An introduction to Gaussian process regression

Andreas Lindholm Fredrik Lindsten Thomas B. Schön Niklas Wahlström

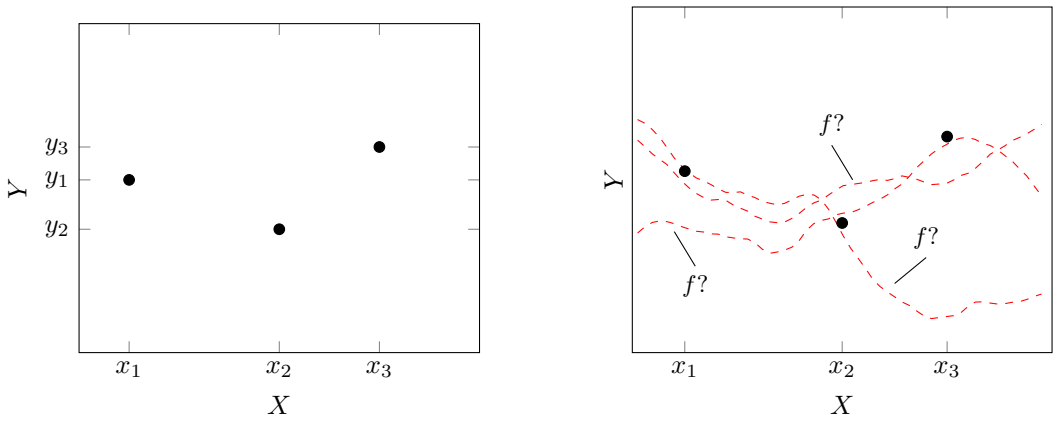
September 26, 2019

Chapter 1

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.

1.1 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 1.1: 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 1.1a. 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 1.1b.

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*

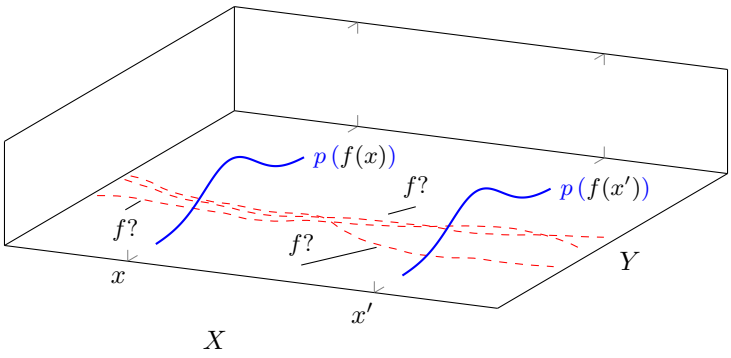


Figure 1.2: The function f is unknown to us, we have given it a pictorial representation by three dashed red lines. The Gaussian process assumption is to model f as random itself, and *assume* that the value of f for any two arbitrary inputs x and x' ($f(x)$ and $f(x')$ respectively) has a joint Gaussian 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:

$$\begin{pmatrix} f(x\,) \\ f(x') \end{pmatrix} \sim \mathcal{N}\left(\mu, \mathbf{K}\right), \tag{1.1}$$

We illustrate this by Figure 1.2. 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:

$$\begin{pmatrix} f(x\,) \\ f(x') \end{pmatrix} \sim \mathcal{N}\Bigg(\underbrace{\begin{pmatrix} m(x\,) \\ m(x') \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} k(x\,, x\,) & k(x\,, x') \\ k(x', x\,) & k(x', x') \end{pmatrix}}_{\mathbf{K}}\Bigg). \tag{1.2}$$

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

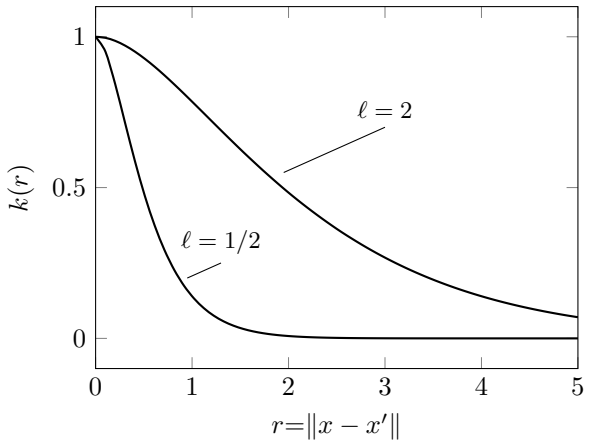


Figure 1.3: The Matérn 3 covariance function (1.3) for two different length scales ℓ.

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

$$k(x, x') = k(\underbrace{\|x - x'\|}_{=r}) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \tag{1.3}$$

where σ_f^2 is a scaling parameter and ℓ is referred to as the length scale, see Figure 1.3. 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 1.3. 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

$$f \sim \mathcal{GP}(m, k). \tag{1.4}$$

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 1.1b and 1.2 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.

1.2 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

$$\left(\begin{array}{c} f(x_1) \\ \vdots \\ f(x_N) \\ \hdashline f(x_\star) \end{array}\right) \sim \mathcal{N}\left(\left(\begin{array}{c} m(x_1) \\ \vdots \\ m(x_N) \\ \hdashline m(x_\star) \end{array}\right), \left(\begin{array}{ccc:c} k(x_1,x_1) & \cdots & k(x_1,x_N) & k(x_1,x_\star) \\ \vdots & \ddots & \vdots & \vdots \\ k(x_N,x_1) & \cdots & k(x_N,x_N) & k(x_N,x_\star) \\ \hdashline k(x_\star,x_1) & \cdots & k(x_\star,x_N) & k(x_\star,x_\star) \end{array}\right)\right), \tag{1.5}$$

or in a more compact notation

$$\begin{pmatrix} \mathbf{f} \\ f(x_\star) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{X}) \\ m(x_\star) \end{pmatrix}, \begin{pmatrix} k(\mathbf{X},\mathbf{X}) & k(\mathbf{X},x_\star) \\ k(x_\star,\mathbf{X}) & k(x_\star,x_\star) \end{pmatrix}\right), \tag{1.6}$$

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 (1.6), the answer follows directly from Theorem 2,

$$f(x_\star) \,|\, \mathbf{f} \sim \mathcal{N}\Big(m(x_\star) + k(x_\star,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}(\mathbf{f} - m(\mathbf{X})),\; k(x_\star,x_\star) - k(x_\star,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x_\star)\Big). \tag{1.7}$$

This result seems rather technical, but the illustration of it in Figures 1.4 and 1.5 is perhaps more intuitive: In Figure 1.4 we show the conditional distribution for $f(x)$ conditioned on the observations $\mathbf{f}$, for three different values of x_*. In Figure 1.5 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.

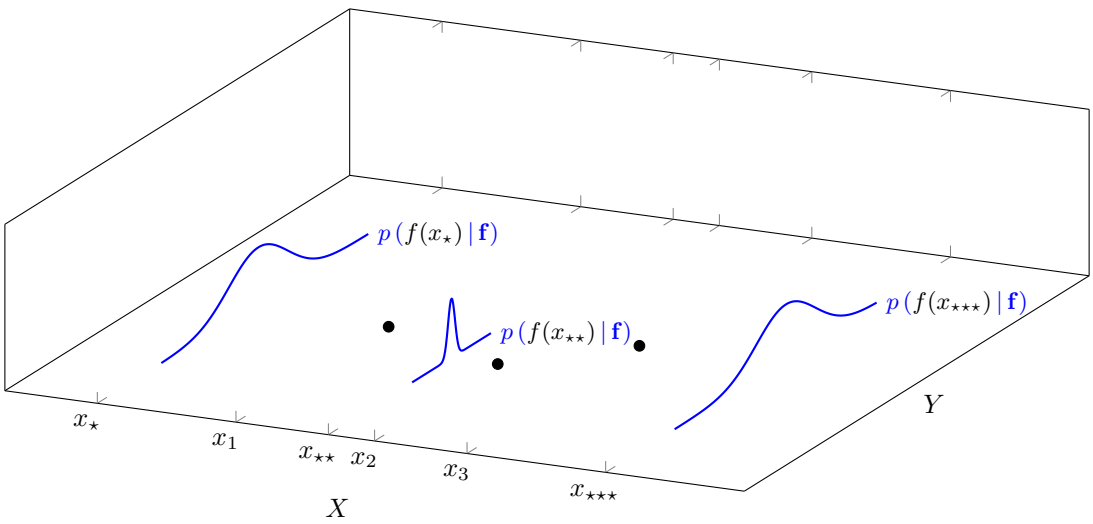


Figure 1.4: The distribution of $f(x_\star)$, $f(x_{\star\star})$ and $f(x_{\star\star\star})$ for the three inputs $x_\star$, $x_{\star\star}$ and $x_{\star\star\star}$, conditional on the observed values $\mathbf{f}$, i.e., $f(x_1)$, $f(x_2)$ and $f(x_3)$.

In the regression problem defined at the beginning of Section 1.1 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

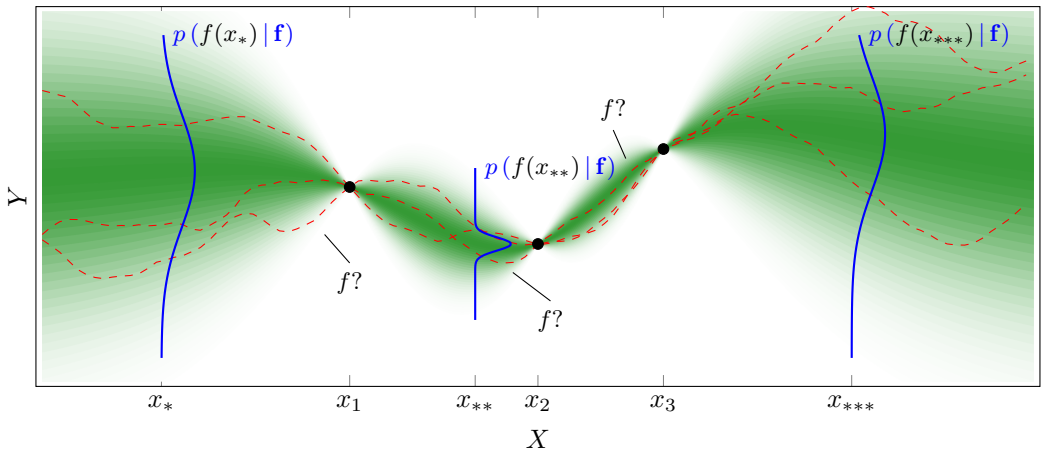


Figure 1.5: The same situation as in Figure 1.4, but we have now evaluated $p(f(x_\star)\,|\,\mathbf{f})$ for every pixel on the screen or every dot in the printer and used the color density to illustrate the Gaussian density. In addition, we have also plotted three samples (dotted red) from the distribution, which all passes through the data points now (cf. Figure 1.1a). The distributions from Figure 1.4 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 (1.6) and (1.7) including the noise ε as

$$\begin{pmatrix} \mathbf{y} \\ f(x_\star) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{X}) \\ m(x_\star) \end{pmatrix}, \begin{pmatrix} k(\mathbf{X},\mathbf{X}) + \sigma^2 I_N & k(\mathbf{X},x_\star) \\ k(x_\star,\mathbf{X}) & k(x_\star,x_\star) \end{pmatrix}\right), \tag{1.8}$$

and

$$f(x_\star)\,|\,\mathbf{y} \sim \mathcal{N}\big(m(x_\star) + \mathbf{s}^\mathsf{T}(\mathbf{y} - m(\mathbf{X})),\; k(x_\star,x_\star) - \mathbf{s}^\mathsf{T} k(\mathbf{X},x_\star)\big), \tag{1.9}$$

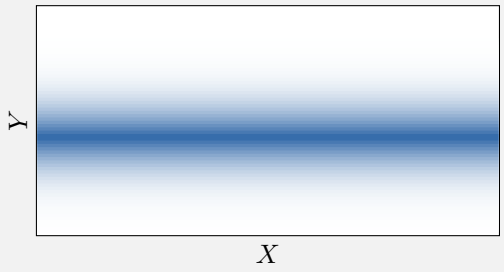
where, for notational brevity, we have introduced the vector $\mathbf{s}$ as

$$\mathbf{s}^\mathsf{T} = k(x_\star,\mathbf{X})(k(\mathbf{X},\mathbf{X}) + \sigma^2 I_N)^{-1}. \tag{1.10}$$

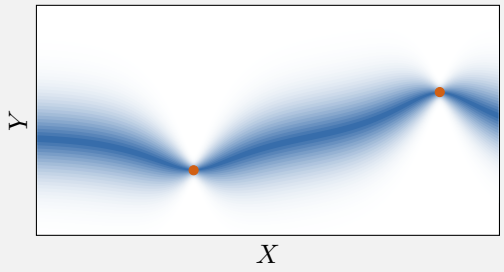
This equation, (1.9), is the real workhorse in Gaussian process regression. We illustrate the use of it in practice with Example 1.1.

Example 1.1: The Gaussian process as a regression model

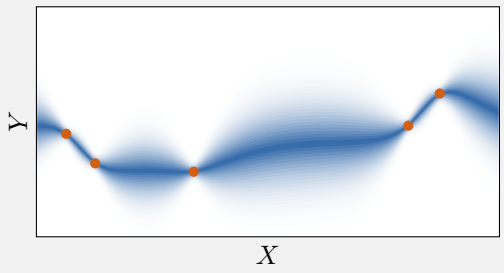
We start with a Gaussian process prior over the unknown function f, illustrated with a shaded blue plot (the darker blue, the higher probability density). The prior is completely determined by the mean and covariance functions, here takes as zero and the Matérn 3, respectively.



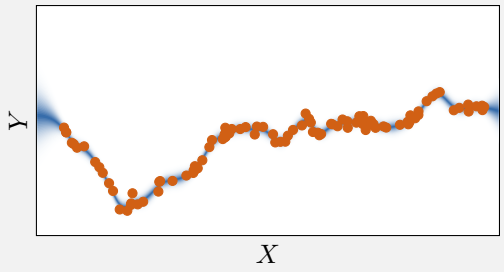
After having observed two data points $\{x_1, y_1\}$ and $\{x_2, y_2\}$ (orange dots), we condition the Gaussian process distribution over f on the observed data. We now have a distribution which looks like



Note that the posterior distribution is obtained by evaluating (1.7) 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.



1.3 Design choices: covariance functions

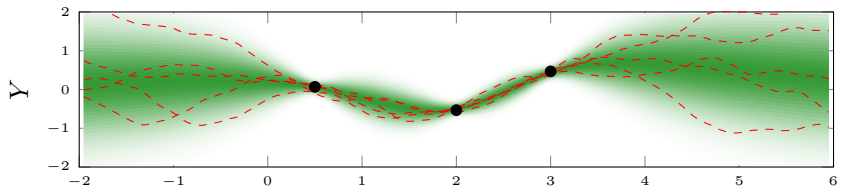
The choice of covariance function is important, as it encodes assumptions made about f. Some common covariance functions are listed in Table 1.1, and exemplified in Figure 1.6. New covariance functions can be constructed by adding or multiplying the covariance functions in the table.

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.
$(r = \|x - x'\|)$		

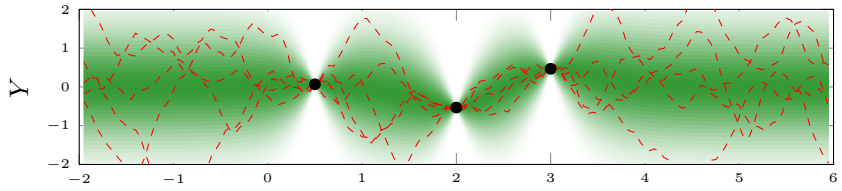
Table 1.1: Some commonly used covariance functions. (The words "continuous" and "differentiable" above should be interpreted in a mean-square sense, as f is a stochastic process.)

1.4 Further reading

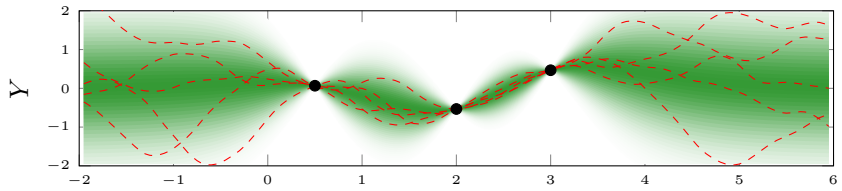
On the historical side it is interesting to mention that the Gaussian process was popularized under the name of *Kriging* within the field of geostatistics. The name stems from the South African Engineer Daniel Krige who made use of the Gaussian process to estimate the distribution of gold based on findings 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).



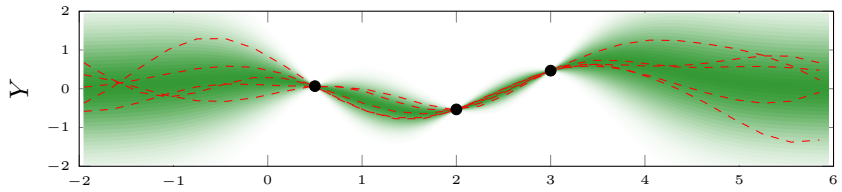
(a) Matérn 3, $\ell = 2$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



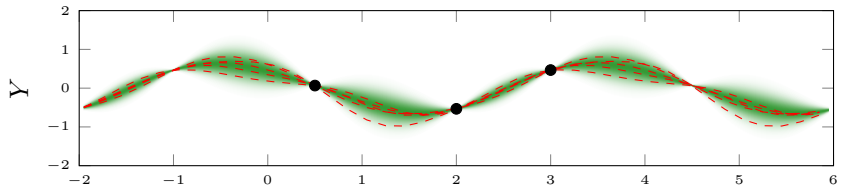
(b) Matérn 3, $\ell = 1/2$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



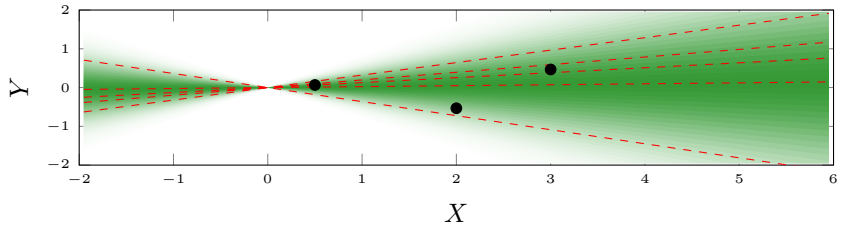
(c) Matérn 5, $\ell = 1$, $\sigma_f^2 = 1$, $\sigma^2 = 0.01$



(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 1.6: 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.

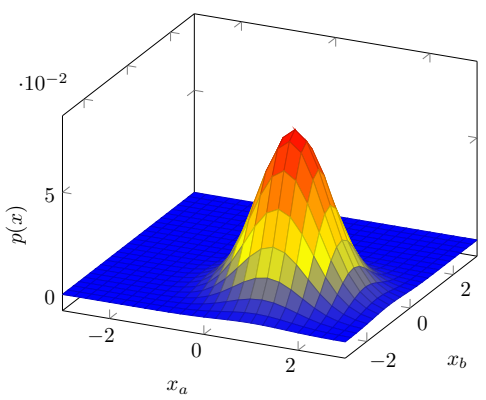
Definition 2 (Multivariate Gaussian)**.** *A random variable $X \in R^p$ with $\mathbb{E}(X) = \mu$ and $\mathrm{Cov}(X) = \Sigma$ such that $\det\Sigma > 0$ is a multivariate Gaussian if and only if the density is*

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

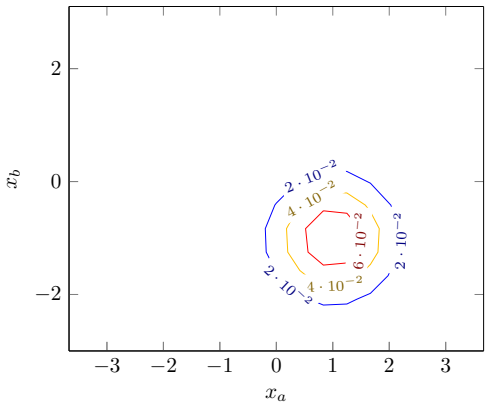
The Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is uniquely determined by its mean vector μ and covariance matrix Σ. For intuition it is helpful to think of the Gaussian distribution as consisting of a normalization constant $Z = 1/(2\pi)^{p/2}\sqrt{\det\Sigma}$ times the exponential of a quadratic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

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

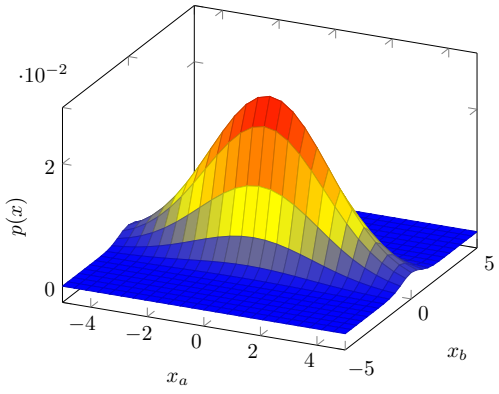
In Figure A.1 we provide a plot of the multivariate Gaussian that was examined in Example 1.1 for particular values of μ and Σ.



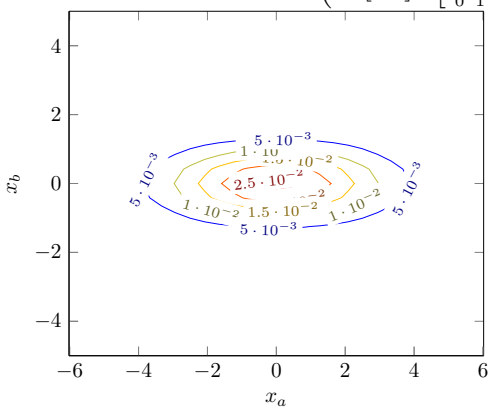
(a) 3D plot of $p(x) = \mathcal{N}\left(x\middle|\ \begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}1^2 & 0\\0 & 1^2\end{bmatrix}\right)$



(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

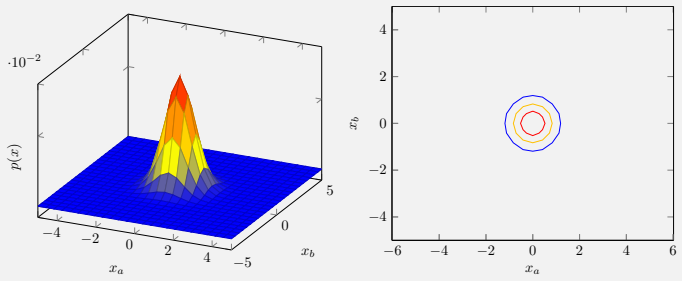
$$\mathbb{E}(Y) = \mathbb{E}(AX+b) = A\mathbb{E}(X) + b = A\mu + b, \tag{A.7a}$$

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

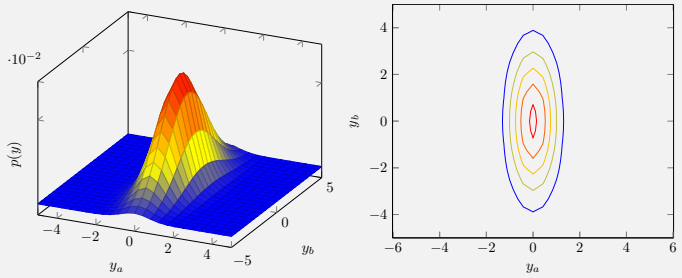
This is illustrated in Example 1.2.

Example 1.2: The geometry of the Gaussian distribution

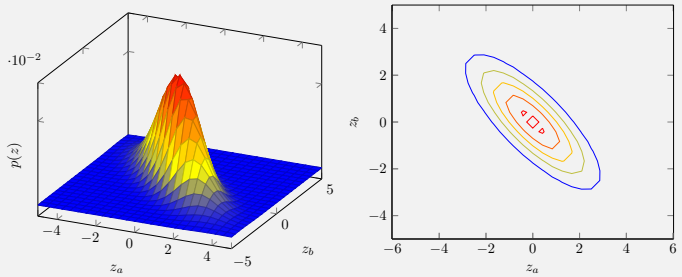
Consider a two-dimensional Gaussian random variable $X \sim \mathcal{N}(x|\ \mu,\ \Sigma)$ where $\mu = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$ and $\Sigma = \left[\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right]$.



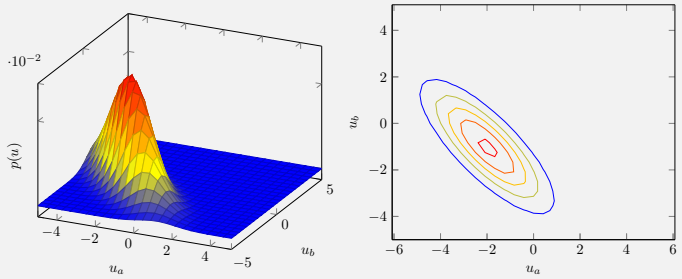
Perform a linear transformation $Y = A_1 X$ where $A_1 = \left[\begin{smallmatrix}1 & 0\\0 & 3\end{smallmatrix}\right]$. The random variable Y will then also be Gaussian distributed with $Y \sim \mathcal{N}\left(y|\ \mu,\ A_1 \Sigma A_1^{\mathsf{T}}\right) = \mathcal{N}\left(y|\ \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right],\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

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

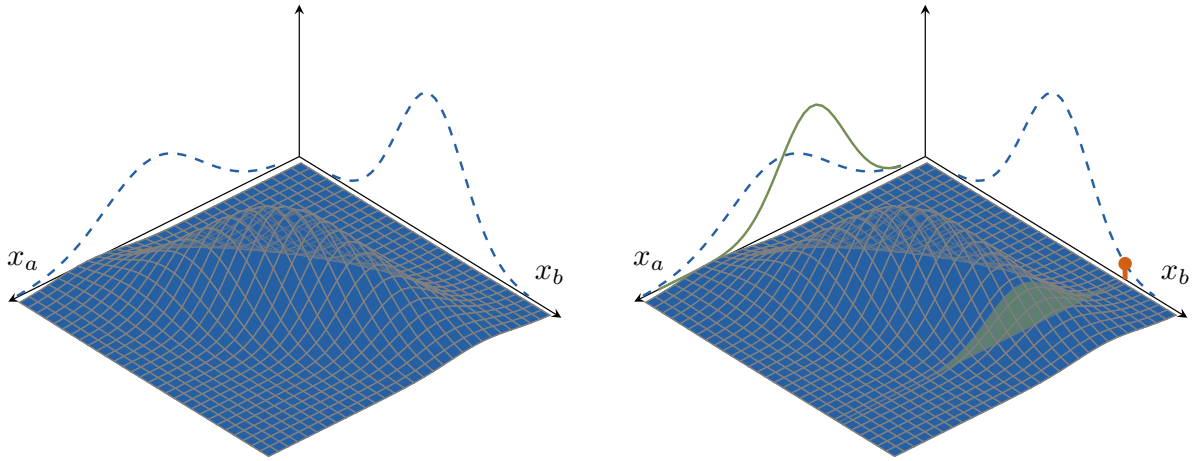
If we measure one variable that in turns depends on another variable, we are often interested in knowing what this measurement can tell us about the unmeasured variable. This is handled using conditioning and for partitioned Gaussian variables the highly useful result is provided in Theorem 2.

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

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

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