2 Exercises, week 2

The exercise solutions have to be returned at the latest on Sunday March 26'th.

- Pen and paper exercises: you can scan the solutions and combile them into pdf or write them with Latex/word/... Compile all the answers into one pdf file
- Computer exercises: Report the answers to no-coding parts of exercises (if any) in pdf and compile with answers to pen and paper exercises. Additionally, send also the code used to solve the exercises. Note!
  - Only code should be returned. **Do not send data files**!
  - Write and comment the code so that it can be run by using your code only and the data provided in the course web pages.
  - If the lecturer cannot understand or run your code you will not get points from coding part even if the results were correct.
- zip all files into one folder to reduce the size of submission.

  Send the zipped files to jarno.vanhatalo@helsinki.fi.

For basic properties and results concerning Gaussian distributions and processes see e.g.
`https://en.wikipedia.org/wiki/Multivariate_normal_distribution`
`http://www.gaussianprocess.org/gpml/chapters/`

2.1 On posterior predictive distributions

a) Consider the case where we have direct observations from a Gaussian process. Show that posterior predictive distribution reduces to the observations if it is calculated at observation locations. Show also that this does not happen in a case where we make noisy observations $y(\mathbf{s}) = f(\mathbf{s}) + \epsilon(\mathbf{s})$ where $\epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon^2)$ are i.i.d. noise terms.

b) Consider we make observations $\mathbf{y} = [y_1, ..., y_n]^\mathsf{T}$ and model them with $y_i = f_+\epsilon_i$ where $\mathbf{f} = [f_1, ..., f_n]$ is a vector of latent variables with a multivariate Gaussian prior distribution, $\mathbf{f} \sim N(0, \mathbf{K}_{\mathrm{f,f}})$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ are i.i.d. noise terms. We can now write this as a hierarchical model

$$p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} N(y_i | f_i, \sigma_\epsilon^2)$$
$$p(\mathbf{f}) = N(\mathbf{f} \,| 0, \mathbf{K}_{\mathrm{f,f}}).$$

Derive the posterior distribution of the latent variables by using the Bayes theorem $p(\mathbf{f} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f})}{p(\mathbf{y})}$.

c) Consider a GP with zero mean and any covariance function of the form $k(\mathrm{s}, \mathrm{s}') = k(||\,\mathrm{s} - \mathrm{s}'\,||)$ (for example, the exponential, Matern class of covariance functions and the squared exponential covariance functions). What are the posterior predictive mean and variance of $f(\mathrm{s})$ as $||\,\mathrm{s} - \mathrm{s}'\,|| \to \infty$, where $\mathrm{s}' \in \mathrm{S}$ is the collection of data locations.

d) Assume $y(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\beta + \epsilon$ where $\mathbf{x}$ is an $p \times 1$ vector of covariates and $\beta \sim N(0, \Sigma)$ a $p \times 1$ vector of linear weights and $\epsilon(\mathbf{s}) \sim N(0, \sigma_\epsilon^2)$ are i.i.d. noise terms. Given a set of observations $\mathbf{y} = [y(\mathbf{x}_1), ..., y(\mathbf{x}_n)]^{\mathsf{T}}$, derive the equations for posterior predictive mean and variance of $f(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\beta$. Derive also the equations for the posterior distribution of β and what are its posterior mean and covariance? Consider next $p = 1$, how do the posterior predictive mean and variance of $f(x)$ behave when the prediction points are close to training data or far from it?

2.2 Exponential covariance function and AR(1) model

Consider an autoregressive time series model of order one (AR(1) model)

$$X_t = \phi X_{t-1} + \epsilon_t \tag{6}$$

where $X_t \in \Re$ is a random variable and the error terms are independent and indentically distributed $\epsilon_t \sim N(0, \sigma^2)$ and $X_{t_0} = 0$. Show that, with some range for ϕ this model induces an exponential covariance function between the random variables, that is $Cov[X_t, X_{t-k}] = \sigma_e^2 e^{-k/l}$ where σ_e^2 and l can be represented by ϕ and σ^2.

2.3 More covariances and means

a) Show that, if $k_1(\mathrm{s}, \mathrm{s}')$ and $k_2(\mathrm{s}, \mathrm{s}')$ are valid covariance functions, so are $k_1(\mathrm{s}, \mathrm{s}') + k_2(\mathrm{s}, \mathrm{s}')$ and $k_1(\mathrm{s}, \mathrm{s}')k_2(\mathrm{s}, \mathrm{s}')$.

b) Suppose $f(\mathrm{s})$ is a Gaussian process with mean surface $\mu(\mathrm{s})$ and covariance function $k(\mathrm{s}, \mathrm{s}')$. Let $z(\mathrm{s})$ be the induced log Gaussian process, i.e., $f(\mathrm{s}) = log z(\mathrm{s})$. Find the mean surface and the covariance function for the $z(\mathrm{s})$ process. If $f(\mathrm{s})$ is stationary, is $z(\mathrm{s})$ necessarily stationary?

2.4 Computer: Simulations from a Gaussian process

Let $f(\mathrm{s})$ be a Gaussian process with an exponential covariance function

$$Cov\,(f(\mathrm{s}), f(\mathrm{s}')) = k(\mathrm{s}, \mathrm{s}') = \sigma^2 e^{-||\,\mathrm{s} - \mathrm{s}'\,||/l}, \tag{7}$$

where $||\,\mathrm{s} - \mathrm{s}'\,||$ is the euclidean distance between locations s and s′, σ^2 the process variance and l the length-scale.
a) What is the range after which the correlation has dropped to 5% of its maximum?

b) Write an R/Matlab function to construct the covariance matrix between function values at arbitrary sets of locations. That is, let $\mathrm{S} = \{\mathrm{s}_1, ..., \mathrm{s}_{n_1}\}$ and $\mathrm{S}' = \{\mathrm{s}'_1, ..., \mathrm{s}'_{n_2}\}$ be collections of spatial locations. The function should construct a matrix $\mathbf{K}$ so that $\mathbf{K}_{i,j} = k(\mathbf{s}_i, \mathbf{s}'_j)$ for any $\mathbf{S}$, $\mathbf{S}'$, l and σ^2. Use the function to

- plot the exponential covariance as a function of distance
- plot the variogram related to the exponential covariance. The variogram of a (stationary and isotropic) covariance function is $\gamma(h) = k(0) - k(h)$, where $k(h) = k(s + h, s)$ is the covariance function written as a function of distance, h, between two locations, s and $s + h$.
- draw realizations of functions from a GP with a zero mean and an exponential covariance function from range $\mathbf{s} \in [0, 10]$ (that is $f(\mathbf{s}) : \Re \rightarrow \Re$) by using the results from exercise 1.5 and plot them.
- draw realizations of functions from a GP with a zero mean and an exponential covariance function from range $\mathbf{s} \in [0, 10] \times [0, 10]$ (that is $f(\mathbf{s}) : \Re^2 \rightarrow \Re$) and plot them either as contour plots, 3D surfaces or colored images.

Test $l = 2$ and $\sigma^2 = 1$ and at least two other combinations of parameters.

c) Repeat the steps a) and b) for a squared exponential covariance function

$$k(\mathbf{s}, \mathbf{s}') = \sigma^2 e^{-||\,\mathbf{s}-\mathbf{s}'\,||^2/l^2}. \tag{8}$$

What are the differences between the two models.

Hints. To draw realizations from a GP in 1D ($f(\mathbf{s}) : \Re \rightarrow \Re$) divide the input space into, e.g. 100 equally spaced intervals and use the ends of these intervals as the set of locations, $\tilde{\mathbf{S}} = [\tilde{s}_1, ..., \tilde{s}_n]^\mathsf{T}$, where GP is simulated at. Calculate the covariance matrix between function values at the $\mathbf{s} \in \mathbf{S}$, use this covariance matrix to draw the realizations and then plot them with respect to $\mathbf{S}$. In 2D do otherwise similarly but divide the input space into a lattice grid of, e.g. 50×50 grid cells (see exercise template for an example how to construct a lattice grid) . Squared exponential covariance function often leads to numerically unstable Cholesky decomposition. Hence, if Cholesky decomposition does not remain positive definite add jitter to it (see lecture notes).

2.5 Computer: Prediction with a Gaussian process

a) Assume, we have observed $\mathbf{f} = [1, -1, 0, 2]^\mathsf{T}$ at $\mathbf{S} = [0.7, 1.3, 2.4, 3.9]^\mathsf{T}$ from the Gaussian processes defined in exercise 2.4. Write a Matlab/R function that calculates the marginal posterior mean, $E[f(s)|\,\mathbf{f}, \mathbf{S}]$, and variance, $Var[f(s)|\,\mathbf{f}, \mathbf{S}]$, at any location $s \in \Re$. Use the code to visualize the posterior mean, 95% central credible interval of $f(s)$ and random draws from $p(f(s)|\,\mathbf{f}, \mathbf{S})$ in the range $s \in [0, 5]$.

b) Continue from exercise 1.2. You need the raster maps, polygons and nutrient consentration data files:

- GoFgrids2000.csv (raster maps from the Gulf of Finland)

- GoFpolygon.txt (polygon coordinates to plot the shore line of the GoF)
- GoFnutrients_2000_2004.csv (a data file with measurements of nutrients in the GoF from 2001-2004)

Build a Gaussian process model with additive Gaussian noise for the average winter nitrogen concentration. That is, let $f(\mathbf{s}_i)$ denote the average winter nitrogen consentration at location $\mathbf{s}_i$ and let

$$y_i = f(\mathbf{s}_i) + \epsilon_i, \tag{9}$$

be the measurement with i.i.d. noise $\epsilon_i \sim N(0, \sigma^2)$. Give a Gaussian process prior for the nitrogen concentration, that is $f(s) \sim GP(0, k(\mathbf{s}, \mathbf{s}'))$, where $k(\mathbf{s}, \mathbf{s}')$ is either the squared exponential or exponential covariance function. Solve the posterior distribution for $f(s)$ and visualize its posterior mean and variance over the whole GoF. Test different combinations values for the noise variance, length-scale and magnitude. Report the values with $\sigma^2 = 0.1$, $l = 20$km and $\sigma^2_{\text{cf}} = 1$, where σ^2_{cf} is the maginitude of the covariance function. Note! before modeling it is good to "standardize" the observations so that $y_i = (y_i - \bar{y})/\text{sd}(y)$, where $\bar{y}$ and $\text{sd}(y)$ are the sample mean and standard deviation of observations.

Hints. You have calculated the winter average nitrogen consentrations in the exercise 1.2. GoFgrids2000.csv provides you the prediction grid similar to the grid you used in exercise 2.4. Hence, you need to repeat the 2D task of exercise 2.4 with this real data. After predicting to the grid in GoFgrids2000.csv you can visualize the map as you visualized, e.g., the depth layer in exercise 1.2.