Spatial modelling and Bayesian inference

Lecture notes

Jarno Vanhatalo

March 10, 2017

Abstract

These are lecture notes for the course Spatial modeling and Bayesian inference. These notes are not comprehensive list of all coarse content but summarize key issues covered during the course. These notes will be updated during the course. The update history is the following:

- **10.3.2017** First version of the notes published

1 Preliminaries on spatial data problems and cartography

Spatial statistics considers analysis of spatially indexed data. Typical problems are related to *inference* and *prediction* of spatially indexed phenomena. For example, what is the temperature at a spatial location $\mathbf{s} = [s_1, s_2]^\mathsf{T}$ and how can we use temperature measurements to predict the temperature at another location $\tilde{\mathbf{s}}$. Similarly we might be interested in inferring and forecasting temporal trends in spatial phenomena, such as the temporal change of annual average temperature in Europe.

Spatial problems involve spatially indexed data and traditionally these data are classified into three types

- *Point referenced data* are measured at disjoint locations in space. That is each datum contains the information, $y(\mathbf{s})$, at location $\mathbf{s} \in D$, where D is a spatial(temporal) area of interest. For example, the temperature at a specific location on the earth.
- *Areal data* describe phenomena over areal regions. That is, a datum y_i describes, for example, the average temperature over region $A_i \subset D$
- *Point pattern data* describes the spatial presence pattern of a phenomenon. Classical example is the spatial pattern of trees in a forest. Here, each datum is a location of a tree, s_i, and the aim is to analyze the process that leads to a specific presence pattern.

In order to analyze spatial data we need a coordinate system for the area of interest. Here we consider problems on the surface of the earth. There are several coordinate systems that can be used to describe the location on the earth, the simplest one being the spherical system where the location is described by the degrees in latitude and longitude (see exercises for more examples of coordinate systems). However, often the purpose is to analyze only a subset of the earth's surface. If this subset is small enough, it is typically practical to use a map projection. There are two main reasons for this. The map projections allow easy visualization on two dimensional plane and they allow the use of Euclidean metric to measure distances between locations (see also section 3).

A map projection is a systematic representation of all or part of earth's surface on a plane. It is well known fact from topology that it is impossible to construct a distortion-free representation of a globe on a flat map. Hence, when building maps decision has to be made which aspects of the reality we want to reconstruct well and which parts of earth's surface the map should represent well. For example the map can be planned to be area or direction preserving. However, we cannot produce a map projection that is distance preserving[1]. Hence, a good projection depends on application and there are numerous projections published. The general strategy to build maps is to use an intermediate surface that can be flattened. The globe (or part of it) is projected onto this intermediate surface, *developable surface*, after which it is flattened to a plane to produce a map. The most commonly used developable surfaces are the cylinder, the cone, the plane and the sinusoidal.

2 Gaussian processes

2.1 Definition and basic properties

Consider a collection of random variables $\{f(\mathbf{s}) : \mathbf{s} \in D\}$ for some region D. We will typically assume that $D \subset \Re^2$ so that $\mathbf{s}$ is a 2×1 vector of spatial coordinates. However, any other dimension is equally possible. We can model $f(\mathbf{s})$ as a stochastic process indexed by $\mathbf{s}$. Moreover, since we are interested in modelling spatial phenomena the variables $f(\mathbf{s})$ should be pairwise dependent with strength of dependence that is specified by their location. See figure 1. We will be using Gaussian processes which can be defined as follows (e.g. Rasmussen and Williams, 2006; Banerjee et al., 2015):

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Hence, if $f(\mathbf{s})$ follows a Gaussian process, any collection of random variables $\mathbf{f} = [f_1, ..., f_n]^\mathsf{T} = [f(\mathbf{s}_1), ..., f(\mathbf{s}_n)]^\mathsf{T}$ at a set of n locations, $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_n]^\mathsf{T}$, has a multivariate Gaussian distribution

$$\mathbf{f} \sim N(\mu, \mathbf{K}_{\mathrm{f,f}}) \tag{1}$$

where μ is the $n \times 1$ mean vector and $\mathbf{K}_{\mathrm{f,f}}$ is the $n \times n$ covariance matrix. We may call a Gaussian process, $f(\mathbf{s})$ interchangeably also a *latent function* or Gaussian random field and a set of function values, $\mathbf{f}$, Gaussian random variables or *latent variables*. The

[1] for a very short introduction see e.g. `https://en.wikipedia.org/wiki/Theorema_Egregium`

rationale for this nomenclature will become clear in section 4 when we build hierarchical models.

The mean vector is formed by a mean function $\mu(\mathbf{s})$ which defines the expected value of a random variable $f(\mathbf{s})$ at any location $\mathbf{s}$. For notational simplicity we will assume $\mu(\mathbf{s}) \equiv 0$ if not otherwise stated. The covariance matrix is constructed from a covariance function, $[\mathbf{K}_{\mathrm{f,f}}]_{i,j} = k(\mathbf{s}_i, \mathbf{s}_j \,|\theta)$, which characterizes the covariances between process realizations at different locations, $Cov\,(f(\mathbf{s}_i), f(\mathbf{s}_j)) = k(\mathbf{s}_i, \mathbf{s}_j \,|\theta)$. The parameter vector θ collects all the parameters of the covariance function. Covariance function encodes prior assumptions of the latent function, such as the smoothness and scale of the variation, and can be chosen freely as long as the covariance matrices produced are symmetric and *positive semi-definite*, satisfying

$$\mathbf{v}^{\mathrm{T}}\, \mathbf{K}_{\mathrm{f,f}}\, \mathbf{v} \geq 0, \forall \mathbf{v} \in \Re^n. \tag{2}$$

An example of a covariance function is the exponential

$$k_{\exp}(\mathbf{s}_i, \mathbf{s}_j \,|\theta) = \sigma^2_{\exp} e^{(-||\,\mathbf{s}_i - \mathbf{s}_j\,||/l)}, \tag{3}$$

where $||\,\mathbf{s}_i - s_j||$ is the euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$, $\sigma^2_{\exp}$ is the process variance, and l is the length-scale, which governs how fast the correlation decreases as a function of distance. Covariance functions are discussed more in section 3 and, for example, in (Diggle and Ribeiro, 2007; Finkenstädt et al., 2007; Rasmussen and Williams, 2006).

Imagine, that we have made observations of a realization of a Gaussian process $\mathbf{f}$ at a set of locations $\mathbf{S}$ and we want to use this information to update our knowledge concerning the values of the Gaussian process at some other locations $\tilde{\mathbf{S}} = [\tilde{\mathbf{s}}_1, ..., \tilde{\mathbf{s}}_{\tilde{n}}]^{\mathsf{T}}, \tilde{\mathbf{s}}_i \in D$. This is a classical problem which is called *Kriging* in tradiotional *geostatistics*. However we will use the Bayesian terminology and call this *prediction*. Notice, prediction is here a statistical term and refers to probabilistic statement at a location from where we do not have observations. Hence, prediction does not necessarily refer to statements about future as in some other fields of science. Other way of stating the problem is that we have a latent function $f(\mathbf{s})$ for which we have given a Gaussian process prior. We have made observations of the function in finite number of locations and want to predict its value at other locations $\tilde{s}$.

By definition of a Gaussian process, the marginal distribution of any subset of latent variables, the function values at fixed input locations, can be constructed by simply taking the appropriate submatrix of the covariance and subvector of the mean. (See also exercises.) Hence, the joint prior for latent variables at observation $\mathbf{S}$ and prediction locations $\tilde{\mathbf{S}}$ is

$$\begin{bmatrix} \mathbf{f} \\ \tilde{\mathbf{f}} \end{bmatrix} |\mathbf{S}, \tilde{\mathbf{S}}, \theta \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathrm{f,f}} & \mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}} \\ \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}} & \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} \end{bmatrix}\right), \tag{4}$$

where $\mathbf{K}_{\mathrm{f,f}} = k(\mathbf{S}, \mathbf{S}|\theta)$, $\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}} = \mathbf{K}^{\mathsf{T}}_{\tilde{f},f} = k(\mathbf{S}, \tilde{\mathbf{S}}|\theta)$ and $\mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} = k(\tilde{\mathbf{S}}, \tilde{\mathbf{S}}|\theta)$. Here, the covariance function $k(\cdot,\cdot)$ denotes also vector and matrix valued functions $k(\mathbf{s}, \mathbf{S}) : \Re^d \times \Re^{d\times n} \rightarrow \Re^{1\times n}$, and $k(\mathbf{S}, \mathbf{S}) : \Re^{d\times n} \times \Re^{d\times n} \rightarrow \Re^{n\times n}$. The marginal distribution of $\tilde{\mathbf{f}}$ is $p(\tilde{\mathbf{f}}|\tilde{\mathbf{S}}, \theta) = \mathrm{N}(\tilde{\mathbf{f}}|\mathbf{0}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}})$ like the marginal distribution of $\mathbf{f}$ given in (1). This marginal is also called a *prior predictive* distribution since it is not conditioned to any observations. The conditional distribution of a set of latent variables given other set of latent variables

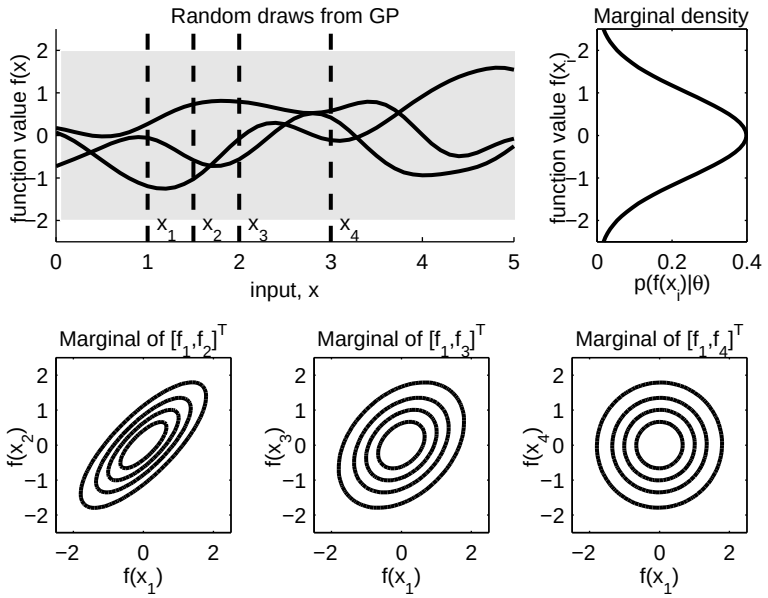

Figure 1: An illustration of a Gaussian process. The upper left figure presents three functions drawn randomly from a zero mean Gaussian process with squared exponential covariance function. The hyperparameters are $l = 1$ and $\sigma^2 = 1$ and the grey shading represents central 95% probability interval. The upper right subfigure presents the marginal distribution for a single function value. The lower subfigures present three marginal distributions between two function values at distinct input locations shown in the upper left subfigure by dashed line. It can be seen that the correlation between function values $f(\mathbf{s}_i)$ and $f(\mathbf{s}_j)$ is the greater the closer $\mathbf{s}_i$ and $\mathbf{s}_j$ are to each others.

is Gaussian as well. For example, the distribution of $\tilde{\mathbf{f}}$ given $\mathbf{f}$ is

$$\tilde{\mathbf{f}}|\,\mathbf{f}, \mathbf{X}, \tilde{\mathbf{X}}, \theta \sim \mathrm{N}(\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\,\mathbf{K}_{\mathrm{f},\mathrm{f}}^{-1}\,\mathbf{f}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\,\mathbf{K}_{\mathrm{f},\mathrm{f}}^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}), \tag{5}$$

which is called the (conditional) *posterior predictive distribution* for $\tilde{\mathbf{f}}$ after observing the function values at locations $\mathbf{S}$. Notice that the mean and covariance of the conditional (posterior predictive) distribution are functions of input vector $\tilde{\mathbf{s}}$ (through dependency in $\mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}}$, $\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}$) and the observation locations, $\mathbf{S}$ as well as the observed function values are fixed. Hence, the distribution 5 generalizes to any number of prediction locations and defines a Gaussian process with mean and covariance functions

$$m_{\mathrm{p}}(\tilde{\mathbf{s}}) = k(\tilde{\mathbf{s}}, \mathbf{S})\,\mathbf{K}_{\mathrm{f},\mathrm{f}}^{-1}\,\mathbf{f} \tag{6}$$

$$k_{\mathrm{p}}(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') = k(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') - k(\tilde{\mathbf{s}}, \mathbf{S})\,\mathbf{K}_{\mathrm{f},\mathrm{f}}^{-1}\,k(\mathbf{S}, \tilde{\mathbf{s}}'). \tag{7}$$

This can be called also the (conditional) posterior distribution of the latent function $f(\tilde{\mathbf{x}})$. We call the Gaussian process defined by (6) and (7) *conditional posterior distribution* since it is conditioned to the values of parameters θ which we will later infer along the latent variables. The conditional posterior GP is illustrated in Figure 2.

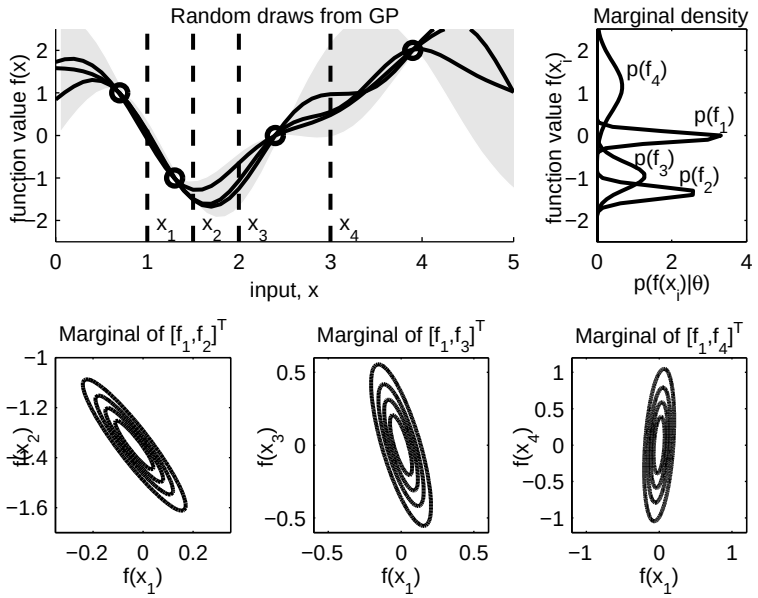

Figure 2: A conditional (posterior) GP $p(\tilde{f}|\,\mathbf{f},\theta)$. The observations $\mathbf{f} = [f(0.7) = 1, f(1.3) = -1, f(2.4) = 0, f(3.9) = 2]^{\mathrm{T}}$ are plotted with circles in the upper left subfigure and the prior GP is illustrated in the figure 1. When comparing the subfigures to the equivalent ones in Figure 1 we can see clear distinction between the marginal and the conditional GP. Here, all the function samples travel through the observations, the mean is no longer zero and the covariance is non-stationary.

2.2 Noisy observations

Typically we do not have direct observations from the Gaussian process but we use it to model the latent variables (process level) in a hierarchical Bayesian model. Possible the simplest example is a model with additive Gaussian noise

$$y(\mathbf{s}) = f(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{8}$$

where $f(\mathbf{s})$ is a Gaussian process with covariance function $k(\mathbf{s},\mathbf{s}')$ and $\epsilon(\mathbf{s})$ follows a zero mean Gaussian distribution with variance σ^2_ϵ independently at each location $\mathbf{s}$. Since the sum of two Gaussian variables is also Gaussian $y(\mathbf{s})$ follows a Gaussian process with covariance function $k(\mathbf{s},\mathbf{s}') + \delta_{\mathbf{s}}(\mathbf{s}')\sigma^2_\epsilon$, where $\delta_{\mathbf{s}}(\mathbf{s}') = 1$ if $\mathbf{s} = \mathbf{s}'$ and zero otherwise. Consider that we make now observations $\mathbf{y} = [y_1, ..., y_n]^{\mathrm{T}}$ at locations $\mathbf{S}$. In this case the (conditional) posterior predictive mean and variance of the Gaussian process are

$$m_{\mathrm{p}}(\tilde{\mathbf{s}}) = k(\tilde{\mathbf{s}},\mathbf{S})(\mathbf{K}_{\mathrm{f,f}} + \sigma^2_\epsilon\mathbf{I})^{-1}\,\mathbf{y} \tag{9}$$

$$k_{\mathrm{p}}(\tilde{\mathbf{s}},\tilde{\mathbf{s}}') = k(\tilde{\mathbf{s}},\tilde{\mathbf{s}}') - k(\tilde{\mathbf{s}},\mathbf{S})(\mathbf{K}_{\mathrm{f,f}} + \sigma^2_\epsilon\mathbf{I})^{-1}k(\mathbf{S},\tilde{\mathbf{s}}'). \tag{10}$$

See also exercises.

2.3 Additive Gaussian processes

More generally, let $f(\mathbf{s}) = h(\mathbf{s}) + g(\mathbf{s})$, where $h(\mathbf{s})$ and $g(\mathbf{s})$ are mutually independent Gaussian processes with covariance functions $k_h(\mathbf{s}, \mathbf{s}')$ and $k_g(\mathbf{s}, \mathbf{s}')$. Then, $f(\mathbf{s})$ follows a Gaussian process with covariance function $k_h(\mathbf{s}, \mathbf{s}') + k_g(\mathbf{s}, \mathbf{s}')$. Consider now that we have made observations of $f(\mathbf{s})$ at locations $\mathbf{S}$. Then the (conditional) posterior distribution of for example $h(\mathbf{s})$ is a Gaussian process with mean and covariance functions

$$m_{h|\,\mathbf{f}}(\tilde{\mathbf{s}}) = k_h(\tilde{\mathbf{s}}, \mathbf{S})(\mathbf{K}_{\mathrm{g,g}} + \mathbf{K}_{\mathrm{h,h}})^{-1}\,\mathbf{y} \tag{11}$$

$$k_{h|\,\mathbf{f}}(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') = k_h(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') - k_h(\tilde{\mathbf{s}}, \mathbf{S})(\mathbf{K}_{\mathrm{g,g}} + \mathbf{K}_{\mathrm{h,h}})^{-1} k_h(\mathbf{S}, \tilde{\mathbf{s}}'), \tag{12}$$

where $[\mathbf{K}_{\mathrm{g,g}}]_{i,j} = k_g(\mathbf{s}_i, \mathbf{s}_j)$ and $[\mathbf{K}_{\mathrm{h,h}}]_{i,j} = k_h(\mathbf{s}_i, \mathbf{s}_j)$. Naturally, this extends also to the case of noisy observations (section 2.2).

2.4 Linear transformations of (multivariate) Gaussians and sampling from a Gaussian process

Consider a multivariate Gaussian $\mathbf{f} \sim N(0, \mathbf{K}_{\mathrm{f,f}})$ and a linar transformation $\mathbf{z} = \mathbf{c} + \mathbf{A}\,\mathbf{f}$ where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{c}$ an $m \times 1$ vector. The vector $\mathbf{z}$ is then Gaussian distributed, $\mathbf{z} \sim N(\mathbf{c}, \mathbf{A}\,\mathbf{K}_{\mathrm{f,f}}\,\mathbf{A}^{\mathsf{T}})$. If the matrix $\mathbf{A}\,\mathbf{K}_{\mathrm{f,f}}\,\mathbf{A}$ is not full rank (for example, if $m > n$) then the multivariate normal is degenerate and does not have density. The density for the transformed vector can be formed by considering a subset of rank$(\mathbf{A}\,\mathbf{K}_{\mathrm{f,f}}\,\mathbf{A})$ coordinates of $\mathbf{z}$ and treating the other co-ordinates as their transformation.

The above property allows an efficient way to simulate from a Gaussian process. Assume we have a way to simulate i.i.d. Gaussian random variables (all computing programs have Gaussian random number generator). We can simulate from a Gaussian process with mean function $\mu(\mathbf{s})$ and covariance function $k(\mathbf{s}, \mathbf{s}')$ at locations $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_n]^{\mathsf{T}}$ as follows. Construt a vector $\mu = [\mu(\mathbf{s}_1), ..., \mu(\mathbf{s}_n)]^{\mathsf{T}}$ and a covariance matrix $[\mathbf{K}_{\mathrm{f,f}}]_{i,j} = k(\mathbf{s}_i, \mathbf{s}_j)$. Form a Cholesky decomposition of the covariance matrix $\mathbf{L}\mathbf{L}^{\mathsf{T}}$. Form an $n \times 1$ vector of i.i.d. zero mean and unit variance Gaussian random variables, $\mathbf{z} \sim N(0, \mathbf{I})$. After this form a vector $\mathbf{f} = \mu + \mathbf{L}\mathbf{z}$. The vector $\mathbf{f}$ is then a sample from the Gaussian process at locations $\mathbf{S}$. By repeating this procedure you can construct mutliple realizations from the same process. (See also exercises). Note! In some cases the constructed covariance matrix $\mathbf{K}_{\mathrm{f,f}}$ may be numerically unstable so that the Cholesky decomposition does not remain positive definite. In this case adding small constant ("jitter"; typically $< 10^{-6}$ is enough) to the diagonal helps.

Consider also a linear model $f(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\,\beta$, where $\mathbf{x}$ is a $p \times 1$ vector of covariates and β a $p \times 1$ vector of coeffcients with Gaussian prior $\beta \sim N(0, \Sigma)$. Since $f(\mathbf{x})$ is linear transformation of β the model can be thought to define a Gaussian process whose realizations are all linear functions of $\mathbf{x}$. See exercises.

2.5 Spatial misalignment (change of support)

3 On covariance functions and relation to classical geostatistics

4 Hierarchical spatial models

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman Hall/CRC, second edition.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer Science+Business Media, LLC.

Finkenstädt, B., Held, L., and Isham, V. (2007). *Statistical Methods for Spatio-Temporal Systems*. Chapman & Hall/CRC.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.