San José State University

Math 263: Stochastic Processes

Gaussian processes

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- Section 10.7

Outline of the presentation

- Multivariate normal distributions
- Definition of Gaussian Processes
- Examples
- Gaussian Processes Regression

First, we recall the definition of multivariate normal distributions.

Def 0.1 **($\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$).** We say that a k-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)^T$ has a multivariate normal distribution, if their joint density has the form

$$f(x_1, \ldots, x_k) = (2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where

- $\boldsymbol{\mu} \in \mathbb{R}^k$: mean vector;
- $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$: covariance matrix.

Remark. In a multivariate normal variable, $\mathbf{X} = (X_1, \ldots, X_k)^T$,

- Each component $X_i \sim N(\mu_i, \mathbf{\Sigma}_{ii})$.
- Each pair of components $\text{Cov}(X_i, X_j) = \mathbf{\Sigma}_{ij}$
- Every linear combination of components has a univariate normal distribution (note that this is also a sufficient condition for the joint normality):

$$Y = \mathbf{a}^T \mathbf{X} = a_1 X_1 + \cdots + a_k X_k \sim N(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \mathbf{\Sigma} \mathbf{a})$$

- Another sufficient and necessary condition for joint normality is that the random variables $X_1, \ldots, X_k$ are linear combinations of several independent normal random variables.

Remark. When $k=2$ (bivariate normal), the above density reduces to

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot$$

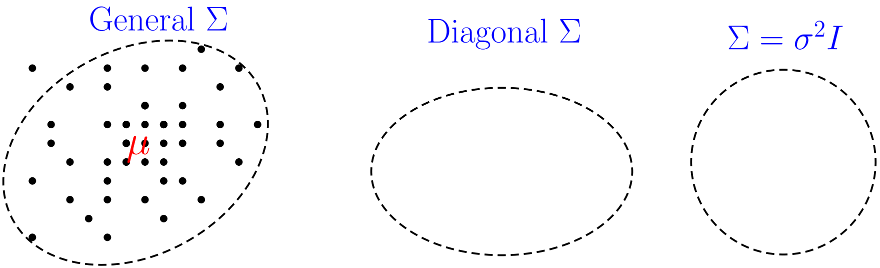
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right)$$

Here, the mean and covariance of the bivariate normal are

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

The conditional distribution of X_2 given X_1 is

$$X_2 \mid X_1 = x_1 \quad \sim \quad N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1-\mu_1),\ (1-\rho^2)\sigma_2^2\right).$$


General Σ
μ
Diagonal Σ
$\Sigma = \sigma^2 I$

More generally, we have the following result.

Theorem 0.1. If $\mathbf{X}_A \in \mathbb{R}^a$ and $\mathbf{X}_B \in \mathbb{R}^b$ jointly have a multivariate normal distribution, i.e.,

$$\begin{pmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix} \right),$$

then the conditional distribution of $\mathbf{X}_B$ given $\mathbf{X}_A = \mathbf{x}_A$ is also multivariate normal:

$$\mathbf{X}_B \mid \mathbf{X}_A = \mathbf{x}_A \quad \sim \quad N\left(\boldsymbol{\mu}_B + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}(\mathbf{x}_A - \boldsymbol{\mu}_A),\ \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{AB}\right).$$

We are now ready to present the definition of Gaussian processes.

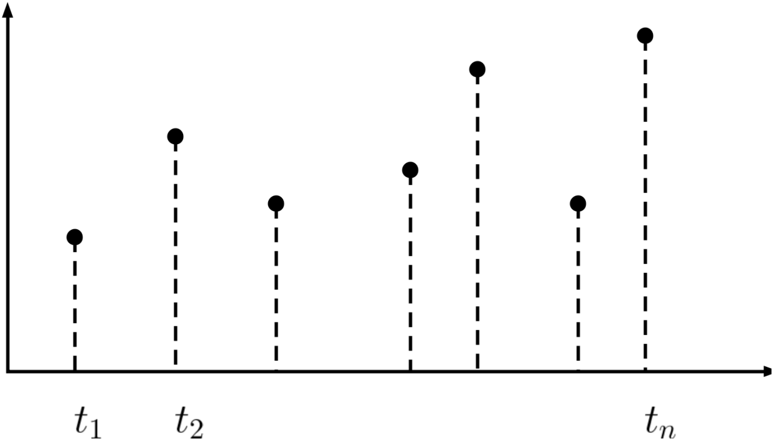
Def 0.2. A stochastic process $X(t), t \geq 0$ is called a **Gaussian process** with mean function $\mu(\cdot)$ and covariance function $\kappa(\cdot,\cdot)$ if for all $n \in \mathbb{Z}^+$ and all $t_1, \dots, t_n > 0$, the collection $X(t_1), \dots, X(t_n)$ have a multivariate normal distribution:

$$(X(t_1), \dots, X(t_n))^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\mu} = (\mu(t_1), \dots, \mu(t_n))^T, \quad \boldsymbol{\Sigma} = (\kappa(t_i, t_j))_{1 \leq i,j \leq n}$$

(This is the same as saying that every linear combination of $X(t_1), \dots, X(t_n)$, including each of them individually, has a univariate Gaussian distribution)


t_1
t_2
t_n

Example 0.1. Brownian motion processes are Gaussian processes.

Proof. For all $t_1, \ldots, t_n > 0$, each $X(t_i)$ is a linear combination of the independent normal random variables $X(t_1), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$. Thus, $X(t_1), \ldots, X(t_n)$ collectively have a multivariate normal distribution.

Remark. For the standard Brownian motion treated as a Gaussian process, the above collection have zero mean $\boldsymbol{\mu} = \mathbf{0}$ and covariances: For $t_i < t_j$,

$$\mathrm{Cov}(X(t_i), X(t_j)) = \mathrm{Cov}(X(t_i), X(t_i) + X(t_j) - X(t_i)) = \mathrm{Cov}(X(t_i), X(t_i)) = t_i$$

For this model, the underlying mean and covariance functions are

$$\mu(t) = 0, \quad \kappa(s,t) = \mathrm{Cov}(X(s), X(t)) = \min(s,t)$$

In general, one can define new Gaussian processes by choosing proper covariance functions $\kappa(s,t) = \mathrm{Cov}(X(s), X(t))$:

- $\kappa(s,t) = \min(s,t)$
- $\kappa(s,t) = \exp\left(-\frac{(s-t)^2}{2\tau^2}\right)$
- $\kappa(s,t) = (st+c)^2$

For simplicity, we always set the mean function to zero: $\mu(t) = 0$ (if it is not zero, then we can always remove it from the process: $X(t) \leftarrow X(t) - \mu(t)$)

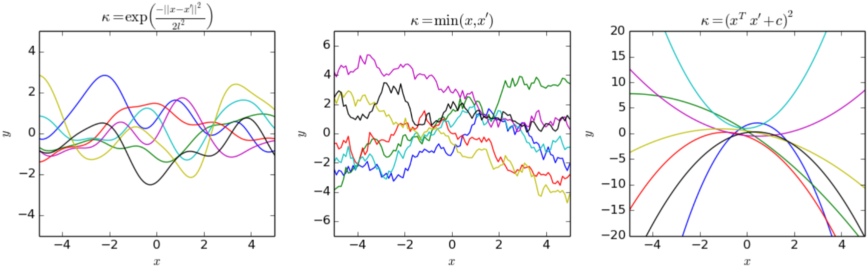


Figure 1: Samples from different Gaussian processes (corresponding to different covariance functions)

Def 0.3. Let $\{X(t), t \geq 0\}$ be a standard Brownian motion process. The conditional stochastic process $\{X(t), 0 \leq t \leq 1 \mid X(1) = 0\}$ is called the **Brownian bridge** process.

Theorem 0.2. Brownian bridge is a Gaussian process with mean $\mu(t) = 0$ and covariance function $\mathrm{Cov}(X(s), X(t)) = s(1-t)$ for $s < t < 1$.

Proof. For all $0 < t_1 < \cdots < t_n < 1$, the joint distribution of

$$X(t_1), \ldots, X(t_n), X(1)$$

is multivariate normal.

It follows that the the conditional joint distribution of $X(t_1), \ldots, X(t_n)$ given $X(1) = 0$ is also multivariate normal.

This shows that Brownian bridge is a Gaussian process.

We compute the mean and covariance of the process below.

First, for any $s < 1$,

$$\mathrm{E}(X(s) \mid X(1) = 0) = 0$$

where we used the result

$$X(s) \mid X(t) = B \quad \sim \quad N(Bs/t, s(t-s)/t).$$

Next, for any $s < t < 1$,

$$
\begin{aligned}
\mathrm{Cov}[(X(s), X(t)) \mid X(1) = 0] &= \mathrm{E}(X(s)X(t) \mid X(1) = 0) \\
&= \mathrm{E}[\mathrm{E}(X(s)X(t) \mid X(t), X(1) = 0) \mid X(1) = 0] \\
&= \mathrm{E}[X(t)\mathrm{E}(X(s) \mid X(t)) \mid X(1) = 0] \\
&= \mathrm{E}[X(t)\frac{s}{t}X(t) \mid X(1) = 0] \\
&= \frac{s}{t}\mathrm{E}[X(t)^2 \mid X(1) = 0] \\
&= \frac{s}{t}\frac{t(1-t)}{1} = s(1-t).
\end{aligned}
$$

Another way of obtaining a Brownian bridge process is below.

Theorem 0.3. Let $\{X(t), t \geq 0\}$ be a standard Brownian motion process. Then

$$Z(t) = X(t) - tX(1),\ 0 \leq t \leq 1$$

is a Brownian bridge process.

Proof. For all $0 < t_1, \ldots, t_n < 1$, the random variables

$$Z(t_1) = X(t_1) - t_1 X(1),\ \ldots,\ Z(t_n) = X(t_n) - t_n X(1)$$

are all linear combinations of the $X(t_1), \ldots, X(t_n), X(1)$, thus jointly having a multivariate normal distribution.

It suffices to show that it has the same mean and covariance functions with the Brownian bridge:

$$
\begin{aligned}
\mathrm{E}(Z(t)) &= \mathrm{E}(X(t)) - t\mathrm{E}(X(1)) = 0 \\
\mathrm{Cov}(Z(s), Z(t)) &= \mathrm{Cov}(X(s) - sX(1), X(t) - tX(1)) \\
&= \mathrm{Cov}(X(s), X(t)) - t\,\mathrm{Cov}(X(s), X(1)) - s\,\mathrm{Cov}(X(1), X(t)) \\
&\qquad + st\,\mathrm{Cov}(X(1), X(1)) \\
&= s - ts - st + st \\
&= s(1-t), \quad \text{for } s < t.
\end{aligned}
$$

Def 0.4. Let $\{X(t), t \geq 0\}$ be a Brownian motion process. The process $\{Z(t), t \geq 0\}$ defined by

$$Z(t) = \int_0^t X(s)\,\mathrm{d}s, \quad \text{for all } t \geq 0$$

is called **Integrated Brownian motion**.

Interpretation:

- $Z(t)$: price of certain commodity at time t
- $X(t) = \frac{\mathrm{d}}{\mathrm{d}t} Z(t)$: rate of change of price at time t ← Brownian motion

Theorem 0.4. The integrated Brownian motion $\{Z(t), t \geq 0\}$ is also a Gaussian process. When it is defined on the standard Brownian motion, we have

$$\mathrm{E}(Z(t)) = 0, \quad \mathrm{Cov}(Z(s), Z(t)) = s^2\left(\frac{t}{2} - \frac{s}{6}\right), \; s < t$$

Proof. The joint normality of $Z(t)$ in different locations $t_1, \ldots, t_n$ can be shown by writing the integral as a limit of approximating sums.

We now verify the mean and covariance functions for the integrated standard Brownian motion: $Z(t) = \int_0^t X(s)\,\mathrm{d}s$, where $X(t)$ is standard Brownian motion.

First,

$$\mathrm{E}(Z(t)) = \mathrm{E}\left(\int_0^t X(s)\,\mathrm{d}s\right) = \int_0^t \mathrm{E}(X(s))\,\mathrm{d}s = 0$$

For the covariance part, suppose $s < t$. Then

$$\begin{aligned}
\mathrm{Cov}(Z(s), Z(t)) &= \mathrm{E}(Z(s)Z(t)) \\
&= \mathrm{E}\left(\int_0^s X(u)\,\mathrm{d}u \int_0^t X(v)\,\mathrm{d}v\right) \\
&= \mathrm{E}\left(\int_0^s \int_0^t X(u)X(v)\,\mathrm{d}v\,\mathrm{d}u\right) \\
&= \int_0^s \int_0^t \mathrm{E}\left(X(u)X(v)\right)\,\mathrm{d}v\,\mathrm{d}u \\
&= \int_0^s \int_0^t \min(u, v)\,\mathrm{d}v\,\mathrm{d}u \\
&= s^2\left(\frac{t}{2} - \frac{s}{6}\right).
\end{aligned}$$

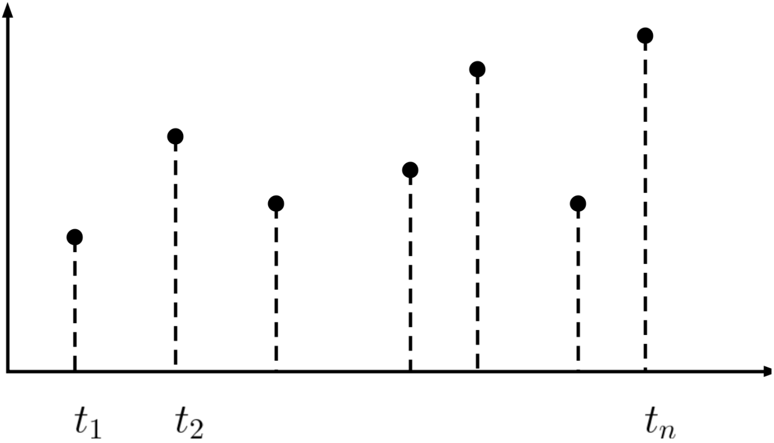
Gaussian process regression

Gaussian processes provide a way to model the probability distribution of functions $X : [0,\infty) \mapsto \mathbb{R}$ (which are regarded as sampled trajectories of a Gaussian process):

$$X(\cdot) \ \sim \ \mathtt{GP}(\mu(\cdot), \kappa(\cdot,\cdot))$$

That is, for any finite number of sampled locations $t_1, t_2, \ldots, t_n > 0$, the restriction of the function to those fixed locations is assumed to have a multivariate normal distribution:

$$\vec{X} = (X(t_1), X(t_2), \ldots, X(t_n))^T \sim N((\mu(t_i))_{1 \le i \le n}, (\kappa(t_i, t_j))_{1 \le i, j \le n})$$


t_1
t_2
t_n

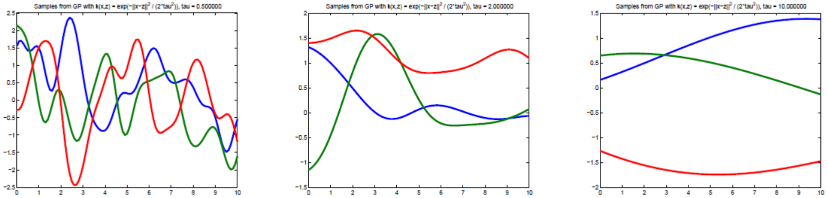
We assume $\mu = 0$ and first consider the squared exponential kernel function

$$\kappa_{\mathrm{SE}}(t, t') = \exp(-\|t - t'\|^2 / 2\tau^2)$$

Functions drawn from such a Gaussian process will tend to be distributed around zero and "locally smooth"with high probability; i.e.,

- nearby function values are highly correlated, and
- the correlation drops off as a function of distance in the input space.

This can be thought of as a prior distribution for functions in the context of Bayesian inference.

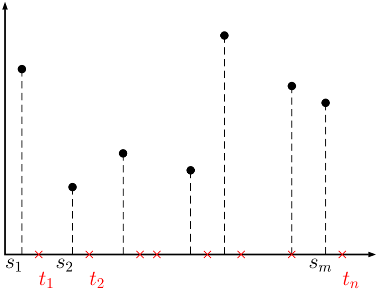


(a) (b) (c)

Figure 2: Samples from a zero-mean Gaussian process prior with $k_{SE}(\cdot,\cdot)$ covariance function, using (a) $\tau = 0.5$, (b) $\tau = 2$, and (c) $\tau = 10$. Note that as the bandwidth parameter τ increases, then points which are farther away will have higher correlations than before, and hence the sampled functions tend to be smoother overall.

Now consider the regression setting where we have a set of (noiseless) observations $(s_i, x_i), i = 1, \ldots, m$, from some unknown function.

Given new locations $t_i, i = 1, \ldots, n$, the goal is to predict the values of the function at those locations.



This turns out to be equivalent to finding the conditional distribution of $X(t_1), \ldots, X(t_n)$ given $X(s_1) = x_1, \ldots, X(s_m) = x_m$.

First, the union of the two sets of random variables have a joint multivariate normal distribution

$$(X(t_1), \ldots, X(t_n), X(s_1), \ldots, X(s_m)) \sim N(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} \Sigma^{(tt)} & \Sigma^{(ts)} \\ \Sigma^{(st)} & \Sigma^{(ss)} \end{pmatrix}.$$

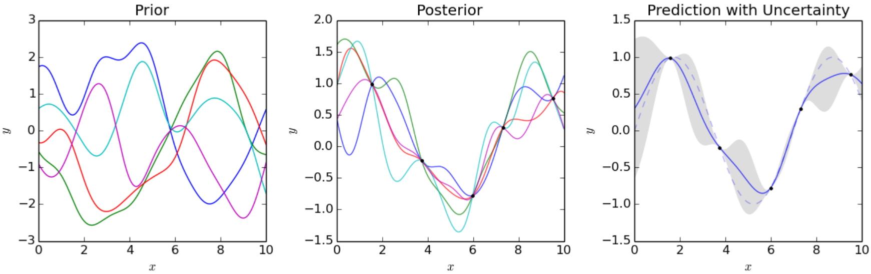
It follows that

$$\begin{aligned} &X(t_1), \ldots, X(t_n) \mid X(s_1) = x_1, \ldots, X(s_m) = x_m \\ &\quad \sim N\left(\Sigma^{(ts)} \left(\Sigma^{(ss)}\right)^{-1} \mathbf{x},\ \Sigma^{(tt)} - \Sigma^{(ts)} \left(\Sigma^{(ss)}\right)^{-1} \Sigma^{(st)}\right) \end{aligned}$$

Thus, our prediction would be

$$\mathrm{E}[X(t_1),\ldots,X(t_n)\,|\,X(s_1)=x_1,\ldots,X(s_m)=x_m]=\Sigma^{(ts)}\left(\Sigma^{(ss)}\right)^{-1}\mathbf{x}.$$

See the figure below for a demonstration.



Remark. Special cases:

- $m = n = 1$:

$$\mathrm{E}[X(t) \mid X(s) = x] = \exp\left(-\frac{|t-s|^2}{2\tau^2}\right) x.$$

- $m = 1, n > 1$:

$$\mathrm{E}[X(t_1), \ldots, X(t_n) \mid X(s) = x] = \left(e^{-|t_1-s|^2/2\tau^2}, \ldots, e^{-|t_n-s|^2/2\tau^2}\right) x.$$

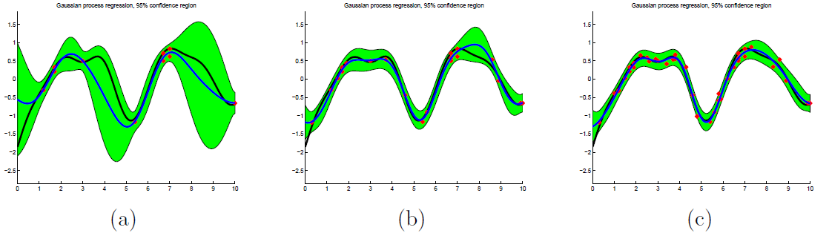
- $m = 2, n = 1$:

$$\mathrm{E}[X(t) \mid X(s_1) = x_1, X(s_2) = x_2]$$

$$= \left(e^{-|t-s_1|^2/2\tau^2}, e^{-|t-s_2|^2/2\tau^2}\right) \begin{pmatrix} 1 & e^{-|s_1-s_2|^2/2\tau^2} \\ e^{-|s_2-s_1|^2/2\tau^2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Remark. When having independent $N(0,\sigma^2)$ noise, the formula becomes

$$\mathrm{E}[X(t_1),\ldots,X(t_n) \mid X(s_1)=x_1,\ldots,X(s_m)=x_m] = \Sigma^{(ts)}\left(\Sigma^{(ss)}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{x}.$$



(a) (b) (c)

Figure 3: Gaussian process regression using a zero-mean Gaussian process prior with $k_{SE}(\cdot,\cdot)$ covariance function (where $\tau = 0.1$), with noise level $\sigma = 1$, and (a) $m = 10$, (b) $m = 20$, and (c) $m = 40$ training examples. The blue line denotes the mean of the posterior predictive distribution, and the green shaded region denotes the 95% confidence region based on the model's variance estimates. As the number of training examples increases, the size of the confidence region shrinks to reflect the diminishing uncertainty in the model estimates. Note also that in panel (a), the 95% confidence region shrinks near training points but is much larger far away from training points, as one would expect.

A MATLAB demonstration

https://www.mathworks.com/help/stats/gaussian-process-regression-models.html

Comments on GP regression

- Nonparametric (lazy learning)
- Flexible, powerful
- Computationally intensive for high dimensional data (overfitting could occur as well).

Further learning on GP regression

- Stanford CS 229 Lecture Notes[1]
- Gaussian Processes for Machine Learning (textbook)[2]
- The Gaussian Processes Website[3]

[1] http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf

[2] http://www.gaussianprocess.org/gpml/chapters/RW.pdf

[3] http://www.gaussianprocess.org