ICMU Mini-course on Gaussian Processes

Homework 3: Inference for Gaussian Processes

Please upload your completed homework using [this link] by the end of the day on Wednesday, February 18. Please [email me] if you have questions or would like to request an extension. The document name should include your last name. Scans or clear photographs of hand-written answers are acceptable if they are uploaded as a single document. If you upload multiple documents with the same name (e.g., if you wish to update a previously submitted document), we will grade the latest version.

1. The covariance of a Gaussian process (GP) posterior at inputs $t^*, s^* \in D \subset \mathbb{R}$ is given by,

$$C_n(t^*, s^*) = C_0(t^*, s^*) - C_0(t^*, t)[C_0(t, t) + \sigma^2 I_{n\times n}]^{-1} C_0(t, s^*),$$

where $t = (t_1, \ldots, t_n)^\top$ are observation locations.

   (a) Explain why the second term in this equation is often referred to as the "information gain" from the data.

   (b) Show that the posterior variance $C_n(t^*, t^*)$ is always less than or equal to the prior variance $C_0(t^*, t^*)$. Under what condition does the posterior variance equal the prior variance?

2. This question compares two approaches to fitting a Gaussian process regression model: *batch updating* where we condition on the full data simultaneously, and *sequential updating* where we condition on disjoint sets of the data in a sequence.

Suppose that we observe two disjoint sets of data $y^{(i)} \in \mathbb{R}^{n_i}$ measured at $t^{(i)} \in [0, 1]^{n_i}$ for $i = 1, 2$, which are realizations of

$$Y^{(i)} = X(t^{(i)}) + \epsilon^{(i)}, \quad \epsilon^{(i)} \overset{ind}{\sim} N_{n_i}(0, \sigma^2 I_{n_i \times n_i}),$$

where X is an unknown function. The analyst chooses the following prior model for X,

$$X \sim GP\left(0, C_0(t, s)\right).$$

Show that the distribution of $X(t^*) \mid Y_1 = y_1, Y_2 = y_2$, obtained by processing all data simultaneously (batch updating) is identical to the distribution obtained by first finding the distribution of $X(t) \mid Y_1 = y_1$ and then using this as a prior in a subsequent update performed by conditioning on $Y_2 = y_2$. *Hint: you may use the block matrix inversion formula for one of the steps.*

3. Consider the R script `GP_regression_temperature_Ukraine.R` from the course GitHub directory. The prior model for the temperature field, X, over Ukraine is

$$X \sim GP\left(0, C_0(t, s)\right),$$

where the prior covariance is the square exponential,

$$C_0(t, s) = \alpha_0^2 \exp\left\{-\frac{||t - s||_2^2}{2\ell^2}\right\},$$

and the *prior variance* factor α_0^2 controls the scale of the posterior variance at a given input location. A large prior variance leads to a more variable GP prior. The settings in the R script are $\ell = 2.0$, $\alpha_0^2 = 10.0$, and $\sigma^2 = 0.5$ (note that, in practice, we would also estimate the error variance rather than holding it fixed at an arbitrary value).

(a) Decrease the length-scale ℓ to 0.5 and then increase it to 10.0. What do you notice about the posterior mean and variance? Explain the differences by relating it to role of the length-scale in defining the strength of the correlation between the process at nearby input locations.

(b) Set the error variance σ^2 to a very small value (e.g., 0.0001) and then a large value (e.g., 5.0) and describe the results. Note that when $\sigma^2 = 0$, GP regression interpolates the data.