Mattias Villani
Department of Statistics
Stockholm University

Computer Lab 1 - GP Regression and Classification

The labs are the only examination, so you should do the labs **individually**.
You can use any programming language you prefer, but do **submit the code**.
Submit a readable report in:
- **PDF** (no Word documents!)
- **JuPyteR/Quarto notebook compiled to PDF/HTML**

---

1. *Homoscedastic GP regression*

   (a) Consider the following GP regression

   $$\begin{aligned} y_i &= f(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right) \\ f(\boldsymbol{x}) &\sim \text{GP}\left(0, k(\boldsymbol{x}, \boldsymbol{x}')\right) \end{aligned}$$

   with a squared exponential kernel

   $$k(x, x') = \sigma_f^2 \cdot \exp\left(-\frac{(x-x')^2}{2\ell^2}\right).$$

   Write your own code to compute the posterior distribution of $\boldsymbol{f}_*$

   $$p(\boldsymbol{f}_* | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*)$$

   where $\boldsymbol{y}$ $(n \times 1)$ is the training response data and $\mathbf{X}$ $(n \times p)$ is the training covariate data for p covariates, $\boldsymbol{X}_*$ $(n_* \times p)$ is a matrix of test inputs and $\boldsymbol{f}_*$ $(n_* \times 1)$ is the corresponding function values at the test points. Let the kernel hyperparameters σ_f and ℓ, and the noise standard deviation σ_ε be inputs to your posterior function.

   (b) Use your code to analyze the `Lidar` data (available on the course web page) with the `Distance` variable as the only covariate/feature in both the mean and variance. Set $\ell = 1$, $\sigma_f = 0.5$ and $\sigma_\varepsilon = 0.05$ (but also play around to learn!). In particular, do a scatterplot of the data and overlay:

   - the posterior mean of $f(\cdot)$ (computed over a suitable grid)
   - 95% credible bands for $f(\cdot)$
   - 95% predictive bands for y.

2. *Heteroscedastic GP regression*

(a) Consider the following heteroscedastic GP regression

$$\begin{aligned} y_i &= f(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2_{\varepsilon,i}\right) \\ f(\boldsymbol{x}) &\sim \mathrm{GP}\left(0, k(\boldsymbol{x}, \boldsymbol{x}')\right) \\ \log \sigma^2_{\varepsilon,i} &= w_0 + \boldsymbol{w}_1^T \boldsymbol{x}_i \end{aligned}$$

Implement an algorithm that samples from the joint posterior

$$p(\boldsymbol{f}_*, w_0, \boldsymbol{w}_1 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*) = p(\boldsymbol{f}_* | w_0, \boldsymbol{w}_1, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*) p(w_0, \boldsymbol{w}_1 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*).$$

Use the prior $(w_0, \boldsymbol{w}_1)^T \sim N(0, \tau^2 I)$, independent of f.
Hint: $p(\boldsymbol{f}_* | w_0, \boldsymbol{w}_1, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*)$ is really close to the formulas (2.22) to (2.24) in the GPML book, and to what you coded in Problem 1 above. The distribution

$$p(w_0, \boldsymbol{w}_1 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*)$$

is available in closed form and its expression is close to an expression in the GPML book (if you think a little ...). However, $p(w_0, \boldsymbol{w}_1 | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X}_*)$ is not a standard (known) distribution, and you need to use use MCMC or HMC, or something else, to simulate from it. It is OK to use a package for MCMC/HMC.

(b) Re-analyze the Lidar data using the heteroscedastic GP regression. Plot the marginal posterior distributions for w_0 and w_1. Do a similar plot as the one in 1(b) using the heteroscedastic model.

HAVE FUN!