Introduction to Gaussian processes

Note to course TMA4265 Stochastic modeling, NTNU, 2017

Jo Eidsvik
Department of Mathematical Sciences, NTNU, Norway,
(email: jo.eidsvik@ntnu.no)

1 Introduction

The Gaussian distribution is tremendously popular because of its theoretical properties and the attractive computational features in multivariable settings. In the following we first present background material on the multivariate Gaussian distribution, and next apply these to describe stationary Gaussian processes and Brownian motion in the time domain.

There are numerous textbooks covering Brownian motion and continuous time and state models from the mathematical point of view, see e.g. Øksendal (2003). Others cover the more practical modeling aspects of Gaussian processes, see e.g. Rasmussen and Williams (2006). There are also lots of online web resources such as GPstuff (Vanhatalo et al., 2013) (http://research.cs.aalto.fi/pml/software/gpstuff/).

2 Background on the Gaussian distribution

This section provides some definitions and properties of the Gaussian distribution. Consult e.g. Johnson et al. (2014) for more detailed statistical discussions. Proofs of properties would rely on transformation of variables or the use of moment generating functions.

2.1 Univariate case

For a random variable x, with mean $E(x) = \mu$ and variance $\text{Var}(x) = \sigma^2$, the univariate Gaussian probability density function (pdf) is defined by

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

For short, this is often denoted $p(x) = N(\mu, \sigma^2)$.

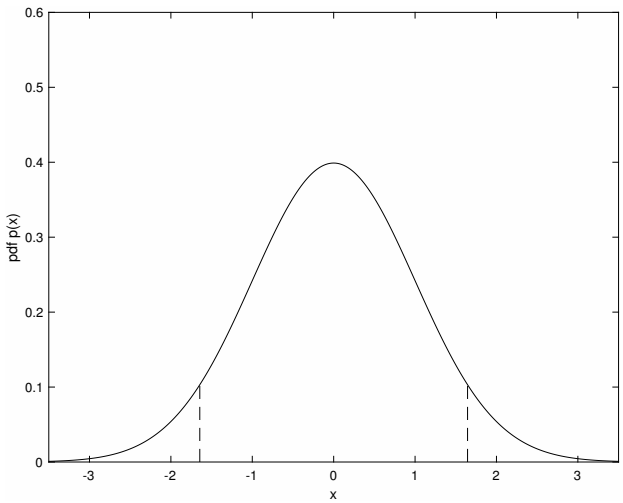

Figure 1: Illustration of a univariate Gaussian pdf with mean $\mu = 0$ and variance $\sigma^2 = 1^2$. The vertical dashed lines at ± 1.64 indicate the 0.9 centered prediction interval for the random variable x.

By a transformation of variable $z = (x - \mu)/\sigma$, with inverse $x = \mu + \sigma z$, and derivative (Jacobian) $J_z = \left|\frac{dx}{dz}\right| = \sigma$, we get the standard Gaussian pdf with zero-mean and unit-variance:

$$p(z) = |J_z| p(x(z)) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \tag{2}$$

Figure 1 illustrates this standard Gaussian pdf.

2.2 Definition of the multivariate case

The multivariate Gaussian pdf for a random variable $\boldsymbol{x} = (x_1, \ldots, x_n)$, viewed as an $n \times 1$ vector, with model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{3}$$

This multivariate Gaussian pdf is a direct extension of the univariate pdf in equation (1), and it is recognized by the quadratic form in the exponent. For short, this pdf is often denoted $p(\boldsymbol{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

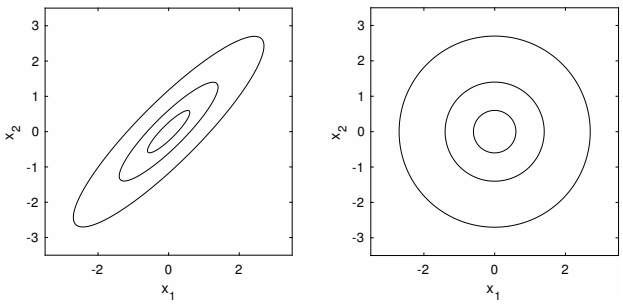

Figure 2: Contour plots illustration of Gaussian pdfs. In both displays the means are 0 and the variances 1. Left: Correlation is 0.9. Right: Independent variables.

The size $n \times 1$ mean vector is $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, $E(x_i) = \mu_i$, and the covariance matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & \ldots & \Sigma_{1,n} \\ \ldots & \ldots & \ldots \\ \Sigma_{n,1} & \ldots & \Sigma_{n,n} \end{bmatrix}, \tag{4}$$

where $\Sigma_{i,i} = \sigma_i^2 = \mathrm{Var}(x_i)$, $i = 1, \ldots, n$. This parameterization of the Gaussian pdf, with the particular quadratic form, thus defines the **marginal** distributions directly; $p(x_i) = N(\mu_i, \sigma_i^2)$. Off-diagonal entries $\Sigma_{i,j} = \mathrm{Cov}(x_i, x_j)$, and the **correlation** between x_i and x_j is $\mathrm{Corr}(x_i, x_j) = \Sigma_{i,j}/(\sigma_i \sigma_j)$.

In the simplest multivariate case there are two random variables x_1 and x_2. If they are **independent**, the covariance matrix in equation (4) is diagonal. Then $\boldsymbol{\Sigma}^{-1}$ is also diagonal, and this means that the quadratic form has no cross-terms in this case. The joint pdf then simplifies to a product of the two marginal distributions; $p(\boldsymbol{x}) = p(x_1)p(x_2)$. If, in addition, the mean and variance terms of x_1 and x_2 are the same, the two variables are said to be independent and **identically distributed** (i.i.d.). The joint distribution of $\boldsymbol{x} = (x_1, x_2)$ is then

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2}\frac{(x_1 - \mu)^2}{\sigma^2} - \frac{1}{2}\frac{(x_2 - \mu)^2}{\sigma^2} \right) = p(x_1)p(x_2), \tag{5}$$

where μ and σ^2 are the common mean and variance of the two variables. For processes, the main interest is in modeling dependent variables.

Figure 2 illustrates two bivariate Gaussian pdfs in a contour plot. In both displays the means are zero and the variables have unit variance. In

the left display the two variables are dependent with $\text{Corr}(x_1, x_2) = 0.9$. This positive correlation means that the two variables have a tendency of being jointly larger than the means, or jointly smaller. In the right display the variables are independent. In such contour plots the ellipses (left) and circles (right) indicate (x_1, x_2) variables where the pdf $p(x_1, x_2)$, with the quadratic form, has constant value. This is also indicated in Figure 1, where the vertical lines going down from the density function indicate an interval for the random variable.

2.3 Linear transformations

A transformation $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{b}$ of size $n \times 1$ random vector $\boldsymbol{x}$, with size $m \times n$ fixed matrix $\boldsymbol{F}$ and fixed $m \times 1$ vector $\boldsymbol{b}$, has Gaussian pdf $p(\boldsymbol{y}) = N(\boldsymbol{F}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}')$. This occurs because a linear transformation of Gaussian variables remains Gaussian distributed. The only parameters are then the mean vector and covariance matrix, which can be computed by direct use of expectation and covariance operations.

In particular, extending what was shown for the univariate case, a Gaussian variable can be **transformed to independent** zero-mean and unit-variance variables $\boldsymbol{z} = (z_1, \ldots, z_n)$ by setting $\boldsymbol{F} = \boldsymbol{L}^{-1}$, $\boldsymbol{b} = -\boldsymbol{L}^{-1}\boldsymbol{\mu}$ as follows:

$$\boldsymbol{z} = \boldsymbol{y} = \boldsymbol{L}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \quad \boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}, \quad \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'. \tag{6}$$

Here $\boldsymbol{L}$ is the lower triangular Cholesky factor of the covariance matrix $\boldsymbol{\Sigma}$. For the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ 0.9 & 0.44 \end{bmatrix}. \tag{7}$$

Another useful transformation is obtained by setting $\boldsymbol{b} = 0$ and $\boldsymbol{F} = \mathbf{1}_n'$, where $\mathbf{1}_n$ is a length $n \times 1$ vector of 1 entries. This transformation results in the **sum of Gaussian variables**; $y = x_1 + x_2 + \ldots + x_n$. If the mean values are 0, the mean of the sum will also be 0. The variance of the sum will depend on the covariances between x_i and x_j, $i, j = 1, \ldots, n$. If the variables are i.i.d. with variance σ^2, the sum has variance $n\sigma^2$.

2.4 Conditioning

If we split the random vector in two blocks of variables, denoted $\boldsymbol{x}_A = (x_{A,1}, \ldots, x_{A,n_A})$ and $\boldsymbol{x}_B = (x_{B,1}, \ldots, x_{B,n_B})$, where $n_A + n_B = n$, the mean

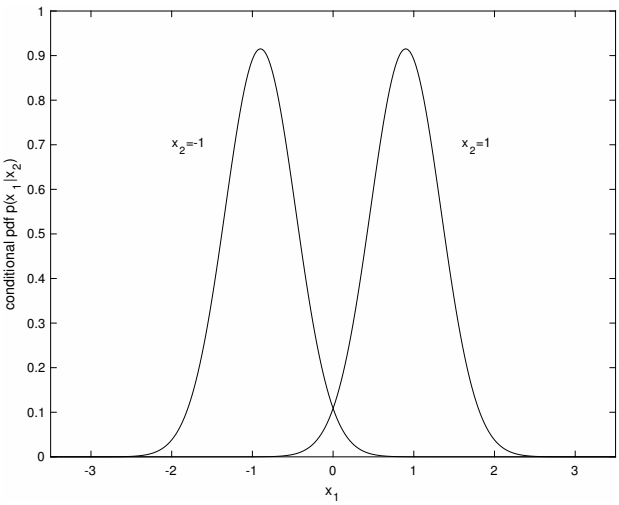

Figure 3: Conditional pdf for x_1 when $x_2 = 1$ or $x_2 = -1$.

and covariance matrix are

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_A, \boldsymbol{\mu}_B), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_A & \boldsymbol{\Sigma}_{A,B} \\ \boldsymbol{\Sigma}_{B,A} & \boldsymbol{\Sigma}_B \end{bmatrix}, \tag{8}$$

where $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are the block mean vectors. Moreover, the matrix $\boldsymbol{\Sigma}_A$ holds the covariance matrix for $\boldsymbol{x}_A$, $\boldsymbol{\Sigma}_B$ the covariance matrix for $\boldsymbol{x}_B$, and $\boldsymbol{\Sigma}_{A,B} = \boldsymbol{\Sigma}'_{B,A}$ is the size $n_A \times n_B$ cross-covariance matrix between $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$.

If we know the variables in the B block, the conditional pdf of $\boldsymbol{x}_A$ is also Gaussian with the following **conditional mean and covariance**:

$$\begin{aligned} E(\boldsymbol{x}_A|\boldsymbol{x}_B) &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{x}_B - \boldsymbol{\mu}_B), \\ \mathrm{Var}(\boldsymbol{x}_A|\boldsymbol{x}_B) &= \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_{B,A}. \end{aligned} \tag{9}$$

For the simplest multivariate pdf with two variables, we let x_1 be the block A variable, while x_2 is the block B variable. We consider the case from Figure 2 left), where there is correlation 0.9 between x_1 and x_2, and they both have zero-mean and unit variance. Figure 3 shows the conditional distribution of the variable x_1 for this case, when we know that $x_2 = -1$ or $x_2 = 1$. Clearly the mean of the pdf is shifted from the unconditional mean of $E(x_1) = 0$. The conditional mean is $E(x_1|x_2) = 0.9x_2$. Further, the variance

is reduced from $\mathrm{Var}(x_1) = 1$ (Figure 1) to $\mathrm{Var}(x_1|x_2) = 1 - 0.9^2 = 0.19$. This shift and reduction in uncertainty would not occur if we had independence between x_1 and x_2. In that case $p(x_1|x_2) = p(x_1)$.

2.5 Sampling

We can generate random **realizations** from the multivariate Gaussian pdf in several ways. One uses sequential sampling of random variables, going from 1 to n. This works because

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)\dots p(x_n|x_{n-1},\dots,x_1). \tag{10}$$

We then start by sampling random variable x_1 according to $p(x_1) = N(\mu_1, \sigma_1^2)$. Next we generate x_2 from the pdf $p(x_2|x_1) = N\left(E(x_2|x_1), \mathrm{Var}(x_2|x_1)\right)$, and so on. For time t we sample x_t from pdf $p(x_t|x_{t-1},\dots,x_1)$. At each stage $t = 2,\dots,n$, the formula for conditional mean and variance is defined in equation (9), for different blocks A and B.

This sequential approach gives the same result as a matrix-vector solution using the Cholesky factorization. It can be explained as follows: The covariance matrix $\boldsymbol{\Sigma}$ is decomposed as $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Sigma}$ as in equation (6). Now, the random simulation can be done by

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}, \quad p(\boldsymbol{z}) = N(\boldsymbol{0}_n, \boldsymbol{I}_n), \tag{11}$$

where $\boldsymbol{0}_n$ is a size $n \times 1$ vector of 0 entries and $\boldsymbol{I}_n$ is the size n identity matrix. Recall from equation (6), or show directly, that equation (11) gives the correct mean $\boldsymbol{\mu}$ and the correct covariance structure since $\boldsymbol{L}\boldsymbol{I}_n\boldsymbol{L}' = \boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Sigma}$. (See Exercise C for more on computing the Cholesky factor.)

3 Stationary Gaussian processes

Consider a random process $x(t) \in \mathbb{R}$ for continuous time or location reference $t \geq 0$. The Gaussian process is defined as follows: For any configuration of n times or locations: $t_1,\dots,t_n$, the variable $\boldsymbol{x} = (x_1,\dots,x_n)$ is multivariate Gaussian distributed. Here, $x_i = x(t_i)$ denotes the random variable at time t_i, $i = 1,\dots,n$.

A Gaussian process is mean (first order) **stationary** if $E\left(x(t)\right) = \mu$ for all times t. In practice one often extends this to model the mean as a

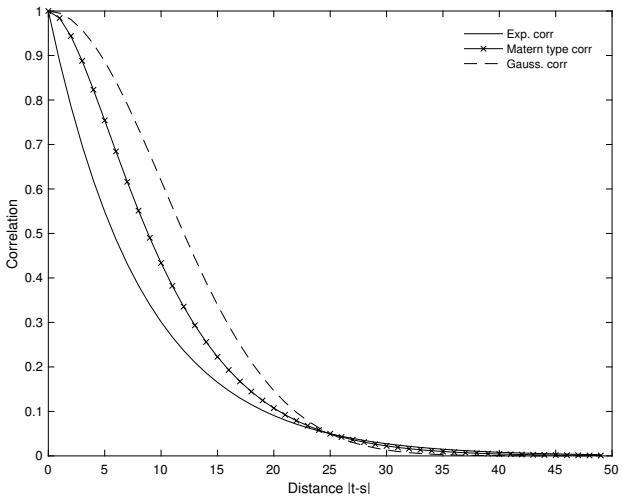

Figure 4: Three different correlation functions.

function of some covariates, like in a regression setting. The process is second order stationary if $\text{Var}\,(x(t)) = \sigma^2$ for all times t and the correlation only depends on the time differences: $\text{Corr}\,(x(t), x(s)) = \text{Corr}\,(x(r+t), x(r+s))$. Since the Gaussian distribution is defined by the first two moments (mean and covariance), the distribution is stationary if the mean and covariance is stationary. This will not hold for other distributions.

3.1 Correlation functions

A key element in Gaussian processes is to model $\text{Corr}\,(x(t), x(s))$, which specifies the dependence in the process over time. There exist famous correlation functions that are used for this purpose. They all start at 1 for $t = s$, and decline to 0 as the distance between times t and s increases. The **exponential correlation function** is defined by $\text{Corr}\,(x(t), x(s)) = \exp(-\phi_E|t-s|)$, a common **Matern type correlation function** is $\text{Corr}(x(t), x(s)) = (1 + \phi_M|t-s|)\exp(-\phi_M|t-s|)$, and the **squared exponential or Gaussian correlation function** is defined by $\text{Corr}(x(t), x(s)) = \exp(-\phi_G|t-s|^2)$. Here, ϕ_E, ϕ_M and ϕ_G are fixed parameters that determine the decay in the different correlation functions.

Figure 4 shows these three different correlation functions. In this display they are all constructed to equal 0.05 at correlation distance $|t-s| = 25$.

This means that the exponential decay parameter $\phi_E = 3/25$ and the squared exponential decay parameter is $\phi_G = 3/25^2$. For the Matern type, we approximated the decay parameter to be about $\phi_M = 0.19$.

Note the differences near $|t - s| = 0$ in Figure 4, where the exponential function comes down much faster than the Matern type, which is again much faster than the squared exponential correlation function. This decay near 0 distance is indicative of the smoothness (differentiability) of the process $x(t)$.

In the special case of the exponential correlation function, the Gaussian process satisfies the Markov property (see Exercise E), and the conditional distribution becomes $p(x_i|x_{i-1}, \ldots, x_1) = p(x_i|x_{i-1})$ for times $t_1 < \ldots < t_{i-1} < t_i$.

There are other correlation functions that are useful in different applications. For instance, some correlation functions go below 0 and have negative correlation for some intermediate distances, and then approach 0 from below. There are also a larger class of Matern correlation functions, with the exponential and squared exponential as special cases. Correlation functions further extend to higher dimensional locations such as the spatial case with $(t_{i,E}, t_{i,N}, t_{i,D})$ being east, north and depth coordinates. However, it is not easy to just come up with a parametric model that gives a valid correlation function. The reason for this challenge is that one must get a positive definite covariance matrix $\boldsymbol{\Sigma}$ for any configuration of times $t_1, \ldots, t_n$. Positive definite here means that any linear combination of variables must have positive variance: for any non-zero size $1 \times n$ vector $\boldsymbol{f}$, $\text{Var}(\boldsymbol{fx}) = \boldsymbol{f\Sigma f}' > 0$. Equivalently, from linear algebra, this means that the smallest eigenvalue of $\boldsymbol{\Sigma}$ is positive. It is common practice to use established correlation functions, like the ones shown in Figure 4, or various combinations of these.

3.2 Sampling Gaussian processes

Gaussian processes are simulated on a defined grid of time points which has the resolution required for the particular application. In the following we illustrate a Gaussian process model on times $t \in [0, 100]$. A regular grid set of times $t = 1, \ldots, 100$ is defined, and the process is simulated on this **discretized grid** of the time domain. One can simulate $\boldsymbol{x} = (x_1, \ldots, x_{100})$, where $x(t_i) = x_i$, using the approach in equation (11). This builds on the Cholesky factorization $\boldsymbol{LL}' = \boldsymbol{\Sigma}$ of the covariance matrix $\boldsymbol{\Sigma}$.

For the illustration we specify a constant mean $\mu = 0$ and variance $\sigma^2 = 1^2$ on the grid of time values. This entails that the 100×100 covariance matrix $\boldsymbol{\Sigma}$

has 1 entries on the diagonal and correlation terms on the off-diagonal entries. The correlation entries depend on the distances and the choice of correlation function. An exponential correlation function with $\phi_E = 3/25$ and a Matern type correlation function with parameter $\phi_M = 0.19$ is used. Figure 4 shows that these parameter values give very small correlation (< 0.05) for points that are more than 25 distance units away from each other. For the Matern type we have $\Sigma_{i,j} = (1 + \phi_M|t_j - t_i|)\exp(-\phi_M|t_j - t_i|)$, $i, j = 1, \ldots, 100$. The covariance matrix can be effectively computed by first forming a matrix of distances, e.g. by setting $\boldsymbol{H} = |\boldsymbol{t}\boldsymbol{1}'_{100} - \boldsymbol{1}_{100}\boldsymbol{t}'|$ for size 100×1 vector of time points $\boldsymbol{t} = (1, 2, \ldots, 100)'$. (There are often established coding routines for extracting distances in software such as R, Matlab or Python. See Exercise D.) The covariance matrix for a Matern-type correlation function is then $\boldsymbol{\Sigma} = (1 + \phi_M\boldsymbol{H}) \otimes \exp(-\phi_M\boldsymbol{H})$, where $\otimes$ means elementwise multiplication.

Figure 5 shows realizations from the exponential and Matern type correlation functions. In this display the realization of independent Gaussian variables, denoted $\boldsymbol{z}$ in equation (11), is identical for the two plots. This means that differences are only due to the correlation structure of the two models, and the way this influences the $\boldsymbol{L}$ matrix for the exponential and the Matern type. The plot shows that the Matern type correlation function gives a smoother process than the exponential function.

Algorithm 1 summarizes the main steps for sampling a random realization from a Gaussian process with Matern type correlation function. If several samples are needed, from the same model, only the last two steps of the algorithm must be re-done.

Algorithm 1 Simulation of a Gaussian process with constant variance and Matern type correlation function

Require: Time points $\boldsymbol{t} = (t_1, \ldots, t_n)$, viewed as a size $n \times 1$ vector. Mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, variance σ^2, correlation function parameter ϕ_M.

1: Build the distance matrix $\boldsymbol{H}$, where $H_{ij} = |t_i - t_j|$.
2: Compute the covariance matrix $\boldsymbol{\Sigma} = \sigma^2(1 + \phi_M\boldsymbol{H}) \otimes \exp(-\phi_M\boldsymbol{H})$.
3: Factorize $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$.
4: Draw n independent standard normal variables $\boldsymbol{z} = (z_1, \ldots, z_n)$.
5: **return** $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}$.

3.3 Conditional process

Assume the process is known at some time points $\boldsymbol{t}_B = (t_{B,1}, \ldots, t_{B,n_B})$, and denote the variables at these time points by $\boldsymbol{x}_B = (x_{B,1}, \ldots, x_{B,n_B})$. When

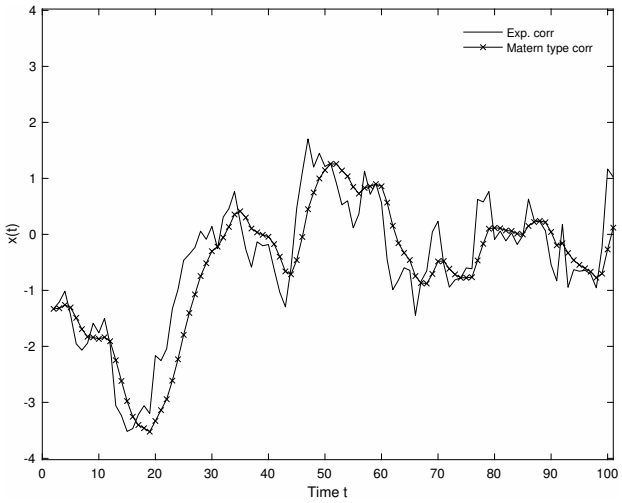

Figure 5: One realization from the Gaussian process with exponential covariance function and one with Matern type correlation function. The mean is 0 and variance 1. The correlation decay parameters are $\phi_E = 3/25$ and $\phi_M = 0.19$.

we now know the outcome of the Gaussian process at some locations, we can use equation (9) to compute the conditional mean and covariance of the process, which defines the conditional Gaussian process. Again, this relies on a discretization of the domain of interest, and we let $\boldsymbol{x}_A = (x_{A,1}, \ldots, x_{A,n_A})$ denote the process at a grid of time points $\boldsymbol{t}_A = (t_{A,1}, \ldots, t_{A,n_A})$ covering the domain of interest.

The conditional formulas in equation (9) require the block covariance matrices, which in this process context will depend on the time differences or distances, as described above. The distance matrix for block B can be defined on vector-matrix form $\boldsymbol{H}_B = |\boldsymbol{t}_B \mathbf{1}'_{n_B} - \mathbf{1}_{n_B} \boldsymbol{t}'_B|$ for size $n_B \times 1$ vector $\boldsymbol{t}_B$, like was done in the previous subsection, or by built-in methods for extracting distances. Similarly, block A covariance matrix $\boldsymbol{\Sigma}_A$ is constructed from distance matrix $\boldsymbol{H}_A$. The size $n_A \times n_B$ cross-covariance matrix $\boldsymbol{\Sigma}_{A,B}$ is

a function of the distances between time points in the A and B sets, e.g. by:

$$\begin{aligned}\boldsymbol{H}_{A,B} &= |\boldsymbol{t}_A \mathbf{1}'_{n_B} - \mathbf{1}_{n_A}\boldsymbol{t}'_B| = \left| \begin{bmatrix} t_{A,1} & \dots & t_{A,1} \\ \dots & \dots & \dots \\ t_{A,n_A} & \dots & t_{A,n_A} \end{bmatrix} - \begin{bmatrix} t_{B,1} & \dots & t_{B,n_B} \\ \dots & \dots & \dots \\ t_{B,1} & \dots & t_{B,n_B} \end{bmatrix} \right| \\ &= \begin{bmatrix} |t_{A,1} - t_{B,1}| & \dots & |t_{A,1} - t_{B,n_B}| \\ \dots & \dots & \dots \\ |t_{A,n_A} - t_{B,1}| & \dots & |t_{A,n_A} - t_{B,n_B}| \end{bmatrix}. \end{aligned} \tag{12}$$

See also Exercise D. When these distance matrices have been built, we can compute the required covariance matrices, and use equation (9) to get the conditional Gaussian distribution.

Algorithm 2 summarizes the main steps of computing the conditional mean and covariance of a Gaussian process, conditional on data $\boldsymbol{x}_B$ at time points $\boldsymbol{t}_B$.

Algorithm 2 Conditional mean and covariance of a Gaussian process with Matern type correlation function

Require: Time points $\boldsymbol{t}_A = (t_{A,1}, \dots, t_{A,n_A})$, viewed as a size $n_A \times 1$ vector. Conditioning variable or data $\boldsymbol{x}_B = (x_{B,1}, \dots, x_{B,n_B})$ at time points $\boldsymbol{t}_B = (t_{B,1}, \dots, t_{B,n_B})$, mean vectors $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$, variance σ^2, correlation function parameter ϕ_M.

1: Build the distance matrices $\boldsymbol{H}_A$, $\boldsymbol{H}_B$ and $\boldsymbol{H}_{A,B}$, e.g. by equation (12).

2: Build the covariance matrices $\boldsymbol{\Sigma}_A = \sigma^2(1 + \phi_M \boldsymbol{H}_A) \otimes \exp(-\phi_M \boldsymbol{H}_A)$, $\boldsymbol{\Sigma}_B = \sigma^2(1 + \phi_M \boldsymbol{H}_B) \otimes \exp(-\phi_M \boldsymbol{H}_B)$ and $\boldsymbol{\Sigma}_{A,B} = \sigma^2(1 + \phi_M \boldsymbol{H}_{A,B}) \otimes \exp(-\phi_M \boldsymbol{H}_{A,B})$.

3: **return** $E(\boldsymbol{x}_A|\boldsymbol{x}_B) = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{x}_B - \boldsymbol{\mu}_B)$, $\mathrm{Var}(\boldsymbol{x}_A|\boldsymbol{x}_B) = \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}'_{A,B}$.

The conditional variance terms are on the diagonal of the conditional covariance matrix $\mathrm{Var}(\boldsymbol{x}_A|\boldsymbol{x}_B)$. Along with the conditional mean, they define the conditional marginal distribution $p(x_{A,i}|\boldsymbol{x}_B)$, $i = 1, \dots, n_A$. The 90 % **conditional prediction interval** at this time point, given $\boldsymbol{x}_B$, is thus

$$\left\{ E(x_{A,i}|\boldsymbol{x}_B) \pm z_{0.05}\sqrt{\mathrm{Var}(x_{A,i}|\boldsymbol{x}_B)} \right\}, \tag{13}$$

where $z_{0.05} = 1.64$ is the upper 5 percentile of the standard Gaussian pdf. If we want to predict the probability that the process is below a threshold a we have

$$p(x_{A,i} < a|\boldsymbol{x}_B) = \Phi\left(\frac{a - E(x_{A,i}|\boldsymbol{x}_B)}{\sqrt{\mathrm{Var}(x_{A,i}|\boldsymbol{x}_B)}} \right), \tag{14}$$

where $\Phi(b) = \int_{-\infty}^{b} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} dz$ is the cumulative distribution function of the standard Gaussian pdf evaluated at b. (See Exercise G.)

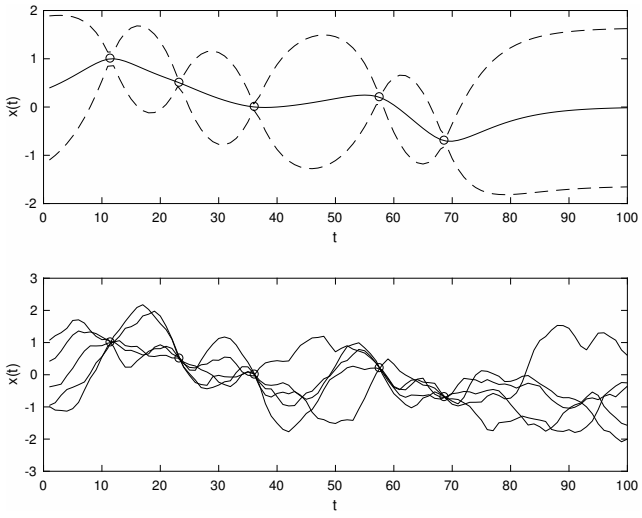

Figure 6: The process is known at selected times (circles). Top: Conditional mean (solid) and 90 % prediction interval (dashed). Bottom: Five realizations from the conditional Gaussian process.

The conditional process can be simulated by applying the Cholesky factorization method described in equation (11) and in Algorithm 1, now applied to the conditional covariance matrix $\mathrm{Var}(\boldsymbol{x}_A|\boldsymbol{x}_B)$. With this sampling method one can get realizations that account for the joint properties in the process model, conditional on $\boldsymbol{x}_B$. In contrast, equation (13) provides only marginal predictions for all $i = 1, \ldots, n$, conditional on $\boldsymbol{x}_B$.

Figure 6 (top) shows conditional prediction intervals as in equation (13), given $\boldsymbol{x}_B = (1, 0.5, 0, 0.2, -0.7)$ at points $\boldsymbol{t}_B = (11.5, 23.3, 36.2, 57.6, 68.7)$. Figure 6 (bottom) shows five conditional realizations of a Gaussian process, given the same data. In this display the discretization grid for the process is again set to $(1, \ldots, 100)$, like in the previous section. The correlation function used is the Matern type with parameter $\phi_M = 0.19$.

4 Brownian motion

The Brownian motion is a continuous time and continuous state model with special requirements to the process increments. Process increments are defined by $x(t_i) - x(t_{i-1})$, for any configuration of times $t_0 = 0 < t_1 < t_2 < \ldots$.

We define the following for the **process increments**:

- $x(t_i) - x(t_{i-1})$ and $x(t_j) - x(t_{j-1})$ are **independent** for all $i \neq j$.
- **Stationarity**, i.e. the distribution of $x(t_i) - x(t_{i-1})$ is identical to that of $x(t_i + s) - x(t_{i-1} + s)$, for any s.
- $x(t_i) - x(t_{i-1})$ is **Gaussian** distributed with 0 mean and variance $\sigma^2(t_i - t_{i-1})$.

The process is assumed to start at $x(0) = 0$. If this is different, the process can simply be shifted to 0.

With $\sigma = 1$ this zero-mean process is sometimes referred to as the standard Brownian motion. There exists various extensions, such as the Brownian motion with drift which has mean $(t_i - t_{i-1})\mu$ for the independent increments, or the geometric Brownian motion which is defined by $y(t) = \exp(x(t))$, where $x(t)$ satisfy the definition for the Brownian motion.

4.1 Properties

Since $x(0) = 0$, the marginal pdf of $x(t)$ is Gaussian with mean 0 and variance $t\sigma^2$. The probability that $p(x(t) > 0) = 1/2$ at any point, because of the symmetry of the Gaussian distribution.

The Brownian motion is a Markov process because of the independent increments, and we have **conditional** pdf

$$p(x_i|x_{i-1}, \ldots, x_0) = p(x_i|x_{i-1}) = N\left(x_{i-1}, (t_i - t_{i-1})\sigma^2\right), \tag{15}$$

where $x_i = x(t_i)$, for any i. The mean is defined by the conditioning variable x_{i-1}, and the variance increases linearly with the time difference. We note that for a stationary zero-mean process with exponential correlation function, the related equation is

$$p(x_i|x_{i-1}, \ldots, x_0) = p(x_i|x_{i-1}) = N\left(\xi x_{i-1}, \sigma^2(1 - \xi^2)\right), \tag{16}$$

where $\xi = \exp(-\phi_E(t_i - t_{i-1}))$ (See exercise E). Here, the mean goes to (the unconditional level) 0 and the variance goes to (the unconditional level) σ^2 as the time difference increases.

Like above, we define $x_i = x(t_i)$, $i = 1, \ldots, n$, to be the process values for grid partitioning $t_1, \ldots, t_n$ of the time line. The **joint** distribution is defined

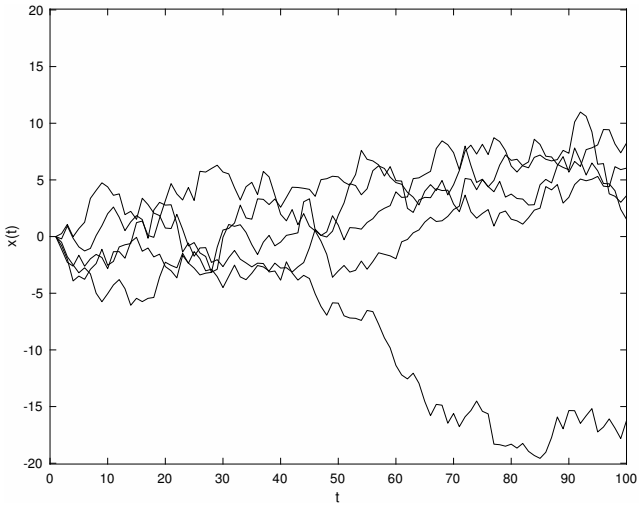

Figure 7: Five realizations of a standard Brownian motion.

by the Gaussian distributed independent increments (see also equation (15)), and we get

$$\begin{aligned} p(\boldsymbol{x}) &= p(x_1)p(x_2|x_1)\dots p(x_n|x_{n-1}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2(t_2-t_1)}}\exp\left(-\frac{1}{2\sigma^2}\frac{x_1^2}{t_2-t_1}\right)\dots \\ &\cdot \quad \frac{1}{\sqrt{2\pi\sigma^2(t_n-t_{n-1})}}\exp\left(-\frac{1}{2\sigma^2}\frac{(x_n-x_{n-1})^2}{t_n-t_{n-1}}\right). \end{aligned} \tag{17}$$

One can **sample**, or simulate, a Brownian motion by using equation (17). On the grid of time values $t_0 = 0 < t_1, \dots < t_n$, with the required resolution, a realization x_i at time t_i is the sum of all independent increments up to that time:

$$x_i = \sum_{k=1}^{i}(t_i - t_{i-1})\sigma z_i = x_{i-1} + (t_i - t_{i-1})\sigma z_i, \tag{18}$$

where z_i, $i = 1, \dots, n$, are independent standard Gaussian variables. Figure 7 shows five realizations of standard Brownian motion variables on the time grid $1, 2, \dots, 100$.

Algorithm 3 summarizes the main steps for sampling a Brownian motion.

Algorithm 3 Simulation of a Brownian motion with mean 0

Require: Time points $t_0 = 0 < t_1 < \ldots < t_n$ at the desired resolution. Initial value $x(0) = 0$. Scale parameter σ.

1: Draw n independent standard normal variables $(z_1, \ldots, z_n)$.

2: **return** $x(t_i) = x(t_{i-1}) + (t_i - t_{i-1})\sigma z_i$, $i = 1, \ldots, n$.

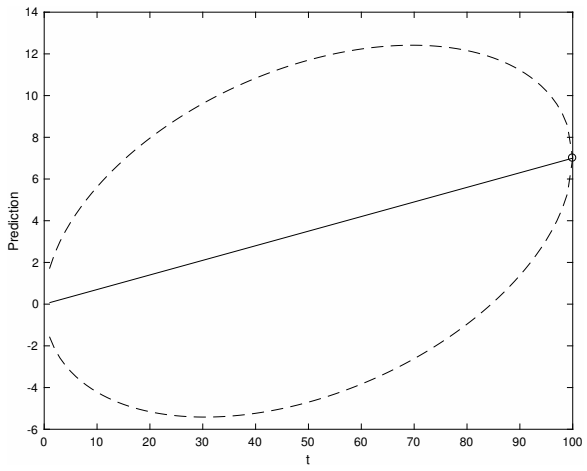

Figure 8: Prediction interval conditional on $x(0) = 0$ and $x(100) = 7$.

Equation (15) defines the forward Markov property for the Brownian motion, where $t_i > t_{i-1}$. The backward process is also Markovian because of the independent increment definition. The **conditional** pdf of x_i given x_j, for $t_j > t_i$ is defined by equation (9). Equation (18) shows that $\text{Cov}(x_i, x_j) = \text{Var}(x_i) = \sigma^2 t_i$, because of the independent increments. This means that the conditional mean and variance are

$$E(x_i|x_j) = \frac{t_i}{t_j}x_j, \quad \text{Var}(x_i|x_j) = \sigma^2 t_i(1 - \frac{t_i}{t_j}). \tag{19}$$

When time t_i is very close to t_j, the conditional mean is close to x_j, and the variance is near 0. The mean decreases linearly from x_j at $t_i = t_j$ to 0 at $t_i = 0$. The variance is 0 at $t_i = 0$ and $t_i = t_j$, and it has a maximum at $t_i = t_j/2$. Figure 8 shows the conditional mean and prediction interval when $x(100) = 7$ for $t \in (0, 100)$, with $\sigma = 1$.

Note that by specifying the time intervals, say $h = t_i - t_{i-1}$ in equation (18), dividing by h on both sides in equation (18), and letting the h approach zero to get a certain limit, the Brownian motion can be seen as a stochastic

differential equation. These formulations are used a lot in several applications involving stochastic dynamics.

4.2 Hitting times

The Brownian motion is used a lot in finance, where it plays an important role in the modeling of random fluctuations in markets. In these situations a common question is related to hitting times. Say, what is the first time that a stock price will exceed threshold a.

Let T_a be the first time the standard Brownian motion, starting at $x(0) = 0$, hits the threshold $a > 0$. The probability that the process exceeds a **threshold** $a > 0$ at time t is

$$\begin{aligned} p(x(t) > a) &= p(x(t) > a | T_a \leq t) P(T_a \leq t) + p(x(t) > a | T_a > t) p(T_a > t) \\ &= p(x(t) > a | T_a \leq t) P(T_a \leq t) = P(T_a \leq t)/2, \end{aligned} \tag{20}$$

since the event $T_a > t$ means that $p(x(t) > a | T_a > t) = 0$, and because of the symmetric Gaussian pdf: $p(x(t) > a | T_a \leq t) = p(x(t) < a | T_a \leq t) = 1/2$. This means that the hitting time distribution is

$$p(T_a \leq t) = 2p(x(t) > a) = 2\left(1 - \Phi\left(a/\sqrt{t\sigma^2}\right)\right). \tag{21}$$

When the time t gets larger $\Phi(a/\sqrt{t\sigma^2})$ approaches $1/2$, and the probability $p(T_a \leq t)$ naturally goes to 1. For fixed t, the probability in equation (21) starts at 1 for $a = 0$, and it goes to 0 when a increases.

Exercises

A: Marginal and conditional probability calculations

Consider a bivariate Gaussian distribution for (x_1, x_2), with mean $(0, 0)$, variance terms equal to σ_1^2, σ_2^2 and correlation ρ.

1. Find $\mathbf{\Sigma}^{-1}$, and write out the quadratic form in the exponent of the joint Gaussian model.

2. Complete the square for x_2, and use this to integrate out this variable x_2. Show that the resulting marginal pdf for x_1 has marginal mean 0 and variance σ_1^2.

3. Use conditional probability, with the joint pdf for (x_1, x_2) in the numerator and the marginal of x_2 in the denominator, to find the conditional distribution for x_1 given x_2.

B: Covariance identities

Consider block variables $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$, where $\boldsymbol{x} = (\boldsymbol{x}_A, \boldsymbol{x}_B)$. Set means equal to 0. The covariance matrix is defined in equation (8). The inverse covariance matrix, also known as the precision matrix, has the same block structure as the covariance matrix:

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_A & \boldsymbol{Q}_{A,B} \\ \boldsymbol{Q}_{B,A} & \boldsymbol{Q}_B \end{bmatrix}. \tag{22}$$

1. Use $\boldsymbol{Q}\boldsymbol{\Sigma} = \boldsymbol{I}$ to find the relations between the block matrix entries of $\boldsymbol{Q}$ and the block matrix entries of $\boldsymbol{\Sigma}$.

2. The conditional pdf of $p(\boldsymbol{x}_A|\boldsymbol{x}_B) \propto p(\boldsymbol{x}_A, \boldsymbol{x}_B)$, because $\boldsymbol{x}_B$ is known. Use the quadratic form to find the conditional mean and variance as a function of $\boldsymbol{Q}$ and $\boldsymbol{x}_B$, and relate this to problem B.1 and equation (9).

C: Cholesky factor

For a covariance matrix $\boldsymbol{\Sigma}$ the Cholesky factorization is defined by $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Sigma}$ (see equation (6)). Here, $\boldsymbol{L}$ is a lower triangular matrix, i.e. $L_{i,j} = 0$ for $i < j$.
(In Matlab and R the matrix $\boldsymbol{L}'$ is found by *chol*, in Python one can use *np.linalg.cholesky*.)

1. Consider a zero mean bivariate Gaussian distributions with the following covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}. \tag{23}$$

   Find the Cholesky factorization of the covariance matrix in equation (23). Check your calculations on the computer.

2. Sample and visualize 1000 variables from the distribution, using the matrix found in C.1. Do the same for the 2×2 matrix in equation (7). Compare the two bivariate distributions.

3. Consider a mean zero trivariate Gaussian distribution with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0.1 \\ 0.9 & 1 & 0.2 \\ 0.1 & 0.2 & 1 \end{bmatrix}. \tag{24}$$

Find the Cholesky factorization and use it to sample 1000 samples from this distribution.
(You can visualize 3D scatterplots by e.g. *scatter3* or *plot3* in Matlab, *scatterplot3d* in R (see also *ggplot2*), or *scatter* in Python.)

D: Distances

Specify a number of points along the line; say $\boldsymbol{t}_A = (3, 5, 9, 10, 13, 20)$, $\boldsymbol{t}_B = (3.5, 5.2, 7.8, 12.1)$.

1. Study ways to build the distance matrices $\boldsymbol{H}_A$, $\boldsymbol{H}_B$ and $\boldsymbol{H}_{A,B}$ between these points. (See equation (12).) Check out various implementation methods; for-loops, the described vectorization method, or other built-in approaches.
(In Matlab you can use the built-in functions *pdist* and *squareform*, and likewise in Python after importing
from scipy.spatial.distance import pdist
from scipy.spatial.distance import squareform
In R you have *dist*.)

E: Exponential correlation and Markov property

Consider a Gaussian process with mean 0 mean and variance 1. Consider three points $r < s < t$ on the real line. The goal is to predict $x(t)$, conditional on $x(s)$ and $x(r)$.

1. Assuming the exponential correlation function is valid, compute the conditional mean and variance of $x(t)$, given $x(s)$ and $x(r)$. Show that this process is Markovian.

2. Assuming the squared exponential correlation function, compute the conditional mean and variance of $x(t)$, given $x(s)$ and $x(r)$. Show that this process is not Markovian.

F: Simulation of Gaussian processes

1. Specify points $t = 1, 2, \ldots, 100$. Sample 10 realizations of the zero-mean unconditional Gaussian process with exponential correlation function for parameter $\phi_E = 3/10$, and 10 others with $\phi_E = 3/30$.

2. Data $(0.58, -1.34, 0.61)$ are provided at points $\boldsymbol{t} = (11.2, 51.8, 81.4)$. Consider again the points $t = 1, 2, \ldots, 100$, and sample 10 realizations of the conditional Gaussian process, given the data, with exponential correlation function for parameter $\phi_E = 3/10$, and 10 others with $\phi_E = 3/30$.

G: Production quality and temperatures

One application of Gaussian processes is optimization of complicated functions or experimental configurations. Examples include logistics or operational planning where the response is a complex function of several input variables, decisions about production controls in natural resources utilization or environmental treatment settings for efficient cleaning of pollutants, and many others. The goal in such applications is to find the input value t that gives the optimal output, that is the value that maximizes some 'profit' according to

$$\hat{t} = \text{argmax}_t \left(x(t) \right).$$

Because it is time-consuming or costly to evaluate $x(t)$, one can only run the experiments for some inputs t_i, resulting in outputs $x(t_i)$ $i = 1, \ldots, n_B$. Conditional on the data, one either chooses the best input value, or one looks for the next most valuable evaluation point using a Gaussian process approximation for the outputs.

We look at an industrial application where product quality $x(t)$ has a very complex relation to the temperature t input variable. At the production unit they are aiming for high quality, and the goal is to produce at the best possible temperature, but they do not know this optimal temperature value. Initially the product quality is assumed to be a Gaussian process with mean $E(x(t)) = 50$ for all temperatures, $\text{Var}(x(t)) = 4^2$ and $\text{Corr}(x(t), x(s)) = (1 - 0.2|t - s|) \exp(-0.2|t - s|)$.

The production unit allows experimentation to get data and guide the search for optimal temperature inputs. Experiments are very costly, so in this case the product quality has only been evaluated for five temperature

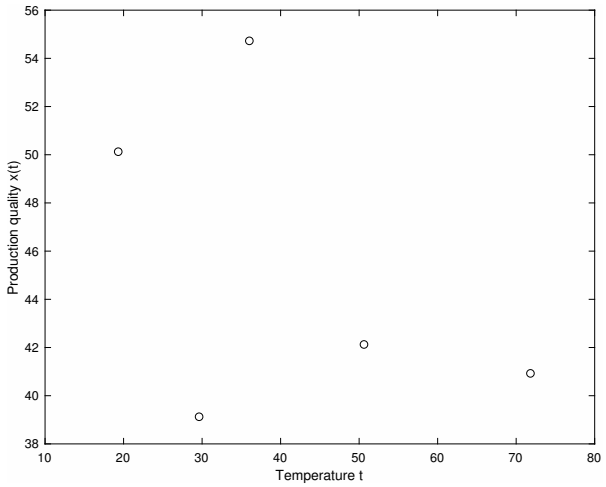

Figure 9: Plot of the production quality at five temperature evaluation points.

values. Figure 9 shows the results from the experiments. The five evaluation points are $(t, x(t))$: $(19.4, 50.1)$, $(29.7, 39.1)$, $(36.1, 54.7)$, $(50.7, 42.1)$ and $(71.9, 40.9)$.

They will use the Gaussian process model (with the specified parameters) along with the available data to find useful temperatures for testing.

1. Define a regular grid of temperature points from $t = 10$ to $t = 80$ degrees, with spacing 0.5 ($n = 141$ point). Construct the required mean vectors and covariance matrices to compute the conditional mean and covariance of the process at the 141 points, given the five evaluation points. Display the prediction as a function of time, along with the 90 % conditional prediction intervals.
   (The cumulative distribution function is built in to Matlab (*normcdf*), R (*pnorm*) and Python (*norm.cdf*, after adding *from scipy.stats import norm* on top of file).)

2. The company has a goal of production quality above 57. Use the predictions from G.1 to compute the probability that $x(t) > 57$ for all t on the grid of 141 points, given the evaluation points. Plot the probability as a function of time.

3. The company decides to run another experiment with temperature $t = 40.7$. The result is $x(40.7) = 49.7$. Augment the evaluation set with this information, i.e. six points in total. Given the new information, compute and visualize the prediction, prediction intervals and the probabilities that $x(t) > 57$, for all t on the grid of 141 points. If the company has budget for yet another experiment, which input temperature would you recommend, considering their goal for production quality?

H : Brownian motion for stock price

Assume the price of a certain stock develops according to a Brownian motion with 0 mean and noise standard error $\sigma = 0.75$ per day. On Jan 1st, the price of the stock is \$ 40. In this exercise the goal is to predict the future stock price.

1. What is the probability that the stock price is larger than \$ 50 on May 1st (120 days ahead)? Find the solution by analytical calculations and approximate the solution by generating and plotting 100 realizations of the random process with a daily resolution until 1 May.

2. On March 2, the price of the stock is \$ 45. What is now the probability that the stock price is larger than \$ 50 on May 1st (60 days ahead)? Find again the solution by analytical calculations and approximate the solution by generating and plotting 100 realizations of the random process with a daily resolution until 1 May.

3. Assume that we know only the stock price of \$ 40 on Jan 1st. There is interest in the waiting time until the stock price has gone up 10 % (to \$ 44). Find the distribution analytically and approximate this distribution with sorted simulated hitting times over 100 realizations. (You can truncate the simulations at some time (say $t = 10000$) to avoid the tail in the hitting time distribution.)

References

Johnson, R. A., D. W. Wichern, et al. (2014). *Applied multivariate statistical analysis*. Prentice-Hall New Jersey.

Øksendal, B. (2003). *Stochastic differential equations*. Springer.

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

Vanhatalo, J., J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari (2013). Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research 14*(Apr), 1175–1179.