# *Exercises 3: Gaussian processes*

## Basics

A *Gaussian process* is a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ such that, for any finite collection of indices $x_1, \ldots, x_N \in \mathcal{X}$, the random vector $[f(x_1), \ldots, f(x_N)]^T$ has a multivariate normal distribution. It is a generalization of the multivariate normal distribution to infinite-dimensional spaces. The set $\mathcal{X}$ is called the index set or the state space of the process, and need not be countable.

A Gaussian process can be thought of as a random function defined over $\mathcal{X}$, often the real line or $\mathbb{R}^p$. We write $f \sim \mathrm{GP}(m, C)$ for some mean function $m : \mathcal{X} \to \mathbb{R}$ and a covariance function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. These functions define the moments[1] of all finite-dimensional marginals of the process, in the sense that

[1] And therefore the entire distribution, because it is normal

$$E\{f(x_1)\} = m(x_1) \quad \text{and} \quad \mathrm{cov}\{f(x_1), f(x_2)\} = C(x_1, x_2)$$

for all $x_1, x_2 \in \mathcal{X}$. More generally, the random vector $[f(x_1), \ldots, f(x_N)]^T$ has covariance matrix whose $(i, j)$ element is $C(x_i, x_j)$. Typical covariance functions are those that decay as a function of increasing distance between points $x_1$ and $x_2$. The notion is that $f(x_1)$ and $f(x_2)$ will have high covariance when $x_1$ and $x_2$ are close to each other.

(A) Define the *squared exponential* covariance function as

$$C_{SE}(x_1, x_2) = \tau_1^2 \exp\left\{-\frac{1}{2}\left(\frac{x_1 - x_2}{b}\right)^2\right\} + \tau_2^2 \delta(x_1, x_2),$$

where $(b, \tau_1^2, \tau_2^2)$ are constants (often called *hyperparameters*), and where $\delta(a, b)$ is the Kronecker delta function that takes the value 1 if $a = b$, and 0 otherwise.

Let's start with the simple case where $\mathcal{X} = [0, 1]$, the unit interval. Write an R function that simulates a mean-zero Gaussian process on $[0, 1]$ under the squared-exponential covariance function. The function will accept as arguments: (1) finite set of points $x_1, \ldots, x_N$ on the unit interval; and (2) a triplet $(b, \tau_1^2, \tau_2^2)$. It will return the value of the random process at each point: $f(x_1), \ldots, f(x_N)$.

Use your function to simulate (and plot) Gaussian processes across a range of values for $h$, $\tau_1^2$, and $\tau_2^2$. Try starting with a very small value of $\tau_2^2$ (say, $10^{-6}$) and playing around with the other two first. On the basis of your experiments, describe the role of these three hyperparameters in controlling the overall behavior of the random functions that result. What happens when you try $\tau_2^2 = 0$? Why? If you can fix this, do—remember our earlier discussion on different ways to simulate the MVN.

(B) Suppose you observe the value of a Gaussian process $f \sim \text{GP}(m, C)$ at points $x_1, \ldots, x_N$. What is the conditional distribution of the value of the process at some new point $x^\star$? For the sake of notational ease simply write the value of the $(i, j)$ element of the covariance matrix as $C_{i,j}$, rather than expanding it in terms of a specific covariance function.

(C) Prove the following lemma.

**Lemma 1** *Suppose that the joint distribution of two vectors $y$ and $\theta$ has the following properties: (1) the conditional distribution for $y$ given $\theta$ is multivariate normal, $(y \mid \theta) \sim N(R\theta, \Sigma)$; and (2) the marginal distribution of $\theta$ is multivariate normal, $\theta \sim N(m, V)$. Assume that $R$, $\Sigma$, $m$, and $V$ are all constants. Then the joint distribution of $y$ and $\theta$ is multivariate normal.*

## In nonparametric regression

(A) Suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for some unknown function $f$. Suppose that the prior distribution for the unknown function is a mean-zero Gaussian process: $f \sim \text{GP}(0, C)$ for some covariance function $C$. Let $x_1, \ldots, x_N$ denote the previously observed $x$ points. Derive the posterior distribution for the random vector $[f(x_1), \ldots, f(x_N)]^T$, given the corresponding outcomes $y_1, \ldots, y_N$, assuming that you know $\sigma^2$.

(B) As before, suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, N$. Now we wish to predict the value of the function $f(x^\star)$ at some new point $x^\star$ where we haven't seen previous data. Suppose that $f$ has a mean-zero Gaussian process prior, $f \sim GP(0, C)$. Show that the posterior mean $E\{f(x^\star) \mid y_1, \ldots, y_N\}$ is a linear smoother, and derive expressions both for the smoothing weights and the posterior variance of $f(x^\star)$.

(C) Go back to the utilities data, and plot the pointwise posterior mean and 95% posterior confidence interval for the value of the function at each of the observed points $x_i$ (again, superimposed on top of the scatter plot of the data itself). Choose $\tau_2^2$ to be very small, say $10^{-6}$, and choose $(b, \tau_1^2)$ that give a sensible-looking answer.

(D) Let $y_i = f(x_i) + \epsilon_i$, and suppose that $f$ has a Gaussian-process prior under the squared-exponential covariance function $C$ with scale $\tau^1_2$, range $b$, and nugget $\tau^2_2$. Derive an expression for the marginal distribution of $y = (y_1 \ldots, y_N)$ in terms of $(\tau^1, b, \tau^2_2)$, integrating out the random function $f$. This is called a marginal likelihood.

(E) Return to the utilities data. Fix $\tau^2_2 = 0$, and evaluate the marginal likelihood function $p(y \mid \tau^2_1, b)$ over a discrete 2-d grid of points. If you're getting errors in your code with $\tau^2_2 = 0$, use something very small instead. Use this plot to choose a set of values $(\hat{\tau^2_1}, \hat{b})$ for the hyperparameters. Then use these hyperparameters to compute the posterior mean for $f$, given $y$.