3 Exercises, week 3

The exercise solutions have to be returned at the latest on Sunday April 2'nd.

- 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/`

3.1 Hyper-parameter inference and spatial prediction, 4 points

We continue from the exercise 2.5 b). 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)

Consider 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

$$y_i = f(\mathbf{s}_i) + \epsilon_i, \tag{10}$$

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. We give hyperpriors for the hyperparameters σ^2, l and σ^2_{cf}, where σ^2_{cf} is the maginitude of the covariance function.

You are provided with R template `exerciseTemplate_week_1.R` and few STAN model files. You need also either the model solutios or your own solutions to the exercise 2.5. Your task is to fill in the missing parts of the exercise template and to answer to the following questions

a) test different priors for the hyperparameters. Compare the priors provided in the .stan file with at least very flat priors and very informative priors. You may choose the priors yourself and you need to report whether the MAP and MCMC (see part d) solutions are sensitive to the priors. If they are, report how.

b) Calculate the posterior predictive distribution of the average nitrogen concentration in each of the water framework directive region using the MAP estimate of the hyperparameters so that you calculate the average over the lattice grid cells that fall into that region. for exapmle, the average nitrogen concentration in area 1 is $N_{A_1} = \frac{1}{M_1}\sum_{i=1}^{M_1} f(\tilde{\mathbf{s}}_i)$ where $\tilde{\mathbf{s}}_i \in A_1$. Visualize the distribution. What is the probability that the average concentration is less than 28 micro.mol / l?

c) Calculate the posterior predictive distribution of the *difference* in the average nitrogen concentrations *between* the water framework directive *region 1 and the other regios*. Use the MAP estimate of the hyperparameters. Visualize the distributions. What is the probability that the average concentration is less in region 1 than in the other regions?

d) Do the MCMC sampling for the hyperparameters, check for convergence, calculate PSRF and effective number of samples. After this marginalize over the posterior of the hyperparameters and calculate the posterior predictive mean and variance of nitrogen concentration at each grid cell using the law of total variance

$$E\left(f(\mathbf{s})\right) = E\left(E\left(f(\mathbf{s})|\theta\right)\right) \tag{11}$$
$$Var\left(f(\mathbf{s})\right) = E\left(Var(f(\mathbf{s})|\theta)\right) + Var\left(E(f(\mathbf{s})|\theta)\right). \tag{12}$$

Redo calculations of part b) but now so that you marginalize over the posterior of the hyperparameters (you can utilize the above laws in this as well). Compare your results to the results that you got with the MAP estimates for the hyperparameters. **Note!** Even if you sample, e.g., 1000 samples with MCMC, it is enough to use about 100 to calculate the predictive summaries.

Hints. You may want to consult `http://mc-stan.org/documentation/`. When calculating areal averages remember that the latent function values at different lattice grid cells are correlated. Remember also that a (weighted) sum of Gaussian random variables is also Gaussian. Hence, $N_{A_i} = \mathbf{w}_i^\mathsf{T}\tilde{\mathbf{f}}_i \sim N(\cdot,\cdot)$, where $\mathbf{w}_i = [1/M_i, \ldots, 1/M_i]^\mathsf{T}$ is a length M_i vector of weights and $\tilde{\mathbf{f}}_i$ is a vector of latent variables at the grid cells of region i. Similarly, the difference between two averages can be written as $N_{A_i} - N_{A_j} = \mathbf{w}_i^\mathsf{T}\tilde{\mathbf{f}}_i - \mathbf{w}_j^\mathsf{T}\tilde{\mathbf{f}}_j = \mathbf{w}_{ij}\tilde{\mathbf{f}}_{ij} \sim N(\cdot,\cdot)$ where $\mathbf{w}_{ij} = [\mathbf{w}_i^\mathsf{T}, -\mathbf{w}_j^\mathsf{T}]^\mathsf{T}$ and $\tilde{\mathbf{f}}_{ij} = [\tilde{\mathbf{f}}_i^\mathsf{T}, \tilde{\mathbf{f}}_j^\mathsf{T}]^\mathsf{T}$. Additional hints are provided in the exercise template.

Note also that this exercise requires care in coding. The asked quantities can be very time consuming to calculate if you calculate them naively. For this reason you are provided hints

in the exercise template. One specific thing to avoid is the construction of large covariance matrices. For example, when calculating the variance of $f(\mathbf{s})$ at lattice grid cells you should not construct the full covariance matrix between all the latent variables. Similarly, when calculating the areal means use only those lattice cells that you really need.

3.2 The rule of total variance

The rule of total variance is handy, if you are able to analytically solve the conditional variances and conditional means. Hence, given a result $E\left(f(\mathbf{s})\right) = E\left(E\left(f(\mathbf{s})|\theta\right)\right)$ show that $Var\left(f(\mathbf{s})\right) = E\left(Var(f(\mathbf{s})|\theta)\right) + Var\left(E(f(\mathbf{s})|\theta)\right)$.

Hints. Start with the equation for the marginal variance $Var[f(\mathbf{s})] = E[\mathbf{f}(\mathbf{s})^2] - E[\mathbf{f}(\mathbf{s})]^2$, solve $E(\mathbf{f}(\mathbf{s})^2|\theta)$, plug it into the marginal variance and simplify.