Introduction to Gaussian Processes - RAW

Filipe P. Farias
Teleinformatics Engineering Department
Federal University of Ceará
filipepfarias@fisica.ufc.br

Abstract

The Gaussian processes have proven to be a powerfull framework for robust estimation and a flexible model for non-linear *regression*, case which will be the main object of this work, with some implementations of real situations.

1 Introduction

First, we'll overview some initial concepts that will set the background for the GP framework. Let's suppose a data set $\mathcal{D} = \{(x_i, t_i)\}_{i=0}^{N-1}$ which we denote the input as x, the output (or target) as t and N as the number of observations. The first step in our workflow is define a *training* set, i.e. some data that is given to make our first assumptions of the *model*. The model y can be defined as a guess of the law that rules the phenomenon of which our data was observed. This law can be, by example a senoid function as represented in Figure 1. Given this training data we wish to make *predictions* for new inputs x_* that we have not observed in the training set.

We'll assume an parametric approach, then the model is said to contain *parameters* $\mathbf{w}$, that will be adjusted during the *training phase*, when those are modified aiming to reduce the mismatch with the training set. In general we define a *loss function* $L(y, t)$ which increases itself as the mismatch becomes larger, in other words the *error* of the model. Then our work will be to reduce this error such that the smallest one will be the which defines when our model has learned the parameters of the law of the phenomenon. This turns possible to make our predictions where the data was not observed.

Unfortunately, in this trying of obtain the model by the smallest error, we may lose the capability of generalize it, i.e. our model could learned well for the training set only. So, if new data arrive or a new realization of the phenomenon occurs, that smallest error may increase for the same model. With this we define that our model isn't flexible. Then we can increase this flexibility by accepting some *uncertainty* above it. More, sometimes a good first assumption can make the difference to the estimation, and one may want to put its beliefs in the model even before to observe the data, i.e. make a *prior* assumption. These both strategies of uncertainty and prior assumptions are well defined by the Bayesian inference and could help if we assume a *probabilistic model*.

The Bayesian inference can handle with the classical approaches of search the model which has the smallest error, but our objective is achieve one step ahead. We can not only obtain one model, but a *distribution* of possible models. And with this, all the probabilistic meaning of distribution is carried with it, that is we can obtain both the model which *minimize* the error or the statistics of the distribution of models. A more explaining view of what this really means will be given in the next sections.

Furthermore, the concept of *infer* is similar to what we have done since beginning. We maded a guess of the law which rules the phenomenon, i.e. a prior assumption. Then we turns our model more plausible by reducing its error, or more *likely*. After, we obtained a

result of these assumptions, a *posterior* assumption. These steps are similar in concept when dealing with Bayesian inference, except that, as we will deal with probability distributions, then some rules must be established for the method to be concise.

Finally, we'll deal with a specific class of models in which we assume not a distribution of parameters but functions in general. By example, in a space with infinite possible functions, we'll evaluate how much possible which one are to be generated the data by its statistics, what is similar to what was done for the parameters. And in this part we make the fully use of the Gaussian process.

2 Linear Regression

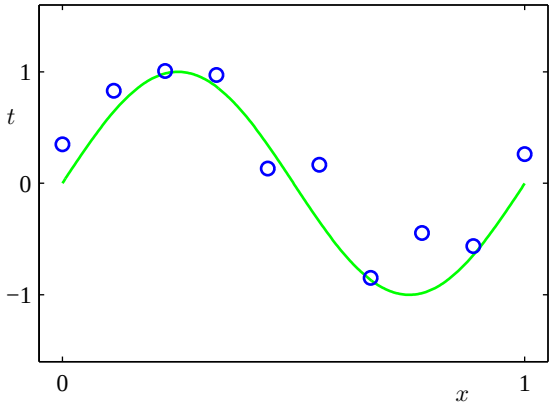

Figure 1: Training data set with $n = 10$ points in blue. The green curve shows the function $sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x, without knowledge of the green curve [Bishop, 2006].

The following scenario is given. The data set of observations is given by $\mathcal{D} = \{(x_i, t_i)\}_{i=0}^{N-1}$, hence we make a guess about the law that rules the phenomenon behind the data. In general, we define that model as a mathematical function y whose parameters $\mathbf{w}$ will be adjusted in trying of learn the phenomenon law. By example, we may choose a polynomial as a model and try to adjust its coefficients, that will be the parameters for the model in this case.

$$y\left(x, \mathbf{w}\right) = w_0 x^0 + \cdots + w_{M-1} x^{M-1} \tag{2.1}$$

This approach is not linear in the inputs, i.e. the model output y it isn't a linear transformation of the inputs. But we can say that is *linear on the parameters*. To visualize this, considering the parameters array $\mathbf{w} = \left(w_0, \cdots, w_{M-1}\right)^\top$, we put in the matrix form as

$$y\left(x, \mathbf{w}\right) = \mathbf{w}^\top \boldsymbol{\phi}(x). \tag{2.2}$$

Here we have defined the array $\boldsymbol{\phi}(x) = \left(x^0, \cdots, x^{M-1}\right)^\top$ in order to keep the linearity of the model. This approach becomes more clear if we take $\boldsymbol{\phi}$ for all the space where we have the observations, by this we mean evaluate $\boldsymbol{\phi}$ for all the $\{x_i\}_{i=0}^{N-1}$ inputs, and we'll obtain

$$\begin{pmatrix} x_0^0 & \cdots & x_0^{M-1} \\ x_1^0 & \cdots & x_1^{M-1} \\ \vdots & \ddots & \vdots \\ x_{N-1}^0 & \cdots & x_{N-1}^{M-1} \end{pmatrix}. \tag{2.3}$$

This matrix is called the *design matrix* Φ. Hence for the vector with all the outputs, we would obtain of $\mathbf{y} = \mathbf{w}^\top \Phi$. These tools make possible gain some insight that help us in the follow steps. Backing to the data, let's take the targets in the vector form, this is $\mathbf{t} = \left(t_0, \cdots, t^{N-1}\right)^\top$, so we can consider that $\mathbf{t}$ is a *vector in space* [Bishop, 2006]. If we take the columns of Φ, denoting each one by $\boldsymbol{\varphi}_j$, what defines the vector $\mathbf{y}$ by its linear combination. If we take a polynomial of M^{th} order, being M smaller than the number of points N, than we say that $\boldsymbol{\varphi}$ will span a linear subspace $\mathcal{S}$ of dimensionality M. With this, our problem turns to actually train the parameters, what can be done defining some metric that say to us when the model is missing, as a *loss function*.

Maybe an intuitive way to say to the model if he's going far away from the target is measure this error, i.e. how much the model is *distant* from the target. We may generalize the distance itself by the *Minkowsky distance* L_p between $\mathbf{u}$ and $\mathbf{v}$, defined as

$$L_p\left(\mathbf{u}, \mathbf{v}\right) = \left(\sum_{i=1}^{n} |u_i - v_i|^p\right)^{1/p} \tag{2.4}$$

where we have the Euclidean distance for $p = 2$ and the also known Manhattan distance for $p = 1$. The question here is define what distance use, by example the Euclidean distance is the most common due to its analytical tractability and practical importance. With this choice the error between the model and the targets will be define as the distance between $\mathbf{y}$ and $\mathbf{t}$, and where the orthogonal projection of $\mathbf{t}$ onto the subspace $\mathcal{S}$ is the minimal distance. The *learning* step of the parameters $\mathbf{w}$ lies on choose those whose make that distance minimal. Hence, being $\langle \cdot, \cdot \rangle$ the inner product, we have

$$L_2(\mathbf{t} - \mathbf{y}) = \langle \mathbf{t} - \mathbf{y}, \mathbf{t} - \mathbf{y} \rangle^{1/2}. \tag{2.5}$$

In general we denote L as a *loss function* and add an $1/2$ term multiplying the inner product, what will be shown next. The minimal distance makes itself clear the process of minimization for looking the best choice of parameters. In this case, the minimization is taken by the derivative of L_2 with $\partial L_2 / \partial \mathbf{w} = 0$, then

$$\mathbf{w}^* = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbf{t} \tag{2.6}$$

for $\mathbf{y} = \mathbf{w}^\top \Phi$. Making use of the optimization notation, we denoted $\mathbf{w}^*$ as the best choice of the parameters. Note that, even working with a polynomial model, we do not manipulate it in fact. This approach enable us to generalize for a several types of *basis functions*, i.e. functions that will define our design matrix.

The linear regression is one of the simplest methods for estimation and bring us an idea to find the law of the phenomenon, but one of its problems is its flexibility and non-robustness. There's a method trying to reduce this *non-flexibility* in the model regularizing the parameters, known as regularization, but we'll be not discussed in this work. By the example of the polynomial model, if we construct the polynomial for the learned parameters, we note that its coefficients will have high values. This is because when learning the targets, if we consider the order of the polynomial the same as the number of data, the model starts to interpolate the noise, i.e. the aspects captured by the model may not be precise with the law of the phenomenon and with this, the derivatives of the polynomial may grow, as shown in the Figure 2. Some questions may arise as how know exactly this law, what in fact we may never know. The next step lead us to a probabilistic point of view with some more tools to deal with the problem.

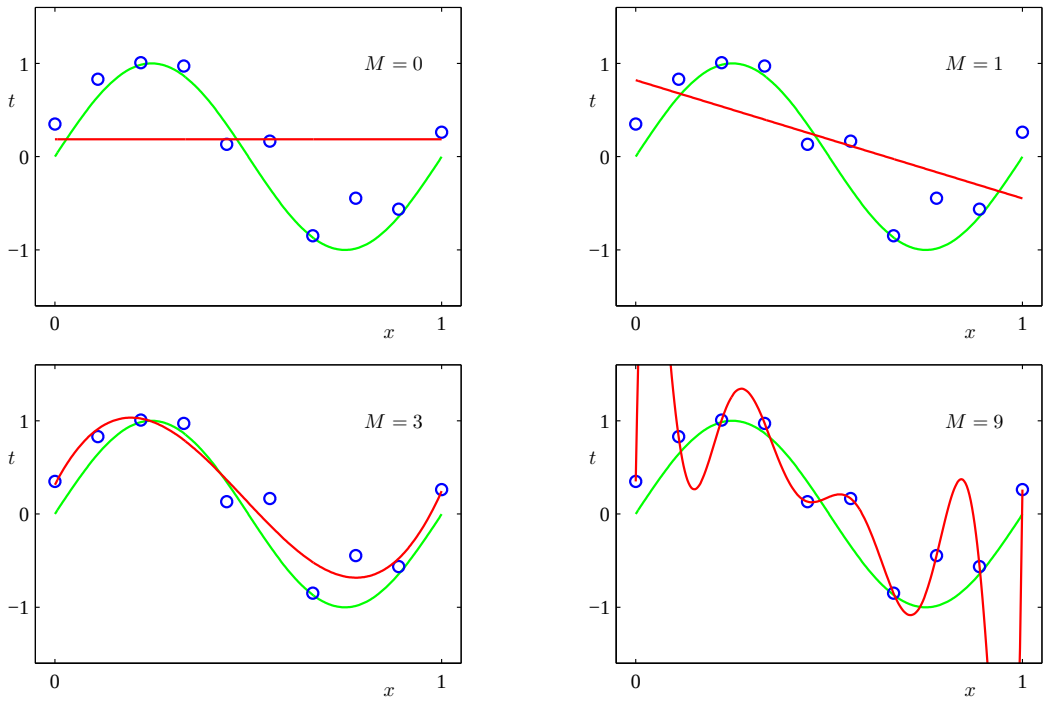

Figure 2: Plots of polynomials for the model in (2.1) having various orders M, shown as red curves [Bishop, 2006].

3 Bayesian Linear Regression

3.1 Bayesian inference

Until now, we see the curve fitting problem in terms of the minimization of the error function. Aiming to apply a "more Bayesian" approach, we not have yet a prior distribution to make the inference using the Bayes' rule.

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{3.1}$$

We can use the Bayes' theorem (3.1) to convert a *prior* probability into a *posterior* probability at the light of some evidence. Here, we represented the distribution probability of $\mathbf{w}$ that could be generated the data $\mathcal{D}$ as $p(\mathbf{w}|\mathcal{D})$. We can make prior assumptions about quantities such as the parameters $\mathbf{w}$ in the form of a prior distribution $p(\mathbf{w})$. The observation of the data $\mathcal{D}$ and what its implies in the parameters is expressed as a conditional probability $p(\mathcal{D}|\mathbf{w})$. Then we can evaluate the uncertainty about $\mathbf{w}$ after observed the data $\mathcal{D}$ as a posterior probability $p(\mathbf{w}|\mathcal{D})$.

For a fixed $\mathcal{D}$, the function $l(\mathbf{w};\mathcal{D}) = p(\mathbf{w}|\mathcal{D})$ gives us the plausibility or *likelihood* of each possible value of $\mathbf{w}$, include Ricardo Ehlers citation. This and the prior distribution are combined leading to the posterior distribution of $\mathbf{w}$. considering $p(\mathcal{D})$ a constant independent of $\mathbf{w}$, the usual for of the Bayes' theorem is

$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) \tag{3.2}$$

The quantity $p(\mathcal{D}|\mathbf{w})$ is called *likelihood function* and expresses how probable the observed data $\mathcal{D}$ is for different settings of $\mathbf{w}$. Then, not being necessarily a probability distribution, but a function over the parameters [DeGroot and Schervish, 2012], its integral with respect

to $\mathbf{w}$ could not be equal one, then to normalize the equation with respect to the left-side there's a term $p(\mathcal{D})$.

Integrating the both sides of (3.2) with respect to $\mathbf{w}$, we obtain the denominator, then considering that integrating a probability distribution over itself is equal to one, we have

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \tag{3.3}$$

This is called *Bayesian inference*.

3.2 Bayesian curve fitting

Let's consider the same data set $\mathcal{D}$ presented before, but now we have some uncertainty over the value of the measured value t. This uncertainty can be represented as a probability distribution function p, in this particular case a Gaussian distribution, with a mean equal to the model $y(x, \mathbf{w})$. Thus we have

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{3.4}$$

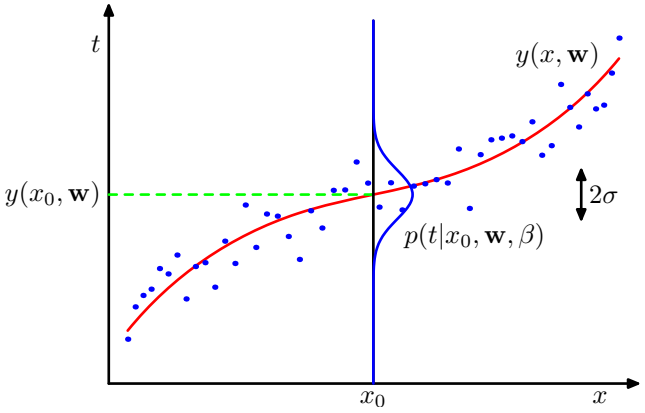

Figure 3: Schematic illustration of a Gaussian conditional distribution for t given x given by (3.4), in which the mean is given by the polynomial function y(x, w), and the precision is given by the parameter β, which is related to the variance $\beta^{-1} = \sigma^2$ [Bishop, 2006].

Where β is the variance of the distribution. Note that a large β will give us more imprecision about the measured value t, then we can call it of *precision parameter*, i.e. how much certain we are about t. As we done in linear regression, we are trying to obtain the parameters for the model. In other words, given a value t, we trying to obtain the *mean* and the *variance* which maximize the probability of the measured value. Assuming the data set being independent and identically distributed, the joint probability of the whole data set will be

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=0}^{N-1} \mathcal{N}\left(t_i|y(x_i, \mathbf{w}), \beta^{-1}\right) \tag{3.5}$$

When viewed as function of $y(x_i, \mathbf{w})$, the model, and β^{-1}, this is the likelihood function for the Gaussian. The parameters of the distribution can be determined by maximizing the likelihood function.

3.3 Predictive distribution

We can now introduce here the probability distribution over the parameters $p(\mathbf{w})$. The choice is arbitrary, but for this particular case we will consider

$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}\right\} \tag{3.6}$$

where α is the variance, or precision parameter, of the distribution and $M+1$ is the number of parameters of the model, i.e. the length of $\mathbf{w}$. We call *hyperparameters* the variables such α who control the model parameters distribution. And now we have by the Bayes' theorem considering that our *posterior* distribution is proportional to the product between the *likelihood function* and the assumed *prior*, as seen in (3.3) then, assuming the observation of the whole data set

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \tag{3.7}$$

The similarity between the maximization and the Bayesian approach mentioned before shows that the second comprise even a model training such as the classical regression, as also the control of the over fitting by the regularization. But to say that our model is in fact Bayesian, we might obtain not just a single value, as in MAP, but its distribution. This requires the application the fully Bayes' theorem as (3.1). Then, we have

$$\overbrace{p(\mathbf{w}|\mathcal{D})}^{posterior} = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} = \frac{\overbrace{p(\mathcal{D}|\mathbf{w})}^{likelihood}\overbrace{p(\mathbf{w})}^{prior}}{\underbrace{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}}_{marginal\ distribution}} \tag{3.8}$$

If we assume that all distributions which we are working are Gaussian, the posterior distribution has closed form. To do that, we make use of the closure under linear transformations, or *affine transformations*, for the Gaussian. For that, we will make use of the corollary below, which theorems are proven in the Appendix B.

Assuming no deviation in the mean, $\mathbf{d} = \mathbf{0}$, and the linear transformation being our design matrix, $\mathbf{M} = \Phi$, we obtain that

$$p(\mathbf{w}|\mathbf{t}, \alpha, \beta) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}}, \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}}\right) \tag{3.9}$$

being

$$\boldsymbol{\mu}_{\mathbf{w}|\mathbf{t}} = \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}}\left(\beta\Phi^\top\mathbf{t} + \boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\mu}_{\mathbf{w}}\right), \quad \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} = \left(\boldsymbol{\Sigma}_{\mathbf{w}}^{-1} + \beta\Phi^\top\Phi\right)^{-1} \tag{3.10}$$

assumed the prior distribution defined in (3.6) and the precision matrix $\boldsymbol{\Sigma}_{\mathbf{t}|\mathbf{w}} = \beta^{-1}\mathbf{I}$. Then we have defined

$$\boldsymbol{\mu}_{\mathbf{w}} = \mathbf{0}, \quad \boldsymbol{\Sigma}_{\mathbf{w}} = \alpha^{-1}\mathbf{I}. \tag{3.11}$$

In practice, some times it is more valuable the information about t itself than its parameters $\mathbf{w}$. We can make this by evaluating the predictions of t for the new values of x by

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\mathrm{d}\mathbf{w} \tag{3.12}$$

what is called *predictive distribution*. The distributions under integration were defined in (3.4) and (3.9). Then we use the *marginalization* defined in Appendix B and obtain that

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right) \tag{3.13}$$

with $\boldsymbol{\mu}_t = \boldsymbol{\phi}(x)^\top \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}}$ and $\boldsymbol{\Sigma}_t = \beta^{-1} + \boldsymbol{\phi}(x)^\top \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{t}} \boldsymbol{\phi}(x)$. Note that we are considering the linear model as $y(x, \mathbf{w}) = \boldsymbol{\phi}(x)^\top \mathbf{w}$, i.e. the affine transformation $\mathbf{M}$ here is $\boldsymbol{\phi}(x)^\top$.

An alternative formulation [Rasmussen and Williams, 2005] is

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\Big(&\boldsymbol{\phi}(x)^\top \Sigma_{\mathbf{w}} \Phi \left(K + \beta I\right)^{-1} \mathbf{t}\ , \\ &\boldsymbol{\phi}(x)^\top \Sigma_{\mathbf{w}} \boldsymbol{\phi}(x) - \boldsymbol{\phi}(x)^\top \Sigma_{\mathbf{w}} \Phi \left(K + \beta I\right)^{-1} \Phi^\top \Sigma_{\mathbf{w}} \boldsymbol{\phi}(x)\Big) \end{aligned} \tag{3.14}$$

with $K = \Phi^\top \boldsymbol{\Sigma}_{\mathbf{w}} \Phi$.

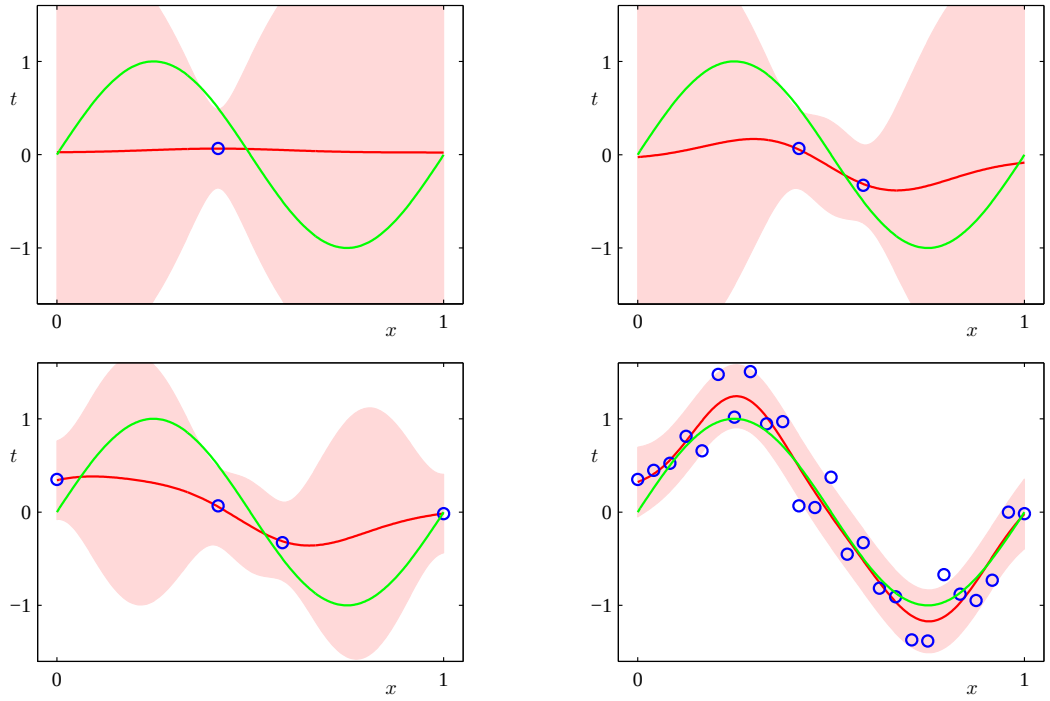

Figure 4: Examples of the predictive distribution for a model consisting of 9 Gaussian basis functions of the form $\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$ [Bishop, 2006].

Thus we need to move from the finite training data $\mathcal{D}$ to a function f that makes predictions for all possible input values.

4 Gaussian processes

Until now we have made the inference in the *feature space*, the space where the parameters $\mathbf{w}$ are. In other words, the strategy was to train our model, obtaining the parameters probability distribution, by Bayesian inference, and then evaluating the predictive distribution with the posterior of the inference. Now, we'll assume a prior distribution of functions. By example, if there's a function $y(\mathbf{x})$, the value of y at each real-valued point $\mathbf{x}$ is a random variable.

With these assumptions, if we take a collection of y in arbitrary points $\mathbf{x}$ and this collection has a Gaussian distribution, we may say that a *Gaussian process* occurs [Rasmussen and

Williams, 2005]. Finally, our Bayesian inference can be done in a similar way to find the posterior distribution of functions.

We define a Gaussian process (GP) as a collection of random variables, such that any finite number of which is normal jointly distributed. In other words, it can be thought of as a generalization of a Gaussian distribution over a finite vector space to a function space of infinite dimension [MacKay, 1998]. As the normal distribution, the GP is completely defined by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ of a real process $y(\mathbf{x})$, these in turn are defined as

$$m(\mathbf{x}) = \mathbb{E}\left\{y(\mathbf{x})\right\}, \quad k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left\{\left(y(\mathbf{x}) - m(\mathbf{x})\right)\left(y(\mathbf{x}') - m(\mathbf{x}')\right)\right\} \tag{4.1}$$

and finally

$$y(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right) \tag{4.2}$$

Such collection definition automatically implies in the marginalization property already present for the multidimensional Gaussian distributions. For the GP this means that the observation of a larger set of variables does not change the distribution of the smaller set.

This is important to obtain our Bayesian linear regression as a GP. Being our model $y(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w}$ with prior $\mathcal{N}\left(\mathbf{w}|\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{w}\right)$. We obtain for the mean and covariance functions

$$\begin{aligned} \mathbb{E}\left\{y(\mathbf{x})\right\} &= \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}\left\{\mathbf{w}\right\} = 0, \\ \mathbb{E}\left\{y(\mathbf{x})y(\mathbf{x}')\right\} &= \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}\left\{\mathbf{w}\mathbf{w}^\top\right\} \boldsymbol{\phi}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_\mathbf{w} \boldsymbol{\phi}(\mathbf{x}') \end{aligned} \tag{4.3}$$

4.1 Prediction with Noisy Observations

The last sections give us an insight about the construction of the learning process for the GP. Analogous to the Bayes' theorem for the Gaussian, we use the idea of the GP as a multidimensional Gaussian distribution and we can make inference with its partitions.

First, to take the concept, we will consider the case where the observations are noise free. We will substitute the covariance matrices from the partitioned Gaussian distributions by the covariance function applied at the points of the observations $\mathbf{t}$ and the prediction $\mathbf{t}_*$, as

$$\begin{pmatrix} \mathbf{t} \\ \mathbf{t}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{x}, \mathbf{x}) & \mathbf{K}(\mathbf{x}, \mathbf{x}_*) \\ \mathbf{K}(\mathbf{x}_*, \mathbf{x}) & \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right) \tag{4.4}$$

where $\mathbf{K}(\cdot, \cdot)$ denotes the covariance matrices evaluated at all pairs of N-dimensional $\mathbf{x}$ training points and N_*-dimensional $\mathbf{x}_*$ test points. Then, making use of Appendix B, we use the *conditioning* to obtain the predictive distribution

$$\begin{aligned} p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{x}, \mathbf{y}) = \mathcal{N}\big(&\mathbf{y}_*|\mathbf{K}(\mathbf{x}_*, \mathbf{x})\mathbf{K}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{y}, \\ &\mathbf{K}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{K}(\mathbf{x}_*, \mathbf{x})\mathbf{K}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{K}(\mathbf{x}, \mathbf{x}_*)\big) \end{aligned} \tag{4.5}$$

Now assuming the noise in the observations, what is a more realistic modelling situation, we have that $t = y(\mathbf{x}) + \varepsilon$, being ε the Gaussian noise with variance β^{-1}. Then we have that $\mathrm{cov}(\mathbf{t}) = \mathbf{K}(\mathbf{x}, \mathbf{x}) + \beta^{-1}\mathbf{I}$.

Deriving the conditional distribution corresponding we arrive at the key predictive equations for Gaussian process regression

$$\begin{aligned} p(\mathbf{y}_*|\mathbf{x}, \mathbf{t}, \mathbf{x}_*) &= \mathcal{N}\left(\overline{\mathbf{y}}_*, \mathrm{cov}\left(\mathbf{y}_*\right)\right), \text{ where} \\ \overline{\mathbf{y}}_* &\triangleq \mathbb{E}\left\{\mathbf{y}_*|\mathbf{x}, \mathbf{t}, \mathbf{x}_*\right\} = \mathbf{K}\left(\mathbf{x}_*, \mathbf{x}\right)\left(\mathbf{K}(\mathbf{x}, \mathbf{x}) + \beta^{-1}\mathbf{I}\right)^{-1}\mathbf{t} \\ \mathrm{cov}\left(\mathbf{y}_*\right) &= \mathbf{K}\left(\mathbf{x}_*, \mathbf{x}_*\right) - \mathbf{K}\left(\mathbf{x}_*, \mathbf{x}\right)\left(\mathbf{K}(\mathbf{x}, \mathbf{x}) + \beta^{-1}\mathbf{I}\right)^{-1}\mathbf{K}\left(\mathbf{x}, \mathbf{x}_*\right) \end{aligned} \tag{4.6}$$

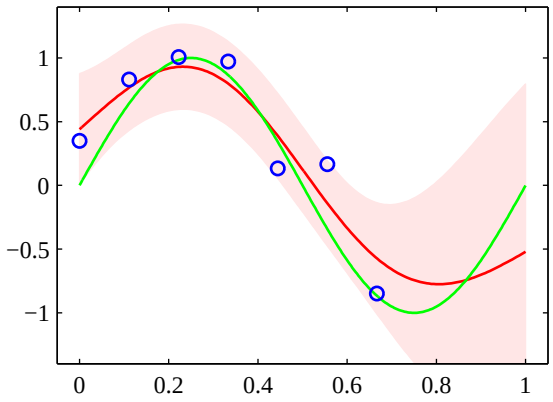

Figure 5: Illustration of Gaussian process regression applied to the sinusoidal data set in which the three right-most data points have been omitted. The green curve shows the sinusoidal function from which the data points, shown in blue, are obtained by sampling and addition of Gaussian noise. The red line shows the mean of the Gaussian process predictive distribution, and the shaded region corresponds to plus and minus two standard deviations. Notice how the uncertainty increases in the region to the right of the data points [Bishop, 2006].

5 Applications in disease mapping

5.1 The model

There's several applications using GP and here we'll resume an example for disease mapping presented by [Vanhatalo et al., 2010]. Then let's assume that our phenomenon is ruled by an function f. But, we interested in the distribution of them, considering the approach presented in this work. So, we may say that we evaluated each observation y_i from an unknown function f_i. With this we assume that our observations and our functions are independent and then we can evaluate our joint distribution for the likehood by the product of each one [Vanhatalo, 2010].

$$\begin{cases} y_1, y_2, \ldots, y_n \sim \prod_{i=1}^{n} Poisson\left(e_i \exp(f_i)\right) & \text{(5.1a)} \\ f(\mathbf{x})|\theta \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'|\theta)\right) & \text{(5.1b)} \\ \theta \sim \text{half-t}(\nu, A)^* & \text{(5.1c)} \end{cases}$$

In this case, we used the Poisson distribution for the likelihood because the nature of the process. The phenomenon here is the relative risk of death μ in a region of the country. So, if we consider y the counting of deaths on this region, we can model the phenomenon with a Poisson process which mean in each region is given by the increasing rate of deaths. At this point we have defined e as the standardized expected number of deaths [Vanhatalo et al., 2010], what multiplied by μ reveals, in mean, the rate of deaths in that region. For numerical reasons, we transform $f = \log(\mu)$. Finally, we assume an uncertainty over the parameters of the kernel functions too, then, our hierarchical model stays for the posterior distribution as

$$p(\mathbf{f}|\mathbf{y}, \mathbf{x}) \propto \int p(\mathbf{y}|\mathbf{f}) \mathcal{GP}\left(\mathbf{f}|m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'|\theta)\right) p(\theta) d\theta. \tag{5.2}$$

This function isn't analytically tractable because of the Poisson process, but it is possible its evaluation with approximation methods.

*The values ν and A are not arbitrary, but deterministic [Vanhatalo et al., 2010].

Appendix

A Derivations

A.1 Matrix Form

Be the linear model $y(x, \mathbf{w}) = \boldsymbol{\phi}(x)^\top \mathbf{w}$. Suppose $\Phi = [\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_N)]^\top$, then Φ will be of the form

$$\Phi = \begin{bmatrix} \phi_0(x_0) & \ldots & \phi_{M-1}(x_0) \\ \vdots & \ddots & \vdots \\ \phi_0(x_{N-1}) & \ldots & \phi_{M-1}(x_{N-1}) \end{bmatrix} \tag{A.1}$$

called *design matrix*. Then the model turns to $\mathbf{y} = \Phi\mathbf{w}$. This will lead us to the matrix form for the quadratic error function

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}(\mathbf{y} - \mathbf{t})^\top(\mathbf{y} - \mathbf{t}) \\ &= \frac{1}{2}(\Phi\mathbf{w} - \mathbf{t})^\top(\Phi\mathbf{w} - \mathbf{t}) \\ &= \frac{1}{2}(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - \mathbf{t}^\top\Phi\mathbf{w} - \mathbf{w}^\top\Phi^\top\mathbf{t} + \mathbf{t}^\top\mathbf{t}) \end{aligned}$$

Observe that even in the matrix form, the error function remains scalar, which implies that $\mathbf{t}^\top\Phi\mathbf{w} = \mathbf{w}^\top\Phi^\top\mathbf{t}$ by the transpose of the product rule. Then

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - 2\mathbf{t}^\top\Phi\mathbf{w} + \mathbf{t}^\top\mathbf{t})$$

Then we proceed by the minimization by $\frac{\partial E}{\partial \mathbf{w}} = 0$

$$\begin{aligned} 0 &= \frac{1}{2}(2\mathbf{w}^\top\Phi^\top\Phi - 2\mathbf{t}^\top\Phi)^\dagger \\ \mathbf{w}^{*\top} &= \mathbf{t}^\top\Phi(\Phi^\top\Phi)^{-1} \\ \mathbf{w}^* &= (\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{t} \end{aligned} \tag{A.2}$$

For the regularized linear regression, we do $\frac{\lambda}{2}||\mathbf{w}||^2 = \mathbf{w}^\top\mathbf{w}$, then

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}(\mathbf{y} - \mathbf{t})^\top(\mathbf{y} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \\ &= \frac{1}{2}(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - \mathbf{t}^\top\Phi\mathbf{w} - \mathbf{w}^\top\Phi^\top\mathbf{t} + \mathbf{t}^\top\mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \end{aligned}$$

[†]Using two facts. First, if $\alpha = \mathbf{x}^\mathrm{T}\mathbf{A}\mathbf{x}$, then $\frac{\partial\alpha}{\partial\mathbf{x}} = 2\mathbf{x}^\top\mathbf{A}$, being α scalar. Second, if $\alpha = \mathbf{t}^\mathrm{T}\mathbf{A}\mathbf{x}$, then $\frac{\partial\alpha}{\partial\mathbf{x}} = \mathbf{t}^\mathrm{T}\mathbf{A}$. For both, $\mathbf{A}$ is independent of $\mathbf{x}$ and $\mathbf{t}$ [Graybill, 1983].

And with the minimization we do $\frac{\partial E}{\partial \mathbf{w}} = 0$, then

$$\begin{aligned} 0 &= \mathbf{w}^\top \Phi^\top \Phi - \mathbf{t}^\top \Phi + \lambda \mathbf{w}^\top \\ \mathbf{w}^{*\top} &= \mathbf{t}^\top \Phi (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \\ \mathbf{w}^* &= (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{t} \end{aligned} \tag{A.3}$$

where $\mathbf{I}$ is the identity matrix.

B Bayes' theorem for Gaussian variables [Schön and Lindsten, 2011]

B.1 Partitioned Gaussian distributions

Be $\mathbf{x}$ a n-dimensional vector with a Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the partitioned will be

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}. \tag{B.1}$$

Preserved the symmetry $\boldsymbol{\Sigma}^\top = \boldsymbol{\Sigma}$, we say the covariance matrix is positive definite. And be the multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\det \boldsymbol{\Sigma})^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{B.2}$$

We define too, just for convenience of work, the precision matrix $\boldsymbol{\Lambda}$ by

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \equiv \boldsymbol{\Sigma}^{-1} \tag{B.3}$$

assuming all matrices have inverses.

Theorem 1 (Marginalization)**.** *Being the random vector* $\mathbf{x}$ *and its partitioned as above, the marginal density* $p(\mathbf{x}_a)$ *is given by*

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

Proof. The marginal density $p(\mathbf{x}_a)$ is obtained by integrating the joint density $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$ with relation to $\mathbf{x}_b$

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) \, \mathrm{d}\mathbf{x}_b \tag{B.4}$$

Then we expand the exponential argument of (B.2) for the partitioned Gaussian

$$-\frac{1}{2} \left(\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \right)^\top \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \left(\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \right) \tag{B.5}$$

What implies

$$
\begin{aligned}
&-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\boldsymbol{\Lambda}_{aa}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)\\
&-\frac{1}{2}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)^{\top}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)-\frac{1}{2}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)^{\top}\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)
\end{aligned}
\tag{B.6}
$$

Here we make use of the *Schur complement*, in which, being the partitioned matrix

$$
\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}\mathbf{M}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{M}^{-1}\\ -\mathbf{M}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{M}^{-1}\end{pmatrix}
\tag{B.7}
$$

the quantity $\mathbf{M}^{-1}$ is the *Schur complement* of the left side matrix with respect to $\mathbf{D}$, defined as

$$
\mathbf{M}=\left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}
\tag{B.8}
$$

This will motivated the term grouping below

$$
\begin{aligned}
&-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\boldsymbol{\Lambda}_{aa}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)\\
&-\frac{1}{2}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)^{\top}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)-\frac{1}{2}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)^{\top}\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)\\
&=-\frac{1}{2}\left(\mathbf{x}_b^{\top}\boldsymbol{\Lambda}_{bb}\mathbf{x}_b-2\mathbf{x}_b^{\top}\boldsymbol{\Lambda}_{bb}\left(\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)\right)-2\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b\right.\\
&\left.+2\boldsymbol{\mu}_a^{\top}\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b+\boldsymbol{\mu}_b^{\top}\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b+\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a-2\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a+\boldsymbol{\mu}_a^{\top}\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a\right)
\end{aligned}
$$

Then we complete the squares resulting independent

$$
\begin{aligned}
&-\frac{1}{2}\left(\mathbf{x}_b-\left(\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)\right)\right)^{\top}\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b-\left(\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)\right)\right)\\
&+\frac{1}{2}\left(\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a-2\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a+\boldsymbol{\mu}_a\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a\right)\\
&-\frac{1}{2}\left(\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a-2\mathbf{x}_a^{\top}\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a+\boldsymbol{\mu}_a^{\top}\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a\right)\\
&=\underbrace{-\frac{1}{2}\left(\mathbf{x}_b-\left(\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)\right)\right)^{\top}\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b-\left(\boldsymbol{\mu}_b-\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)\right)\right)}_{E_1}\\
&\underbrace{-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\left(\boldsymbol{\Lambda}_{aa}-\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)}_{E_2}
\end{aligned}
$$

Now, back to the marginalization, we have

$$
p\left(\mathbf{x}_a\right)=\int\frac{1}{\left(2\pi\right)^{n/2}}\frac{1}{\left(\det\boldsymbol{\Sigma}\right)^{1/2}}\exp\left\{E_1\right\}\exp\left\{E_2\right\}d\mathbf{x}_b
\tag{B.9}
$$

By inspection, we have that, being the integral of the density function equals one and being n_b the dimension of $\mathbf{x}_b$

$$
\int\exp\left\{E_1\right\}d\mathbf{x}_b=\left(2\pi\right)^{n_b/2}\left(\det\boldsymbol{\Lambda}_{bb}^{-1}\right)^{1/2}
\tag{B.10}
$$

Then, substituting (B.9) in (B.10) we have

$$p\left(\mathbf{x}_a\right) = \frac{1}{(2\pi)^{n_a/2}} \frac{\left(\det \boldsymbol{\Lambda}_{bb}^{-1}\right)^{1/2}}{\left(\det \boldsymbol{\Sigma}\right)^{1/2}} \exp\left\{E_2\right\}$$

We will use the determinant property from Schur complement below

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det \mathbf{A} \det\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right) \tag{B.11}$$

Then, substituting the values of $\boldsymbol{\Sigma}$ we have that

$$\det \boldsymbol{\Sigma} = \det \boldsymbol{\Sigma}_{aa} \det\left(\boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\right) \tag{B.12}$$

Using the Schur complement

$$\det \boldsymbol{\Sigma} = \det \boldsymbol{\Sigma}_{aa} \det \boldsymbol{\Lambda}_{bb}^{-1} \tag{B.13}$$

Finally, using the Schur complement again, we obtain that $\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba} = \boldsymbol{\Sigma}_{aa}$, what concludes the proof, substituting the result in E_2. □

Theorem 2 (Conditioning)**.** *Being the random vector* $\mathbf{x}$ *and its partitioned as above, the conditional density* $p(\mathbf{x}_a|\mathbf{x}_b)$ *is given by*

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right)$$

where $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)$ *and* $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$.

Proof. By the product rule we have

$$p(\mathbf{x}_a|\mathbf{x}_b) = \frac{p(\mathbf{x})}{p(\mathbf{x}_b)} \tag{B.14}$$

what is by (B.2),then

$$p\left(\mathbf{x}_a|\mathbf{x}_b\right) = \sqrt{\frac{\det \boldsymbol{\Sigma}_{bb}}{(2\pi)^{n_a/2} \det \boldsymbol{\Sigma}}} \exp(E) \tag{B.15}$$

where

$$E = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)^\top \boldsymbol{\Sigma}_{bb}^{-1}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \tag{B.16}$$

Similarly to what was done in *marginalization* and using the Schur complement, the result

$$\det \boldsymbol{\Sigma} = \det \boldsymbol{\Sigma}_{bb} \det\left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right) = \det \boldsymbol{\Sigma}_{bb} \det \boldsymbol{\Lambda}_{bb}^{-1} \tag{B.17}$$

Using that $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$, we expand the term E

$$E = -\frac{1}{2}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)^\top \boldsymbol{\Lambda}_{aa}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right) - \frac{1}{2}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)^\top \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \tag{B.18}$$

$$-\frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)^\top \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right) - \frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)^\top \left(\boldsymbol{\Lambda}_{bb} - \boldsymbol{\Sigma}_{bb}^{-1}\right)\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \tag{B.19}$$

Reordering the terms, we'll have

$$
\begin{aligned}
E = &-\frac{1}{2}\mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\left(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right) \\
&- \frac{1}{2}\boldsymbol{\mu}_a^\top \boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\mu}_a^\top \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) - \frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)^\top \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)
\end{aligned} \tag{B.20}
$$

Completing the squares, we obtain

$$
E = \left(\mathbf{x}_a - \left(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right)\right)^\top \boldsymbol{\Lambda}_{aa}\left(\mathbf{x}_a - \left(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right)\right) \tag{B.21}
$$

Then by inspection we have for the precision matrix $\boldsymbol{\Lambda}_{aa}$

$$
\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \tag{B.22a}
$$

and analogously

$$
\boldsymbol{\mu}_{b|a} = \boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right) \tag{B.22b}
$$

for $\boldsymbol{\Lambda}_{bb}$. □

B.2 Affine transformation

Making use of the Theorems 1 and 2, we can show that the Gaussian is closed under linear transformations, i.e. an affine transformation of Gaussians results in another Gaussian.

Theorem 1 (Affine transformation)**.** *Assume* $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *are Gaussian distributed and* $\mathbf{x}_b$ *conditioned by* $\mathbf{x}_a$*, as*

$$
p\left(\mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\right), \quad p\left(\mathbf{x}_b|\mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_b|\mathbf{M}\mathbf{x}_a + \mathbf{d}, \boldsymbol{\Sigma}_{b|a}\right) \tag{B.23}
$$

where $\mathbf{M}$ *is a constant matrix and* $\mathbf{d}$ *a constant vector, both with the appropriate dimensions. Then the joint distribution is given by*

$$
p\left(\mathbf{x}_a, \mathbf{x}_b\right) = \mathcal{N}\left(\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \middle| \begin{pmatrix} \boldsymbol{\mu}_a \\ \mathbf{M}\boldsymbol{\mu}_a + \mathbf{d} \end{pmatrix}, \mathbf{R}\right), \tag{B.24}
$$

being

$$
\mathbf{R} = \begin{pmatrix} \mathbf{M}^\top\boldsymbol{\Sigma}_{b|a}^{-1}\mathbf{M} + \boldsymbol{\Sigma}_a^{-1} & -\mathbf{M}^\top\boldsymbol{\Sigma}_{b|a}^{-1} \\ -\boldsymbol{\Sigma}_{b|a}^{-1}\mathbf{M} & \boldsymbol{\Sigma}_{b|a}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_a\mathbf{M}^\top \\ \mathbf{M}\boldsymbol{\Sigma}_a & \boldsymbol{\Sigma}_{b|a} + \mathbf{M}\boldsymbol{\Sigma}_a\mathbf{M}^\top \end{pmatrix}. \tag{B.25}
$$

Proof. Being the vector

$$
\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \tag{B.26}
$$

and its joint distribution, from the Theorems 1 and 2,

$$
p(\mathbf{x}) = p(\mathbf{x}_b|\mathbf{x}_a)p(\mathbf{x}_a) = \frac{(2\pi)^{-(n_a+n_b)/2}}{\sqrt{\det\boldsymbol{\Sigma}_{b|a}\det\boldsymbol{\Sigma}_a}}\exp\left\{-\frac{1}{2}E\right\} \tag{B.27}
$$

where

$$
E = \left(\mathbf{x}_b - \mathbf{M}\mathbf{x}_a - \mathbf{d}\right)^\top \boldsymbol{\Sigma}_{b|a}^{-1}\left(\mathbf{x}_b - \mathbf{M}\mathbf{x}_a - \mathbf{d}\right) + \left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)^\top \boldsymbol{\Sigma}_a^{-1}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right). \tag{B.28}
$$

Rewriting $\mathbf{x}_b - \mathbf{M}\mathbf{x}_a - \mathbf{d}$ as f and $\mathbf{x}_a - \boldsymbol{\mu}_a$ as e, we have

$$\begin{aligned}
E &= (f - \mathbf{M}e)^\top \boldsymbol{\Sigma}_{b|a}^{-1} (f - \mathbf{M}e) + (e)^\top \boldsymbol{\Sigma}_a^{-1} (e) \\
&= e^\top \left(\mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} + \boldsymbol{\Sigma}_a^{-1} \right) e - e^\top \mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} f - f \boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} e + f^\top \boldsymbol{\Sigma}_{b|a}^{-1} f \\
&= \begin{pmatrix} e \\ f \end{pmatrix}^\top \begin{pmatrix} \mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} + \boldsymbol{\Sigma}_a^{-1} & -\mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} \\ -\boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} & \boldsymbol{\Sigma}_{b|a}^{-1} \end{pmatrix} \begin{pmatrix} e \\ f \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \mathbf{M}\boldsymbol{\mu}_a - \mathbf{d} \end{pmatrix}^\top \mathbf{R}^{-1} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \mathbf{M}\boldsymbol{\mu}_a - \mathbf{d} \end{pmatrix}
\end{aligned} \tag{B.29}$$

By (B.11) we have that

$$\begin{aligned}
\det \mathbf{R}^{-1} &= \det \left(\boldsymbol{\Sigma}_{b|a}^{-1} \right) \det \left(\mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} + \boldsymbol{\Sigma}_a^{-1} - \mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} \boldsymbol{\Sigma}_{b|a} \boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} \right) \\
&= \det \left(\boldsymbol{\Sigma}_{b|a}^{-1} \right) \det \left(\boldsymbol{\Sigma}_a^{-1} \right) \\
&= \frac{1}{\det \left(\boldsymbol{\Sigma}_{b|a} \right) \det \left(\boldsymbol{\Sigma}_a \right)}
\end{aligned} \tag{B.30}$$

Finally by inspection of (B.29), we take the mean of $\mathbf{x}_b$ as $\mathbf{M}\boldsymbol{\mu}_a + \mathbf{d}$ which concludes the proof. □

With the results of Theorems 1, 2 and 3 we get the following corollary

Corollary 1. *Being $\mathbf{x}_b$ conditioned on $\mathbf{x}_a$ and Gaussian distributed as*

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \quad p(\mathbf{x}_b | \mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_b | \mathbf{M}\mathbf{x}_a + \mathbf{d}, \boldsymbol{\Sigma}_{b|a}\right) \tag{B.31}$$

$\mathbf{M}$ a constant matrix and $\mathbf{d}$ a constant vector, both with the appropriate dimensions. Then conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is given by

$$p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a | \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right) \tag{B.32a}$$

with

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b} \left(\mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} (\mathbf{x}_b - \mathbf{d}) + \boldsymbol{\Sigma}_a^{-1} \boldsymbol{\mu}_a \right) \tag{B.32b}$$

$$\boldsymbol{\Sigma}_{a|b} = \left(\boldsymbol{\Sigma}_a^{-1} + \mathbf{M}^\top \boldsymbol{\Sigma}_{b|a}^{-1} \mathbf{M} \right)^{-1}. \tag{B.32c}$$

The marginal density of $\mathbf{x}_b$ is given by

$$p(\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_b | \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) \tag{B.33a}$$

with

$$\boldsymbol{\mu}_b = \mathbf{M}\boldsymbol{\mu}_a + \mathbf{d}, \quad \boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_{b|a} + \mathbf{M}\boldsymbol{\Sigma}_a\mathbf{M}^\top. \tag{B.33b}$$

References

[Bishop, 2006] Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

[DeGroot and Schervish, 2012] DeGroot, M. and Schervish, M. (2012). Probability and Statistics. Addison-Wesley.

[Graybill, 1983] Graybill, F. (1983). Matrices with applications in statistics. Wadsworth statistics - probability series. Wadsworth International Group.

[Hennig, 2013] Hennig, P. (2013). Gaussian processes. Machine Learning Summer School 2013.

[MacKay, 1998] MacKay, D. J. (1998). Introduction to gaussian processes. NATO ASI Series F Computer and Systems Sciences, 168:133–166.

[Rasmussen and Williams, 2005] Rasmussen, C. E. and Williams, C. K. I. (2005). Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

[Schön and Lindsten, 2011] Schön, T. B. and Lindsten, F. (2011). Manipulating the multivariate gaussian density. Technical report, Division of Automatic Control, Linköping University, Sweden.

[Vanhatalo, 2010] Vanhatalo, J. (2010). Speeding up the inference in Gaussian process models. PhD thesis, Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Department of Biomedical Engineering and Computational Science, Aalto.

[Vanhatalo et al., 2010] Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse gaussian processes. Statistics in medicine, 29:1580–607.