Stochastic Simulation

Autumn Semester 2024

Prof. Fabio Nobile *Assistant: Matteo Raviola*

Project - 1

Submission deadline: 16 January 2025

Gaussian Process regression

This project is mainly motivated from geoscience, where the modelling of unknown properties of the subsurface as Gaussian random fields is routinely used. The key idea is that an *a priori* chosen Gaussian Process is updated to a conditional distribution, by conditioning on available data.

1 Theoretical background

We start by recalling the well-known formula for the conditional distribution of a subvector $\mathbf{Y} \in \mathbb{R}^{d_y}$ of a Gaussian vector $\mathbf{X} \in \mathbb{R}^d$, conditional on given values of the remaining components $\mathbf{Z} \in \mathbb{R}^{d-d_y}$ such that $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\boldsymbol{\mu}_y, \boldsymbol{\mu}_z)$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{yz}^T & \boldsymbol{\Sigma}_{zz} \end{bmatrix}, \tag{1}$$

we have that

$$\mathbf{Y}|\mathbf{Z} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{2}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\mu}} &= \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_z) && (3) \\ \tilde{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zy}. && (4) \end{aligned}$$

1.1 Conditioned Gaussian random fields

This project focuses on the generalization of the above procedure to Gaussian processes and Gaussian random fields, known as *Gaussian Process (GP) regression* (see [3], Chapter 2). Specifically, if $\{f(\mathbf{x}),\ \mathbf{x} \in I \subset \mathbb{R}^k\}$ is a Gaussian random field with mean $m(\mathbf{x})$ and covariance function $c(\mathbf{x}, \mathbf{x}')$, we have that for any finite set of positions $Z = \{\mathbf{z}_i\}_{i=1}^{N_Z}$, the vector $\mathbf{f}(Z) = (f(\mathbf{z}_1), \ldots, f(\mathbf{z}_{N_Z}))$ follows

$$\mathbf{f}(Z) \sim \mathcal{N}(\mathbf{m}(Z), \mathbf{K}(Z)) \tag{5}$$

where $\mathbf{m}(Z) = (m(\mathbf{z}_1), \ldots, m(\mathbf{z}_{N_Z}))$ and $\mathbf{K}(Z)$ is the matrix with entries $(\mathbf{K})_{ij} = c(\mathbf{z}_i, \mathbf{z}_j)$. Then for any new set of "locations" $Y = \{\mathbf{y}_i\}_{i=1}^{N_Y}$, the conditional distribution of $\mathbf{f}(Y) = (f(\mathbf{y}_1), \ldots, f(\mathbf{y}_{N_Y}))$, given $\mathbf{f}(Z)$ is Gaussian with *posterior* mean and covariance matrix given by

$$\begin{aligned} m(Y|Z) := \mathbb{E}[\mathbf{f}(Y)|\mathbf{f}(Z)] &= \mathbf{m}(Y) + \mathbf{K}(Y,Z)\mathbf{K}^{-1}(Z)\left(\mathbf{f}(Z) - \mathbf{m}(Z)\right) \\ \mathbf{K}(Y|Z) &= \mathbf{K}(Y) - \mathbf{K}(Y,Z)\mathbf{K}(Z)^{-1}\mathbf{K}(Z,Y). \end{aligned} \tag{6}$$

In the above, $\mathbf{K}(Z,Y)$ is the $N_Z \times N_Y$ matrix with entries $\left(\mathbf{K}(Z,Y)\right)_{ij} = c(\mathbf{z}_i, \mathbf{y}_j)$, $i = 1, \ldots, N_Z$, $j = 1, \ldots, N_Y$, while trivially we have that $\mathbf{K}(Y,Z) = \mathbf{K}(Z,Y)^T$.

In real applications, it is a standard practice to assume that the "observations" $\mathbf{f}(Z)$ are contaminated with noise, that is the values that we actually collect are given by

$$\tilde{\mathbf{f}}(Z) = \mathbf{f}(Z) + \epsilon \tag{7}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, s^2\mathbf{I}_{N_Z})$ is independent Gaussian noise added to the actual values of the random field. In this case it is easy to see that the covariance of $\tilde{\mathbf{f}}(Z)$ is $\mathbf{K}(Z) + s^2\mathbf{I}$. Furthermore, one can typically argue about the choice the mean function $\mathbf{m} : I \to \mathbb{R}$ when no information is available about the function $f(\cdot)$ and a standard choice is to set the prior mean to be zero. In this case, the conditional distribution of $\mathbf{f}(Y)$ given $\tilde{\mathbf{f}}(Z)$ is still Gaussian with mean and covariance matrix given by

$$\begin{aligned} \tilde{\mathbf{m}}(Y|Z) := \mathbb{E}[\mathbf{f}(Y)|\tilde{\mathbf{f}}(Z)] &= \mathbf{K}(Y,Z)\left(\mathbf{K}(Z) + s^2\mathbf{I}\right)^{-1}\tilde{\mathbf{f}}(Z) \\ \tilde{\mathbf{K}}(Y|Z) &= \mathbf{K}(Y) - \mathbf{K}(Y,Z)\left(\mathbf{K}(Z) + s^2\mathbf{I}\right)^{-1}\mathbf{K}(Z,Y). \end{aligned} \tag{8}$$

Notice that to compute efficiently $\tilde{\mathbf{m}}(Y|Z)$ and $\tilde{\mathbf{K}}(Y|Z)$, one can "pre-factorize" e.g., by Cholesky decomposition the matrix $\mathbf{K}(Z) + s^2\mathbf{I}$.

1.2 Choice of covariance kernels

For the needs of this project we will make use of two specific models of covariance kernels that are widely used for modeling stationary random fields.

1. The **Exponential (Exp) kernel** given by

$$c(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}'\|_2}{\ell}\right\}, \tag{9}$$

   where $\|\cdot\|_2$ denotes the Euclidean norm of a vector.

2. The **Squared Exponential (SE) kernel** given by

$$c(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}\frac{(x_i - x_i')^2}{\ell_i^2}\right\}. \tag{10}$$

In the above, we will refer to σ^2 as the *variance* and to $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ as the *correlation length* or *lengthscale* of the kernel $c(\cdot, \cdot)$. The SE kernel with $\ell_1 = \cdots = \ell_k$ is called *isotropic* kernel. Note that the Exponential kernel is always isotropic.

1.3 Choosing the parameters

The parameters $\{\sigma^2, \boldsymbol{\ell}, s^2\}$ that appear in the expressions of the covariance kernels affect the performance of the GP regression and are in general problem specific. One way of choosing them, when no prior knowledge is available about their values is to maximize the marginal likelihood

$$p_{\tilde{\mathbf{f}}(Z)}(v) = \int_{\mathbb{R}^{N_z}} p_{\tilde{\mathbf{f}}(Z)|\mathbf{f}(Z)}(v|\mathbf{f}) p_{\mathbf{f}(Z)}(\mathbf{f}) \mathrm{d}\mathbf{f}, \tag{11}$$

where $p_{\tilde{\mathbf{f}}(Z)|\mathbf{f}(Z)}(\cdot|\mathbf{f}) = \mathcal{N}(\mathbf{f}, s^2\mathbf{I})$ and $p_{\mathbf{f}(Z)}(\cdot) = \mathcal{N}(\mathbf{0}, \mathbf{K}(Z))$. From this, one can compute the integral and obtain the marginal log-likelihood

$$\log p_{\tilde{\mathbf{f}}(Z)}(v) = -\frac{1}{2} v^T \left(\mathbf{K}(Z) + s^2\mathbf{I}\right)^{-1} v - \frac{1}{2}\log|\det(\mathbf{K}(Z) + s^2\mathbf{I})| - \frac{N_Z}{2}\log 2\pi, \tag{12}$$

Then, the parameters $\{\sigma^2, \boldsymbol{\ell}, s^2\}$ are determined as

$$\underset{\sigma^2, \boldsymbol{\ell}, s^2}{\arg\min} \left(-\log p_{\tilde{\mathbf{f}}(Z)}(w)\right),$$

where $w = \tilde{\mathbf{f}}(Z)$ is the vector of observed values of the Gaussian process in the points $Z = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_z})$. Note that in the absence of noise in the observations, that is $s^2 = 0$, we have that $\tilde{\mathbf{f}}(Z) = \mathbf{f}(Z)$ and (12) reduces to $\log p_{\mathbf{f}(Z)}(v)$.

2 Goals of the project

2.1 Preliminaries

1. Prove (8) and (12).
2. Compute the gradient of $-\log p_{\tilde{\mathbf{f}}(Z)}(v)$ with respect to the parameters $\{\sigma^2, \boldsymbol{\ell}, s^2\}$. *Hint*: you can find formulas to differentiate the inverse function X^{-1} and $\log\det(X)$ in Examples 4.24 and 4.28 of [1], respectively.
3. Is $-\log p_{\tilde{\mathbf{f}}(Z)}(v)$ convex in s^2 and σ^2?

2.2 Recovering a simple function

The first goal of this project is to recover a given function $f(x) = \sin(x)$ from random noise-free observations using the GP regression technique described in the previous section.

1. Choose randomly N_Z locations $Z = \{z_i\}_{i=1}^{N_Z}$ within the interval $[0, 2\pi]$ (by drawing from a uniform distribution) and generate noise-free data $\tilde{\mathbf{f}}(Z) = \sin(Z)$. Fit a Gaussian Process model on the data as described in the previous section, i.e., compute the updated mean and point-wise variance at a relatively fine uniform partition of the interval $[0, 2\pi]$ ($N_Y = 1000$ points). In such model, you can consider fixed values of the parameters $\{\sigma^2, \ell, s^2\}$. Take in particular, $\sigma^2 = 0.1$, $s^2 = 0.01$ and repeat the GP-regression for both types of covariance kernels (Exponential and SE) for $\ell = 0.1, 0.5, 1$, and two sample sizes $N_Z = 10, 100$. Generate separate plots of the data for each possible case. Add to each plot the conditional mean and the confidence intervals corresponding to ± 2 standard deviations, for each possible case (of different kernel, length scale and sample size). Draw also 3 independent realizations from the conditioned Gaussian process. Note that although the generated data do not have noise, we still take $s^2 > 0$. This has stabilizing effects. Comment on the effect of the correlation length on the quality of the posterior means and the corresponding confidence regions.
2. We now consider the estimation of the parameters $\{\sigma^2, \ell, s^2\}$ from the data. Compute analytically the gradients of $\log p_{\tilde{\mathbf{f}}}$ with respect to the parameters $\boldsymbol{\theta} = (\sigma^2, \ell, s^2)$ and write a script that minimizes the negative marginal log-likeligood $-\log p_{\tilde{\mathbf{f}}}$ with respect to $\boldsymbol{\theta}$ by employing some gradient-based optimization algorithm from `scipy.optimize` (L-BFGS-B or basinhopping are recommended). Report your estimated optimal parameters $\boldsymbol{\theta}$ and generate plots as in 1. Does the Maximum Likelihood Estimate (MLE) eventually suggest that your data is noise-free ?

Remark: To ensure reproducibility of the results and for easy evaluation of your solutions, you are encouraged to fix the seed of your random number generator at the beginning of your scripts, to `np.random.seed(12345)`. This way, the randomly generated locations Z will be the same each time your rerun the script.

2.3 GP regression on a permeability field

Geoscientists commonly face the problem of inferring the unknown properties of the subsurface based on limited data that becomes available through boreholes. Here the task is to compare the performance of GP regression on using available observations from the permeability field shown in Fig. 1.

1. The reference field given as a 110×60 `numpy array` object, can be found under the name `true_perm.npy`. Load the field and extract the values at all locations defined by the cartesian product of:

Data set 1 `10, 20, 30, 40, 50` (x axis) and `15, 35, 55, 75, 95` (y axis).

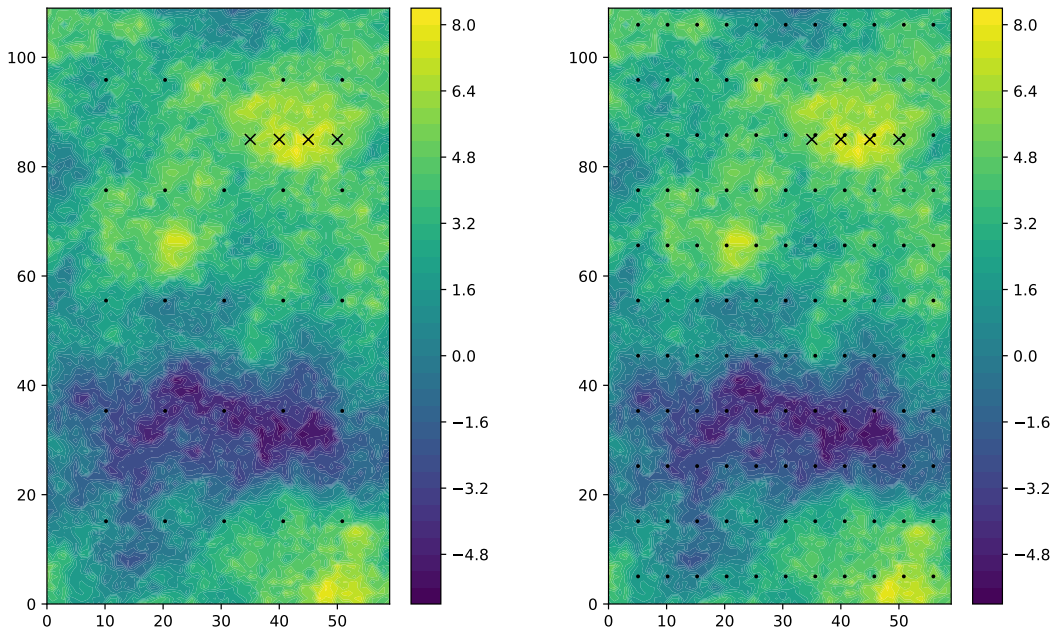

Figure 1: Reference permeability with 25 (left) or 121 (right) locations from where data is extracted.

Data set 2 `5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55` (x axis) and `5, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105` (y axis).

2. Provide plots of the predicted mean and separate plot of the variance after optimizing the parameters $\boldsymbol{\theta}$ for both cases of covariance kernels (Exp, SE). Compare the performance of the two covariance kernels in their ability to predict the true permeability field.

3. Generate and visualize some realizations of the two-dimensional conditional Gaussian random fields for both covariance kernels (Exp, SE) and with the optimized parameters obtained at the previous point on a fine grid $\hat{\Omega}_{257\times257}$ of 257×257 points. For this, utilize the two-dimensional circulant embedding algorithm described in [2, Section 3.3.2]

4. Estimate $\mathbb{P}(\max_{\mathbf{x}\in\hat{\Omega}_{257\times257}} f(\mathbf{x}) \geq 8)$. Devise a standard Monte Carlo estimator to compute such quantity with controlled accuracy.

5. Propose then a possible variance reduction technique to improve the standard MC estimator, and assess its performance.

References

[1] Nicolas Boumal, *An introduction to optimization on smooth manifolds*, Cambridge University Press, 2023.

[2] Andreas Van Barel, *Multilevel monte carlo methods for robust optimization of partial differential equations*, (2021).

[3] Christopher KI Williams and Carl Edward Rasmussen, *Gaussian processes for machine learning*, vol. 2, MIT press Cambridge, MA, 2006.