1 Exercises, week 1

The exercise solutions have to be returned at the latest on Sunday March 19'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/`

1.1 Map projections

a) Search for information about map projections and their reference systems. Include in your search at least the following map projections: WGS84, EUREF-FIN and ETRS89, UTM, GDA94 / Geoscience Australia Lambert conformal projection and the Finnish KKJ coordinate system (Kartastokoordinaattij$\acute{}$arjestelm$\acute{}$a). Which projections are area preserving and which are conformal (locally shape-preserving)? Which are useful for mapping the whole globe and which are good for mapping restricted areas on the globe? In which areas of the globe do they work well.

b) What is the distance in meters between two locations that are 1°longitude apart in WGS84 along a constant latitude of 0°, 66°? What is the distance in meters between 0°and 15°latitude (WGS84) along a constant longitude? What is the distance in meters between 60°and 75°(WGS84) along a constant longitude? Why the two last distances do not match?

c) Consider you would need to conduct spatial modeling in Finland, in the state of Colorado in the USA and in the Great Barrier Reef in Australia. Which map projection would you choose and why?

1.2 Computer: Introduction to useful R tools

Download from course web page the files.zip which contains the following 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)
- exercisesTemplate_week1.R (a template for R scripts to solve the exercises of week 1)
- README.txt general information about the files

Examine the raster maps, polygon and data files. Follow the instructions in the exercisesTemplate_week1.R.

The purpose of this exercise is to introduce few useful R tools for reading and visualization of spatial data. The three types of data considered are raster maps (typically used to extract covariate information), polygons of geographic regions and point wise measurement data (that is used in the inference). We will return to these data several times during the course.

1.3 Marginal and conditional distributions of a multivariate Gaussian distribution

Let's denote $\theta = [\theta_1, \theta_2]^\mathsf{T}$ where θ_1 is a $p \times 1$ vector and θ_2 is a $q \times 1$ vector. Let θ_1 and θ_2 be jointly Gaussian random vectors so that

$$\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}K_{11} & K_{12} \\ K_{21} & K_{22}\end{bmatrix}\right). \tag{1}$$

where μ_1 and μ_2 are the corresponding means and $K = \begin{bmatrix}K_{11} & K_{12} \\ K_{21} & K_{22}\end{bmatrix}$ is the covariance matrix. Show that the marginal and conditional distributions of θ_1 are

$$\theta_1 \sim N\left(\mu_1, K_{11}\right) \tag{2}$$

$$\theta_1|\theta_2 \sim N\left(\mu_1 + K_{12}K_{22}^{-1}(\theta_2 - \mu_2), K_{11} - K_{12}K_{22}^{-1}K_{21}\right). \tag{3}$$

For simplicity you can assume that $\mu_1 = \mu_2 = 0$. How does this result relate to Gaussian processes?

Hint! The proof can be done with "brute force". Write down the probability density of $p(\theta_1, \theta_2)$ explicitly. Note, that you can invert the partitioned covariance matrix to obtain

$$K^{-1} = \tilde{K} = \begin{bmatrix}\tilde{K}_{11} & \tilde{K}_{12} \\ \tilde{K}_{21} & \tilde{K}_{22}\end{bmatrix} \tag{4}$$

where the elements of $\tilde{K}$ can be solved from the elements of K (Rasmussen and Williams, 2006; Appendix A). Open the matrix algebra and rearrange the terms in order to write $p(\theta_1, \theta_2) = f(\theta_1)f(\theta_1, \theta_2)$. After this you can use $p(\theta_1) = \int f(\theta_1)f(\theta_1, \theta_2)d\theta_2$ and $p(\theta_1|\theta_2) = p(\theta_1, \theta_2)/p(\theta_2)$.

1.4 Multivariate Gaussian distribution as a linear combination of univariate i.i.d Gaussian variables

a) Show that a sum of two independent Gaussian random variables is also Gaussian.

b) Let $\theta = [\theta_1, ..., \theta_n]^\mathsf{T}$ so that each θ_i has a univariate standard Gaussian distribution, $\theta_i \sim N(0, 1)$ (independently). Furthermore, let $\mu = [\mu_1, ..., \mu_n]^\mathsf{T}$ be a vector of constants and $LL^\mathsf{T} = K$ be a Cholesky decomposition of a covariance matrix K. Show, that $z = \mu + L\theta$ has a multivariate Gaussian distribution with mean μ and covariance K, that is $z \sim N(\mu, K)$.

Hint. There are several ways to solve these problems. One option to solve b) is the following. Since the dimensions of θ and z are the same, you may use results concerning the dimension preserving transformation of random variables. Let $p_u(u)$ be the probability density of the vector u. We transform to $v = f(u)$ where v has the same number of components as u. If $p_u(u)$ is a continuous distribution and $v = f(u)$ is a one-to-one transformation, then the joint density of the transformed vector is

$$p_v(v) = |J| p_u(f^{-1}(v))$$

where $|J|$ is the determinant of of the Jacobian of the transformation $u = f^{-1}(v)$. Remember also that $|AB| = |A||B|$ and $|A^{-1}| = 1/|A|$.

1.5 Computer: Sampling from a multivariate Gaussian distribution

Using the result of exercise 1.4 sample from the following two-variate Gaussian with the aid of i.i.d. Gaussian random variables,

$$z \sim N\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1.5 & 1 \\ 1 & 3 \end{bmatrix} \right). \tag{5}$$

Plot sample points (z_1 vs. z_2) and histograms of the marginal distributions for z_1 and z_2. Solve the exact marginal distribution of z_1 and z_2 and plot them. Plot a histogram from distribution $p(z_1|z_2 \in [2.975, 3.025])$ (remember to draw enough Monte Carlo samples from the joint distribution in order to reach good approximation for the conditional distribution). Using the results from exercise 1.3 solve the conditional distribution $p(z_1|z_2 = 3)$ and plot it. Do the two last conditional distributions match?

Hint. Remember to check the format of the Cholesky decomposition. For example Matlab and R `chol` return upper triangular Cholesky decompositions. Note also that in R matrix multiplication is done with `%*%` whereas in Matlab it is done with `*`.