Spatial modelling and Bayesian inference

Lecture notes

Jarno Vanhatalo

March 19, 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:

- **19.3.2017** Updated section 2.2, Added section 3.
- **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, $Cov\left(f(\mathbf{s}_i), f(\mathbf{s}_j)\right) = k(\mathbf{s}_i, \mathbf{s}_j \,|\theta)$. 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 $\|\mathbf{s}_i - \mathbf{s}_j\|$ is the euclidean distance (the L_2 norm) 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

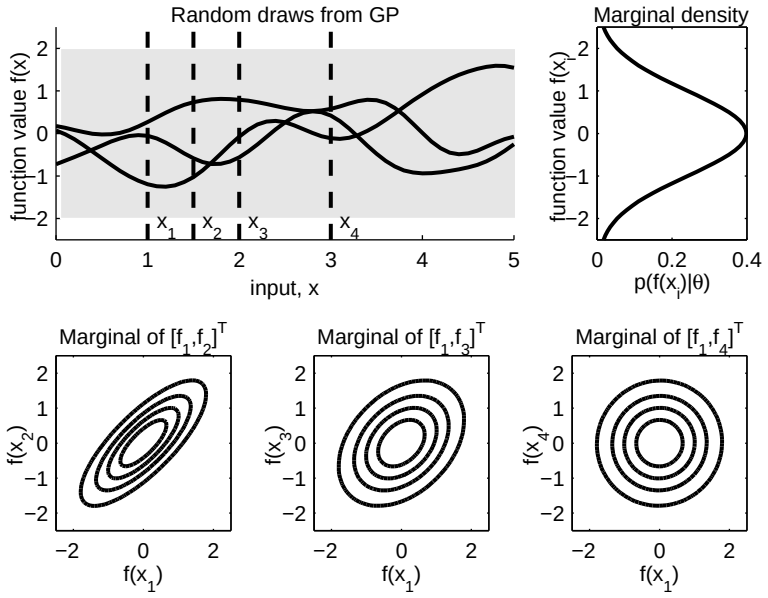

Figure 1: An illustration of a Gaussian process. The upper left figure presents three functions drawn randomly from a zero mean Gaussian process with squared exponential covariance function. The hyperparameters are $l = 1$ and $\sigma^2 = 1$ and the grey shading represents central 95% probability interval. 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.

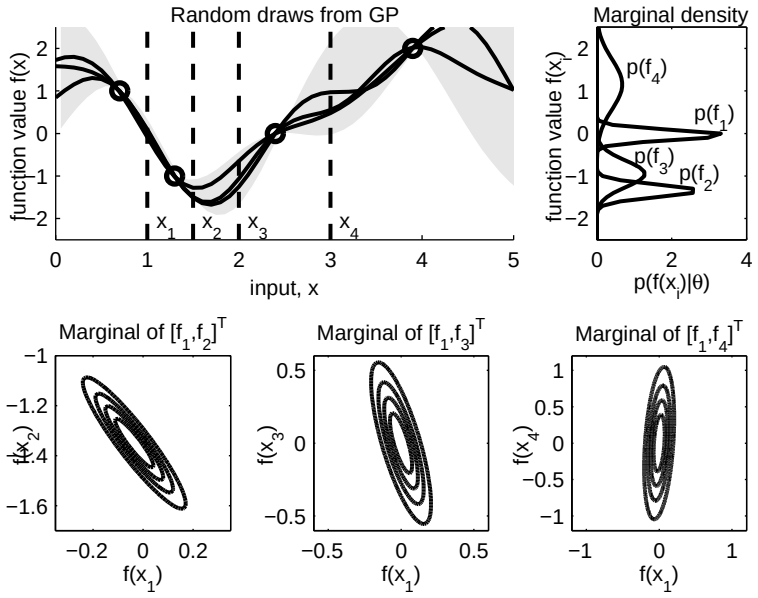

Figure 2: A conditional (posterior) GP $p(\tilde{f}|\,\mathbf{f},\theta)$. The observations $\mathbf{f} = [f(0.7) = 1, f(1.3) = -1, f(2.4) = 0, f(3.9) = 2]^{\mathrm{T}}$ are plotted with circles in the upper left subfigure and the prior GP is illustrated in the figure 1. When comparing the subfigures 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 Observations with Gaussian noise

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

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

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

To derive this result a bit more formally let's define the inference and prediction problem as follows. Consider we have a zero mean Gaussian process $f(\mathbf{s}) : D \to \Re$ where D is the index domain (e.g., a subset of $\Re^2$). Consider further that we have an

observation process $p(y(\mathbf{s}_i)|\,\mathbf{f}(\mathbf{s})) = p(y(\mathbf{s}_i)|\,\mathbf{f}(\mathbf{s}_i))$ where we assume that each observation is conditionally independent of the other observations given the process realization at that location. Now, consider we have made n observations at locations $S = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$ and denote by $\mathbf{y} = [y(\mathbf{s}_i), ..., y(\mathbf{s}_n)]^\mathsf{T}$ the vector of these observations and by $\mathbf{f} = [f(\mathbf{s}_1), ..., f(\mathbf{s}_n)]^\mathsf{T}$ the respective latent variables. Due to the marginalization properties of the Gaussian process the prior distribution of the latent variables is $p(\mathbf{f}) = N(0, \mathbf{K}_{\mathrm{f,f}})$. Hence, we can first solve the posterior distribution for the latent variables at the observation locations

$$p(\mathbf{f}\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})} = \frac{N(\mathbf{f}\,|0, \mathbf{K}_{\mathrm{f,f}})\Pi_{i=1}^n p(y_i|f_i)}{p(\mathbf{y})}. \tag{11}$$

For example, in the case of a Gaussian observation model $p(y_i|f_i) = N(y_i|f_i, \sigma_\epsilon^2)$ the posterior distribution of $\mathbf{f}$ is (see exercises)

$$\begin{aligned} p(\mathbf{f}\,|\,\mathbf{y}) &\propto N(\mathbf{f}\,|0, \mathbf{K}_{\mathrm{f,f}})\Pi_{i=1}^n N(y_i|f_i, \sigma_\epsilon^2) &(12)\\ &= N\left(\mathbf{f}\,|\,\mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma_\epsilon^2\mathbf{I}^{-1})\,\mathbf{y}, (\mathbf{K}_{\mathrm{f,f}}^{-1} + \sigma_\epsilon^{-2}\mathbf{I})^{-1}\right). &(13)\end{aligned}$$

Next, we solve the posterior predictive distribution of the latent function $f(\tilde{\mathbf{s}})$ at a new location $\tilde{\mathbf{s}} \in D$. To do this we utilize the marginalization property of the Gaussian process for a second time to derive the joint distribution of $[\mathbf{f}^\mathsf{T}, \tilde{f}]^\mathsf{T}$, where $\tilde{f} = f(\tilde{\mathbf{s}})$. This is given by equation (4). After this we use the result concerning the conditional distribution $\tilde{f}|\,\mathbf{f}$ in equation (5) and marginalize over the posterior of $\mathbf{f}$ to obtain the posterior predictive distribution for $\tilde{f}$

$$\begin{aligned} p(\tilde{f}|\,\mathbf{y}) &= \int p(\tilde{f}|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{y})d\,\mathbf{f} &(14)\\ &= \int \mathrm{N}(\tilde{f}|\,\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{f}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}) \\ &\quad N\left(\mathbf{f}\,|\,\mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma_\epsilon^2\mathbf{I}^{-1})\,\mathbf{y}, (\mathbf{K}_{\mathrm{f,f}}^{-1} + \sigma_\epsilon^{-2}\mathbf{I})^{-1}\right) d\,\mathbf{f} &(15)\\ &= \mathrm{N}(\tilde{f}|\,\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}) &(16)\end{aligned}$$

Since this is valid for any $\tilde{\mathbf{s}} \in D$ the posterior for $f(\mathbf{s})$ is a Gaussian process with mean and covariance functions as in equations (9)-(10). However, in general, if the observation model is not Gaussian, the posterior distribution of $f(\mathbf{s})$ is not Gaussian process. This will be discussed more in Section 4.

In order to calculate the posterior predictive distribution for a new observation, $\tilde{y} = y(\tilde{\mathbf{s}})$ we can utilize the assumption of conditional independence between $y(\mathbf{s})$ given $f(\mathbf{s})$ to obtain

$$p(\tilde{y}|\,\mathbf{y}) = \int p(\tilde{y}|\tilde{f})p(\tilde{f}|\,\mathbf{y})d\tilde{f}. \tag{17}$$

In the case of Gaussian observation model this will be

$$\tilde{y}|\,\mathbf{y} \sim \mathrm{N}(\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}, \sigma^2 + \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}) \tag{18}$$

which differs from the posterior predictive distribution for $f(\tilde{\mathbf{s}})$ only in the covariance which has now the contribution of the noise variance σ^2 in it.

2.3 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.

3 On construction of Gaussian processes and their covariance functions

In order to build Gaussian process models we need tools to build valid mean and covariance functions. The covariance function has to satisfy the positive definite condition (2). Hence, for any finite set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ the covariance function has to produce a covariance matrix $\mathbf{K}_{\mathrm{f,f}}$ such that if $\mathbf{f} \sim N(0, \mathbf{K})$ the variance of $\mathbf{v}^{\mathsf{T}}\,\mathbf{f}$ is valid for any $\mathbf{v}$; that is $Var(\mathbf{v}\,\mathbf{f}) = \mathbf{v}^{\mathsf{T}}\mathbf{K}\mathbf{v} \geq 0$ with strict inequality if not all v_i are 0. Hence, any function that produces positive definite covariance matrices is valid for constructing Gaussian process. Then a practical problem remains how to construct such functions. After constructing a positive definite covariance function, another practical question is what are the properties of a Gaussian process encoded by a specific covariance function. These questions have motivated a waste literature in statistics and mathematics (see e.g. **?**) and here we will review few common classes of covariance functions and their properties. We will also discuss how certain common models can be extended to Gaussian processes formalism.

3.1 Covariance function terminology and basic results

A covariance function is called *stationary* if it is a function of $\mathbf{h} = \mathbf{s} - \mathbf{s}'$ only[2]. Hence, it is invariant to translations in the index domain. If the covariance function is a function of

[2]Note that this means *weak stationarity* for the corresponding stochastic process whereas *strong stationarity* would mean that all of its finite dimensional distributions are invariant to translations.

distance only $\|\mathbf{s}-\mathbf{s}'\|$ it is called *isotropic*. For example, an exponential covariance function (3) is stationary and isotropic. However, if we modify the calculation of the distance in the input domain D and define an exponential covariance function with dimension scaling

$$k_{\exp}(\mathbf{s}_i, \mathbf{s}_j \,|\theta) = \sigma^2_{\exp} e^{\left(-\Sigma_{d=1}^{D}(s_{i,d}-s_{j,d})^2/l^2\right)^{1/2}}, \tag{19}$$

the resulting covariance remains stationary but is not isotropic if $D > 1$. In higher dimensional index space different length-scales, l_d, per input dimension allows for different smoothness per dimension. Moreover, this example illustrates also that a covariance function that is isotropic in dimension D need not be isotropic in $D+1$.

In case of stationary covariance functions we can calculate a *semivariogram*, $\gamma(\mathbf{h}) = k(\mathbf{0}) - k(\mathbf{h})$, and a *variogram*, $2\gamma(\mathbf{h})$, functions. These terms arise from traditional geostatistics where variograms and semivariograms were empirically estimated from data. After this the covariance function parameters were chosen so that the semivariogram of a chosen covariance function matched the empirical semivariogram points, for example, in root mean square sense. In this course we will not use variograms but the term is good to know since it is still used extensively in some fields of geosciences. In case of stationary covariance function the semivariogram might depend on the direction it is calculated with respect to whereas with isotropic covarance functions the variograms do not depend on the direction. There are three chraracteristics that are traditionally associated with variograms, the *nugget*, the *sill* and the *range*. By definition the nugget is $\lim_{\|\mathbf{h}\|\to 0^+}\gamma(\mathbf{h})$. The sill is defined to be $\lim_{\|\mathbf{h}\|\to\infty}\gamma(\mathbf{h})$. The range is the distance at which $\gamma(\mathbf{h})$ reaches its sill. For example, consider the Gaussian observation model (8) where the Gaussian process has an exponential covariance function (3). The nugget of the variogram of $y(\mathbf{s})$ would be σ^2_ϵ and the sill would be $\sigma^2_{\exp}+\sigma^2_\epsilon$. However, the sill is reached only asymptotically for which reason the range does not exist.

In model based spatial statistics, which is considered in this course, range is typically defined to be the distance at which the covariance has dropped to 5% of its maximum. However, this might vary in the literature for which reason care need to be taken when interpreting the term. (See also exercises.)

In general if we have two valid covariance functions $k_1(\mathbf{s},\mathbf{s}')$ and $k_2(\mathbf{s},\mathbf{s}')$, then the functions $ak_1(\mathbf{s},\mathbf{s}')+bk_2(\mathbf{s},\mathbf{s}')$, $ck_1(\mathbf{s},\mathbf{s}')k_2(\mathbf{s},\mathbf{s}')$ and their combinations are valid covariance functions for all $a,b,c>0$ (see exercises). Similarly, if $k_1(\mathbf{s},\mathbf{s}') = k_1(s_1,s_1')$ is a function of only the first element of $\mathbf{s}$ and $k_2(\mathbf{s},\mathbf{s}') = k_2(s_2,s_2')$ is a function of the second element of $\mathbf{s}$ then any multiplicative or additive combination of $k_1(\mathbf{s},\mathbf{s}')$ and $k_2(\mathbf{s},\mathbf{s}')$ is a valid covariance function. This extents to any combination of elements in $\mathbf{s}$ (however, see also discussion in section 3.4). For example, if $f(\mathbf{s},t):\Re^2\times\Re\to\Re$ is a Gaussian process in space, $\mathbf{s}$, and time, t, one common approach to define a covariance function for this process is to use a separable form $k\left((\mathbf{s},t),(\mathbf{s}',t')\right) = k_1(\mathbf{s},\mathbf{s}')k_2(t,t)$, where $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are some radial basis functions. Convolution is yet another way to construct new covariance functions. If $k_1(\mathbf{h})$ and $k_2(\mathbf{h})$ are valid covariance functions then $k(\mathbf{h}) = \int k_1(\mathbf{h}-\mathbf{t})k_2(\mathbf{t})d\mathbf{t}$ is a valid covariance function as well.

3.2 Stationary covariance functions and Bochner's Theorem

One very influential result for Gaussian process theory is the *Bochner's Theorem* which provides a tool to construct stationary positive definite covariance functions in an arbitrary r-dimensional *Euclidean space*. For real-valued processes, Bochner's Theorem states that

$k(\mathbf{h})$, where $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, is positive definite if and only if

$$k(\mathbf{h}) = \int \cos\left(\mathbf{w}^\mathsf{T}\mathbf{h}\right) G(d\mathbf{w}), \tag{20}$$

where $G(d\mathbf{w})$ is a bounded, positive, symmetric about $\mathbf{0}$ measure in $\Re^r$. Since $G(d\mathbf{w})$ is assumed symmetric and $e^{i\mathbf{w}^\mathsf{T}\mathbf{h}} = \cos(\mathbf{w}^\mathsf{T}\mathbf{h}) + i\sin(\mathbf{w}^\mathsf{T}\mathbf{h})$ we have

$$k(\mathbf{h}) = \int e^{i\mathbf{w}^\mathsf{T}\mathbf{h}} G(d\mathbf{w}). \tag{21}$$

If $G(d\mathbf{w})$ is not assumed symmetric about $\mathbf{0}$, equation (21) still provides a valid covariance function but now for a complex-valued random process on $\Re^r$ (**?**).

Hence, $G(d\mathbf{w})/k(\mathbf{0}) = G(d\mathbf{w})/\int G(d\mathbf{w})$ is referred as the *spectral distribution* of $k(\mathbf{h})$. Typically $G(d\mathbf{w})$ is constructed so that it has a density with respect to Lebesgue measure and $G(d\mathbf{w}) = g(\mathbf{w})d\mathbf{w}$. Then, $g(\mathbf{h})/k(\mathbf{0})$ is referred to *spectral density* of a covariance function $k(\mathbf{h})$. For example, the Matérn class of covariance functions which are widely used in spatial statistics are constructed using Cauchy spectral density. See, e.g., (Rasmussen and Williams, 2006, pp. 84-85) and (Banerjee et al., 2015, p. 62). A more thorough discussion on Bochner's Theorem is provided by, e.g. Banerjee et al. (2015).

Here it should be noticed also that Bochner's Theorem is valid in Euclidean space and we cannot straighforwardly apply covariance functions constructed by Bochner's Theorem in other spaces. Hence, if we want to define a valid covariance function on the surface of a globe we need different tools for that. Further discussion on such covariance functions are provided, for example, by Banerjee et al. (2015); Lindgren et al. (2011); Banerjee (2005).

3.3 Gaussian process interpretation for linear model

Consider the model $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\beta$ where $\mathbf{x}$ is a $p \times 1$ vector of covariates and $\beta \sim N(0, \Sigma_\beta)$. Hence, for any collection of covariate vectors $\mathbf{X} = [\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_n^\mathsf{T}]^\mathsf{T}$ the joint distribution of $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\mathsf{T}$ is a multivariate Gaussian $\mathbf{f} \sim N(0, \mathbf{X}\Sigma_\beta\mathbf{X}^\mathsf{T})$ (section 2.3). Hence, a linear model can be seen as a Gaussian process with covariance function $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T}\Sigma_\beta\mathbf{x}'$ (see (Rasmussen and Williams, 2006) for a more detailed treatment).

3.4 Additive and hierarchical Gaussian processes

In section 2.2 we considered additive Gaussian observation error. 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

$$m_{h|\,\mathbf{f}}(\tilde{\mathbf{s}}) = k_h(\tilde{\mathbf{s}}, \mathbf{S})(\mathbf{K}_{\mathrm{g,g}} + \mathbf{K}_{\mathrm{h,h}})^{-1}\,\mathbf{y} \tag{22}$$

$$k_{h|\,\mathbf{f}}(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') = k_h(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') - k_h(\tilde{\mathbf{s}}, \mathbf{S})(\mathbf{K}_{\mathrm{g,g}} + \mathbf{K}_{\mathrm{h,h}})^{-1}k_h(\mathbf{S}, \tilde{\mathbf{s}}'), \tag{23}$$

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).

Let's next look at the linear Gaussian model in section 3.3 in a bit more detail. Lets rewrite the linear model as $f(\mathbf{z}) = \mathbf{z}^\mathsf{T}\eta$ where $\mathbf{z}$ is a vector of covariates so that $\mathbf{z} = [1, x_1, \ldots, x_p]^\mathsf{T}$ and $\eta = [\alpha, \beta_1, \ldots, \beta_p]^\mathsf{T}$. Let's further define the prior $\eta \sim N(0, \Sigma_\eta)$ where

$$\Sigma_\eta = \begin{bmatrix} \sigma_\alpha^2 & 0 \\ 0 & \Sigma_\beta \end{bmatrix} \tag{24}$$

Now, we can write

$$\begin{aligned} f(\mathbf{z}) &= \mathbf{z}^\mathsf{T}\eta & (25)\\ &= \alpha + \mathbf{x}^\mathsf{T}\beta = f(\mathbf{x}), & (26) \end{aligned}$$

where α is an intercept with prior distribution $\alpha \sim N(0, \sigma_\alpha^2)$ and β is the vector of linear weights as in section 3.3 with the prior $\beta \sim N(0, \Sigma_\beta)$. There are now few ways to interpret this model. One is the Gaussian process interpretation where $f(\mathbf{z}) \sim GP(0, k(\mathbf{z}, \mathbf{z}'))$ where $k(\mathbf{z}, \mathbf{z}') = \sigma_\alpha^2 + \mathbf{x}^\mathsf{T}\Sigma_\beta\mathbf{x}' = k_\alpha(\mathbf{z}, \mathbf{z}') + k_x(\mathbf{z}, \mathbf{z}')$. And hence, for any set of covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ the latent vector $\mathbf{f}$ has a multivariate Gaussian distribution $\mathbf{f} \sim N(0, \mathbf{K}_\alpha + \mathbf{K}_x)$. Hence, the model is a Gaussian process with an additive covariance function. However, the model does not correspond to sum of two Gaussian processes! Even though $k(\mathbf{z}, \mathbf{z}')$ is a valid covariance function the first additive element in it is not a valid covariance function on its own. The matrix $\mathbf{K}_\alpha = \sigma_\alpha^2\mathbf{1}\mathbf{1}^\mathsf{T}$, where $\mathbf{1}$ is an $n\times 1$ vector of ones, is not positive definite. This is natural since α is a random variable with Gaussian distribution. Another way to define this model is through a hierarchical construction

$$\begin{aligned} f(\mathbf{x}) &\sim GP\left(\mu, k(\mathbf{x}, \mathbf{x}')\right) \\ \mu &\sim N(0, \sigma_\alpha^2), \end{aligned}$$

where σ_α^2 is the prior variance of the mean of a Gaussian process. Hence, by a coice of covariance function we can actually implicitly model some hierarchical latent Gaussian models (see Rasmussen and Williams, 2006, for more discussion on mean functions in GPs).

Next we will consider a Gaussian processes in the setting of traditional random effects models (**?**) which are common in many practical applications. Consider a setup where n experiments are conducted at m different experimental plots as illustrated in Figure 3. This could be, for example, aggricultural experiment where each plot, z, corresponds to one field which is divided into experimental units within it. The experimental setup is encoded by covariates $\mathbf{x}$ telling, for example, how much fertilization is used in the experiment. The plots are typically not identical but, for example, the soil composition, depth of the fertile soil etc. may vary. Hence, we are anticipating that, in addition to a covariate effect, there is a plot level effect to the outcome of an experiment y_i, $i = 1, \ldots, n$. Moreover, since each plot and each experiment within a plot is at different spatial location we might anticipate that there is also spatially correlated randomness in the outcomes of the experiments due to, for example, varying weather conditions during the experiments. A typical way to analyze this kind of data is to construct a hierarchical additive model

$$y(\mathbf{x}_i, z_i, \mathbf{s}_i) = \mathbf{x}_i^\mathsf{T}\beta + \epsilon_{z_i} + \phi(\mathbf{s}_i) + \epsilon_i \tag{27}$$

where the effect of experimental treatments are assumed linear with weights $\beta \sim N(0, \Sigma_\beta)$, $\epsilon_{z_i} \sim N(0, \sigma_z^2)$ is a random effect capturing the plot level effect, $\phi(\mathbf{s}_i) \sim GP(0, k(\mathbf{s}, \mathbf{s}'))$ is a spatial random effect (a Gaussian process) and ϵ_i is an i.i.d. random error per measurement.

Let's assume $\mathbf{s}_i \in \mathbf{S} \subset \Re^2$, $\mathbf{x}_i \in \Re^p$ and $z_i \in \mathcal{N}$ is the identifier of plot i. We can then formulate the hierarchical additive model (27) as $\mathbf{y}(\mathbf{x}_i, z_i, \mathbf{s}_i) = f(\mathbf{x}_i, z_i, \mathbf{s}_i) + \epsilon_i$ where $f(\mathbf{x}, z, \mathbf{s})$ is a Gaussian process in a domain $D \subset \mathbf{S} \times \Re^p \times \mathcal{N}$ with an additive covariance function of the form

$$k\left((\mathbf{s}, \mathbf{x}, z), (\mathbf{s}', \mathbf{x}', z')\right) = k_s(\mathbf{s}, \mathbf{s}') + k(\mathbf{x}, \mathbf{x}') + \sigma_z^2 \delta_z(z'), \tag{28}$$

The covariance function $k_s(\mathbf{s}, \mathbf{s}')$ could be any radial basis covariance function suitable for modeling spatial dependence and $k(\mathbf{x}, \mathbf{x}')$ would be the covariance function corresponding to the linear model and $\delta_z(z')$ is a delta function returning 1 if $z = z'$ and zero otherwise. Let's now order the data so that we stack together the observations in ascending order of plot indicator; that is first the observation from plot 1, then from plot 2 and so on all the way to plot m. Then the prior for the latent vector $\mathbf{f}$ would be a zero mean multivariate Gaussian with a covariance matrix

$$\mathbf{K}_{\mathrm{f,f}} = \left[\mathbf{K}_s\right] + \left[\mathbf{X}\Sigma_\beta\mathbf{X}^\mathsf{T}\right] + \begin{bmatrix} \left[\sigma_z^2\mathbf{J}_{m_1}\right] & & & \\ & \left[\sigma_z^2\mathbf{J}_{m_2}\right] & & \\ & & \ddots & \\ & & & \left[\sigma_z^2\mathbf{J}_{m_n}\right] \end{bmatrix} \tag{29}$$

where $\mathbf{J}_{m_1}$ is a matrix of size $m_1 \times m_1$ with one in every element and m_i is the number of measurements in the z_i'th plot. The matrices $\mathbf{K}_s$ and $\mathbf{X}\Sigma_\beta\mathbf{X}^\mathsf{T}$ are full $n \times n$ matrices whereas the rightmost matrix corresponding to covariance function $\sigma_z^2\delta_z(z')$ is a block diagonal matrix. Hence, the plot structure in the data transfers naturally to structured covariance matrix in the prior of the latent variables. Another way to derive the covariance function $\sigma_z^2\delta_z(z')$ would be to define a piece wise constant mean function with Gaussian priors for the constants.

The model structure (27) is present in many other settings as well. Some examples include species distribution modeling (**?**), genetics (**?**), ... However, depending on the application the random effect might be indexed in some other domain than space.

3.5 Spatial misalignment (change of support)

4 Hierarchical spatial models

References

Banerjee, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics*, 61(2):617–625.

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.

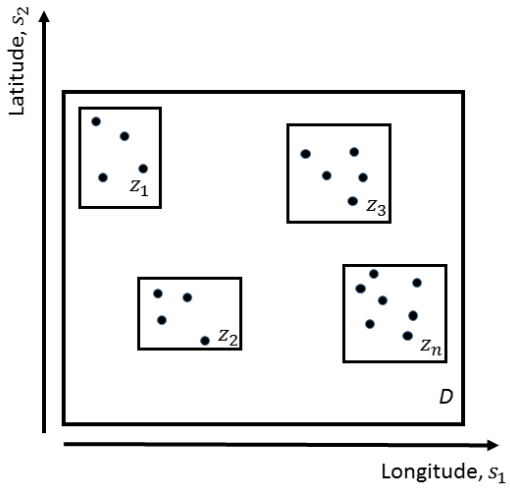

Figure 3: An illustration of an experimental setup where the spatial domain, D is divided into n experimental plots denoted by $z_i = 1, \ldots, n$ and at each plot we have measurements, y_i, (dots) with different experimental treatments, $\mathbf{x}_i$.

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

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(4):423–498.

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