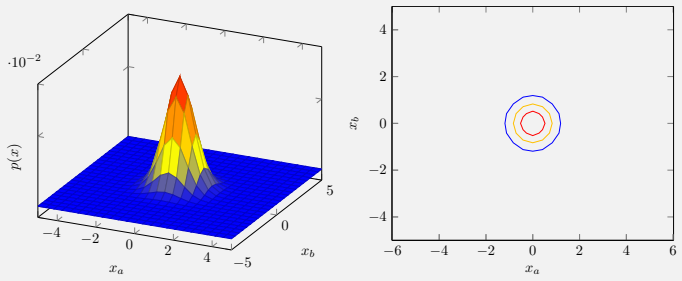
$$\mathbb{E}(Y) = \mathbb{E}(AX + b) = A\mathbb{E}(X) + b = A\mu + b, \tag{A.7a}$$

$$\begin{aligned}\mathrm{Cov}(Y) &= \mathbb{E}\left(Y - \mathbb{E}(Y)\right)\left(Y - \mathbb{E}(Y)\right)^\mathsf{T} = \mathbb{E}\left(AX - A\mu\right)\left(AX - A\mu\right)^\mathsf{T}\\ &= A\mathbb{E}\left((X-\mu)(X-\mu)^\mathsf{T}\right)A^\mathsf{T} = A\Sigma A^\mathsf{T}.\end{aligned} \tag{A.7b}$$

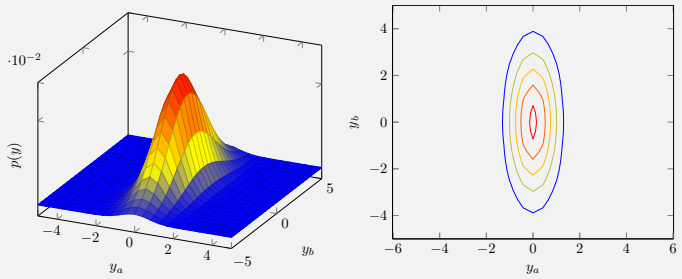
This is illustrated in Example 1.2.

Example 1.2: The geometry of the Gaussian distribution

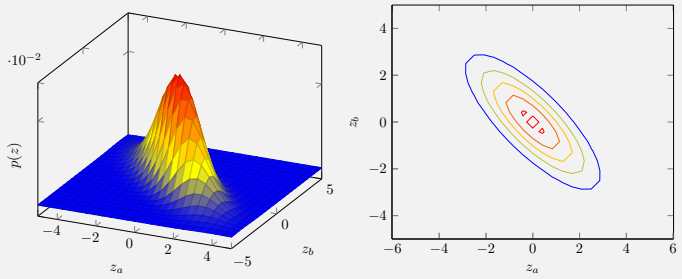
Consider a two-dimensional Gaussian random variable $X \sim \mathcal{N}(x|\ \mu,\ \Sigma)$ where $\mu = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$ and $\Sigma = \left[\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right]$.



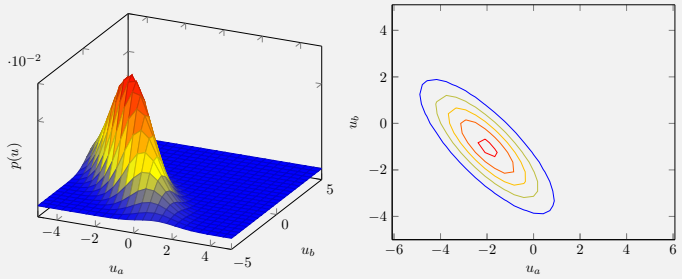
Perform a linear transformation $Y = A_1 X$ where $A_1 = \left[\begin{smallmatrix}1 & 0\\0 & 3\end{smallmatrix}\right]$. The random variable Y will then also be Gaussian distributed with $Y \sim \mathcal{N}\left(y|\ \mu,\ A_1 \Sigma A_1^{\mathsf{T}}\right) = \mathcal{N}\left(y|\ \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right],\ \left[\begin{smallmatrix}1 & 0\\0 & 9\end{smallmatrix}\right]\right)$, i.e., the distribution is scaled in y_b direction.



Perform another linear transformation $Z = A_2 Y$, this time a rotation of $45°$ where $A_2 = \begin{bmatrix}\cos(45°) & -\sin(45°)\\ \sin(45°) & \cos(45°)\end{bmatrix}$. The random variable Y will now be distributed as $Z \sim \mathcal{N}\left(z|\ \mu,\ A_2 A_1 \Sigma A_1^{\mathsf{T}} A_2^{\mathsf{T}}\right) = \mathcal{N}\left(z|\ \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right],\ \left[\begin{smallmatrix}5 & -4\\-4 & 5\end{smallmatrix}\right]\right)$. Consequently, also the distribution will be rotated.



Finally, consider a translation with $U = Z + b$ where $b = \left[\begin{smallmatrix}-2\\-1\end{smallmatrix}\right]$. The final distribution will be $U \sim \mathcal{N}\left(u|\ \mu + b,\ A_2 A_1 \Sigma A_1^{\mathsf{T}} A_2^{\mathsf{T}}\right) = \mathcal{N}\left(u|\ \left[\begin{smallmatrix}-2\\-1\end{smallmatrix}\right],\ \left[\begin{smallmatrix}5 & -4\\-4 & 5\end{smallmatrix}\right]\right)$, i.e., the distribution will be shifted accordingly.



The development in Example 1.2 can alternatively be interpreted as a way of constructing an arbitrary Gaussian from the standard Gaussian $\mathcal{N}(0, I_p)$.

A.2 Marginalization and conditioning of partitioned Gaussians

Given two (possibly vector-valued) random variables $X_a \in R^{n_a}$ and $X_b \in R^{n_b}$ that are jointly Gaussian, we will now establish two important facts. The first fact is that the *marginal distribution of either variable is Gaussian*. The second fact is that the *conditional distribution for one variable given the other variable is Gaussian*. Let us start by assuming the joint distribution $p(x_a, x_b)$ is $X \sim \mathcal{N}(\mu, \Sigma)$, where

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}. \tag{A.8}$$

Since the covariance matrix Σ is symmetric, we must have $\Sigma_{ba} = \Sigma_{ab}^{\mathsf{T}}$. Marginalization amounts to finding the distribution of some of the variables—say X_a—by removing the remaining variables from the joint distribution $p(x_a, x_b)$ by integrating them out according to

$$p(x_a) = \int p(x_a, x_b) \mathrm{d}x_b. \tag{A.9}$$

The simplest way of solving this integral is probably an indirect approach where we start by noting that we can obtain X_a from X by the following linear transformation $X_a = AX$, where $A = \begin{pmatrix} I_{n_a} & 0_{n_b} \end{pmatrix}$. Here I_{n_a} denotes an identity matrix of dimension n_a and 0_{n_b} denotes a matrix full of zeros of dimension n_b. We know that a linear transformation of a Gaussian random variable results in another Gaussian random variable, but with a new mean and covariance according to (A.7). Hence, the prior distribution $p(x_a)$ is given by $\mathcal{N}(A\mu, A\Sigma A^{\mathsf{T}})$, where

$$A\mu = \begin{pmatrix} I_{n_a} & 0_{n_b} \end{pmatrix} \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_a, \tag{A.10}$$

$$A\Sigma A^{\mathsf{T}} = \begin{pmatrix} I_{n_a} & 0_{n_b} \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} I_{n_a} \\ 0_{n_b} \end{pmatrix} = \Sigma_{aa}. \tag{A.11}$$

The above development is summarized in Theorem 1. An alternative way of proving this result is via brute force calculations by inserting (A.4)—with x, μ and Σ according to (A.8)—into (A.9).

Theorem 1. **(Marginalization)** Partition the Gaussian random vector $X \in \mathcal{N}(\mu, \Sigma)$ according to (A.8). The marginal density $p(x_a)$ is then given by

$$p(x_a) = \mathcal{N}\left(x_a \,|\, \mu_a, \Sigma_{aa}\right). \tag{A.12}$$

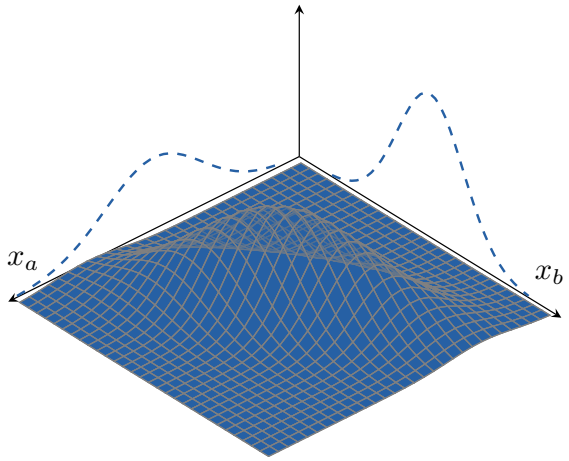
If we measure one variable that in turns depends on another variable, we are often interested in knowing what this measurement can tell us about the unmeasured variable. This is handled using conditioning and for partitioned Gaussian variables the highly useful result is provided in Theorem 2.

Theorem 2. **(Conditioning)** Partition the Gaussian random vector $X \in \mathcal{N}(\mu, \Sigma)$ according to (A.8). The conditional density $p(x_a \,|\, x_b)$ is then given by

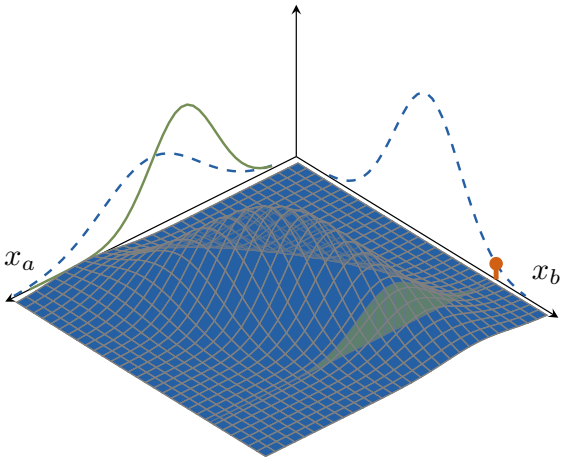
$$p(x_a \,|\, x_b) = \mathcal{N}\left(x_a \,\middle|\, \mu_{a\,|\,b}, \Sigma_{a\,|\,b}\right), \tag{A.13a}$$

$$\mu_{a\,|\,b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \tag{A.13b}$$

$$\Sigma_{a\,|\,b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}. \tag{A.13c}$$



(a) A two-dimensional Gaussian distribution for the random variables X_a and X_b, with a blue surface plot for the density, and the marginal distribution for each component sketched using dashed blue lines along each axis. Note that the marginal distributions do *not* contain all information about the distribution of X_a and X_b, since the covariance information is lacking in that representation.



(b) The conditional distribution of X_a (green line), when X_b is observed (orange dot). The conditional distribution of x_a is given by (A.13), which (apart from a normalizing constant) in this graphical representation also is the green 'slice' of the joint distribution (blue surface). The marginals of the joint distribution from Figure A.2a are kept for reference (blue dashed lines).

Figure A.2: A two-dimensional multivariate Gaussian distribution for x_a and x_b in (a), and the conditional distribution for x_a, when a particular value of x_b is observed, in (b).

A.3 Affine transformations of partitioned Gaussians

In Section A.2 we introduced the idea of partitioned Gaussian densities, and derived the expressions for the marginal and conditional densities expressed in terms of the parameters of the joint density. We shall now take a different starting point, namely that we are given the marginal density $p(x_a)$ and the conditional density $p(x_b \,|\, x_a)$ and derive expressions for the joint density $p(x_a, x_b)$, the marginal density $p(x_b)$ and the conditional density $p(x_a \,|\, x_b)$.

Theorem 3. **(Affine transformation)** Assume that X_a, as well as X_b conditioned on X_a, are Gaussian distributed according to

$$p(x_a) = \mathcal{N}\left(x_a \,|\, \mu_a,\, \Sigma_a\right), \tag{A.14a}$$

$$p(x_b \,|\, x_a) = \mathcal{N}\left(x_b \,\middle|\, Mx_a + b,\, \Sigma_{b\,|\,a}\right), \tag{A.14b}$$

where M is a matrix (of appropriate dimension) and b is a constant vector. The joint distribution of X_a and X_b is then given by

$$p(x_a, x_b) = \mathcal{N}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_a \\ M\mu_a + b \end{pmatrix}, R\right), \tag{A.14c}$$

with

$$R = \begin{pmatrix} M^{\mathsf{T}}\Sigma_{b\,|\,a}^{-1}M + \Sigma_a^{-1} & -M^{\mathsf{T}}\Sigma_{b\,|\,a}^{-1} \\ -\Sigma_{b\,|\,a}^{-1}M & \Sigma_{b\,|\,a}^{-1} \end{pmatrix} = \begin{pmatrix} \Sigma_a & \Sigma_a M^{\mathsf{T}} \\ M\Sigma_a & \Sigma_{b\,|\,a} + M\Sigma_a M^{\mathsf{T}} \end{pmatrix}^{-1}. \tag{A.14d}$$

Combining the results in Theorems 1, 2 and 3 we also get the following corollary.

Corollary 1. **(Affine transformation – marginal and conditional)** Assume that X_a, as well as X_b conditioned on X_a, are Gaussian distributed according to

$$p(x_a) = \mathcal{N}\left(x_a \,|\, \mu_a,\, \Sigma_a\right), \tag{A.15a}$$

$$p(x_b \,|\, x_a) = \mathcal{N}\left(x_b \,\middle|\, Mx_a + b,\, \Sigma_{b\,|\,a}\right), \tag{A.15b}$$

where M is a matrix (of appropriate dimension) and b is a constant vector. The marginal density of X_b is then given by

$$p(x_b) = \mathcal{N}\left(x_b \,|\, \mu_b,\, \Sigma_b\right), \tag{A.15c}$$

with

$$\mu_b = M\mu_a + b, \tag{A.15d}$$

$$\Sigma_b = \Sigma_{b\,|\,a} + M\Sigma_a M^{\mathsf{T}}. \tag{A.15e}$$

The conditional density of X_a given X_b is

$$p(x_a \,|\, x_b) = \mathcal{N}\left(x_a \,\middle|\, \mu_{a\,|\,b},\, \Sigma_{a\,|\,b}\right), \tag{A.15f}$$

with

$$\mu_{a\,|\,b} = \Sigma_{a\,|\,b}\left(M^{\mathsf{T}}\Sigma_{b\,|\,a}^{-1}(x_b - b) + \Sigma_a^{-1}\mu_a\right) = \mu_a + \Sigma_a M^{\mathsf{T}}\Sigma_b^{-1}(x_b - b - M\mu_a), \tag{A.15g}$$

$$\Sigma_{a\,|\,b} = \left(\Sigma_a^{-1} + M^{\mathsf{T}}\Sigma_{b\,|\,a}^{-1}M\right)^{-1} = \Sigma_a - \Sigma_a M^{\mathsf{T}}\Sigma_b^{-1}M\Sigma_a. \tag{A.15h}$$

Bibliography

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Efron, B. and Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press, 3 edition.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

Gut, A. (1995). *An Intermediate Course in Probability*. Springer-Verlag.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer, 2 edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Krige, D. G. (1951). A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master's thesis, University of Witwatersrand.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Murphy, K. P. (2012). *Machine learning – a probabilistic perspective*. MIT Press.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT press.