# Spatial modelling and Bayesian inference

Lecture notes

Jarno Vanhatalo

April 11, 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:

- **11.4.2017** Added section 5
- **3.4.2017** Added section 4.2
- **27.3.2017** Added section 4
- **23.3.2017** Small updates in sections 2.2. and 3.1. Added few proofs for the results concerning Gaussian processes with additive covariance functions to section 3.4.
- **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 $2 \times 1$ 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 multi-

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

variate Gaussian distribution

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

where $\mu$ 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 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 $\theta$ 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

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

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; Cressie, 1993).

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 $k(\mathbf{s}, \mathbf{S}) : \Re^d \times$

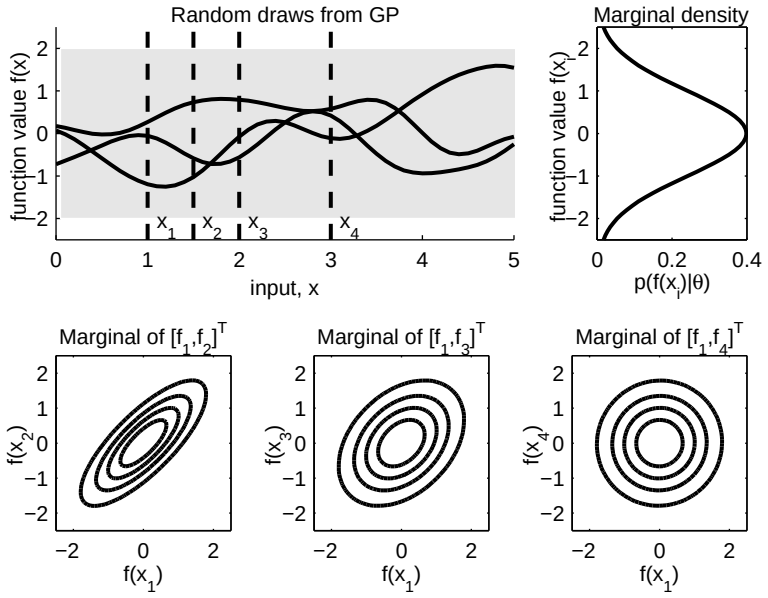



Figure 1: An illustration of a Gaussian process. The upper left figure presents three functions drawn randomly from a zero mean Gaussian process with squared exponential covariance function. The hyperparameters are $l = 1$ and $\sigma^2 = 1$ and the grey shading represents central 95% probability interval.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

$\Re^{d\times n} \rightarrow \Re^{1\times n}$, 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 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*

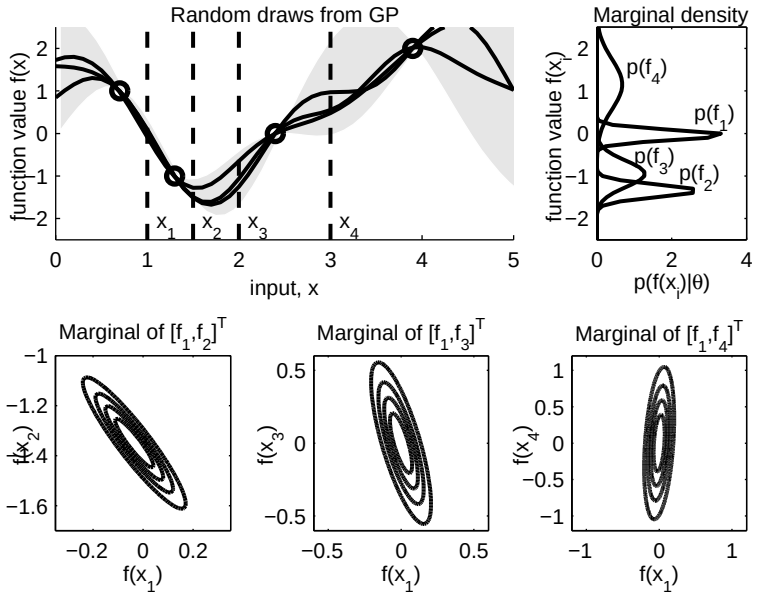



Figure 2: A conditional (posterior) GP $p(\tilde{f}|\,\mathbf{f},\theta)$. The observations $\mathbf{f} = [f(0.7) = 1, f(1.3) = -1, f(2.4) = 0, f(3.9) = 2]^{\mathrm{T}}$ are plotted with circles in the upper left subfigure and the prior GP is illustrated in the figure 1. When comparing the subfigures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

since it is conditioned to the values of parameters $\theta$ which we will later infer along the latent variables. The conditional posterior GP is illustrated in Figure 2.

## 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. Possibly 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 $\sigma^2_\epsilon$ 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 $f(\mathbf{s}) : D \rightarrow \Re$

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

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

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)

$$p(\mathbf{f}\,|\,\mathbf{y}) \propto N(\mathbf{f}\,|0, \mathbf{K}_{\mathrm{f,f}}) \prod_{i=1}^n N(y_i|f_i, \sigma_\epsilon^2) \tag{12}$$

$$= N\left(\mathbf{f}\,|\,\mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma_\epsilon^2\mathbf{I})^{-1}\,\mathbf{y}, (\mathbf{K}_{\mathrm{f,f}}^{-1} + \sigma_\epsilon^{-2}\mathbf{I})^{-1}\right). \tag{13}$$

Next, we solve the posterior predictive distribution of the latent function $f(\tilde{\mathbf{s}})$ at a set of new locations $\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_m \in D$. To do this we utilize the marginalization property of the Gaussian process for a second time to derive the joint distribution of $[\mathbf{f}^\mathsf{T}, \tilde{\mathbf{f}}]^\mathsf{T}$, where $\tilde{\mathbf{f}} = [f(\tilde{\mathbf{s}}_1), \ldots, f(\tilde{s}_m)]^\mathsf{T}$. This is given by equation (4). After this we use the result concerning the conditional distribution $\tilde{\mathbf{f}}|\,\mathbf{f}$ in equation (5) and marginalize over the posterior of $\mathbf{f}$ to obtain the posterior predictive distribution for $\tilde{\mathbf{f}}$

$$p(\tilde{\mathbf{f}}|\,\mathbf{y}) = \int p(\tilde{\mathbf{f}}|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{y})d\,\mathbf{f} \tag{14}$$

$$= \int \mathrm{N}(\tilde{\mathbf{f}}|\,\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{f}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}})$$

$$N\left(\mathbf{f}\,|\,\mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma_\epsilon^2\mathbf{I}^{-1})\,\mathbf{y}, (\mathbf{K}_{\mathrm{f,f}}^{-1} + \sigma_\epsilon^{-2}\mathbf{I})^{-1}\right) d\,\mathbf{f} \tag{15}$$

$$= \mathrm{N}(\tilde{\mathbf{f}}|\,\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}) \tag{16}$$

Since this is valid for any set of $\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_m \in D$ and any $m > 0$ 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 a 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

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

This can be extended also to a set of new observations $\tilde{\mathbf{y}} = [y(\tilde{s}_1), \ldots, y(\tilde{\mathbf{s}}_m)]^\mathsf{T}$. In the case of Gaussian observation model and set this will be

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

which differs from the posterior predictive distribution for $\tilde{\mathbf{f}}$ only in the covariance which has now the contribution of the noise variance $\sigma^2$ 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

[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.

of distance only $\|\mathbf{s}-\mathbf{s}'\| = \|\mathbf{h}\|$ 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*, $\gamma_y(\mathbf{h}) = k_y(\mathbf{0}) - k_y(\mathbf{h})$, and a *variogram*, $2\gamma_y(\mathbf{h})$, functions. Here, the subindex $y$ denotes that the variogram is calculated for the observations; that is the covariance functions are the covariance functions of $y(\mathbf{s})$ in (8). 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 (Gelfand et al., 2010; Banerjee et al., 2015). 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 $\lim_{\|\mathbf{h}\|\to 0^+} \gamma_y(\mathbf{h})$. The sill is defined to be $\lim_{\|\mathbf{h}\|\to\infty} \gamma_y(\mathbf{h})$. The range is the distance at which $\gamma_y(\mathbf{h})$ 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 $\sigma^2_\epsilon$ 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 or validated by Bochner's Theorem in other spaces. Hence, if we want to define a valid covariance function on, for example, 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 $\beta \sim N(0, \Sigma_\beta)$. 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 $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T}\Sigma_\beta\mathbf{x}'$ (a more detailed treatment is given by Rasmussen and Williams (2006)).

## 3.4 Additive and hierarchical Gaussian processes

In section 2.2 we considered an additive Gaussian observation error. More generally, let $f(\mathbf{s}) = h(\mathbf{s}) + g(\mathbf{s})$, where $h(\mathbf{s}) : D \to \Re$ and $g(\mathbf{s}) : D \to \Re$ are mutually independent zero mean Gaussian processes with covariance functions $k_h(\mathbf{s}, \mathbf{s}')$ and $k_g(\mathbf{s}, \mathbf{s}')$. Then, $f(\mathbf{s}) : D \to \Re$ follows a Gaussian process with covariance function $k_h(\mathbf{s}, \mathbf{s}') + k_g(\mathbf{s}, \mathbf{s}')$. To show this, consider any set of input indices $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$. Then, $\mathbf{h} =$

$[h(\mathbf{s}_1), \ldots, h(\mathbf{s}_n)]^\mathsf{T} \sim N(\mathbf{0}, \mathbf{K}_{\mathrm{h,h}})$ and $\mathbf{g} = [g(\mathbf{s}_1), \ldots, g(\mathbf{s}_n)]^\mathsf{T} \sim N(\mathbf{0}, \mathbf{K}_{\mathrm{g,g}})$ where $[\mathbf{K}_{\mathrm{h,h}}]_{i,j} = k_h(\mathbf{s}_i, \mathbf{s}_j)$ and $[\mathbf{K}_{\mathrm{g,g}}]_{i,j} = k_g(\mathbf{s}_i, \mathbf{s}_j)$. Now, define $\mathbf{f} = [f(\mathbf{s}_1), \ldots, f(\mathbf{s}_n)]^\mathsf{T} = \mathbf{h} + \mathbf{g}$. Due to properties of a multivariate Gaussian distribution, $\mathbf{f} \sim N(\mathbf{0}, \mathbf{K}_{\mathrm{f,f}})$ where $[\mathbf{K}_{\mathrm{f,f}}]_{i,j} = [\mathbf{K}_{\mathrm{h,h}} + \mathbf{K}_{\mathrm{g,g}}]_{i,j} = k_h(\mathbf{s}_i, \mathbf{s}_j) + k_g(\mathbf{s}_i, \mathbf{s}_j)$. Hence, for any collection of input indices the values of the function $f(\mathbf{s})$ are multivariate Gaussian with a covariance function $k_h(\mathbf{s}, \mathbf{s}') + k_g(\mathbf{s}, \mathbf{s}')$ and by definition it is then a Gaussian process.

Consider now that we have made observations of $f(\mathbf{s})$ at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$ but we are interested in the posterior distribution of $h(\mathbf{s})$. We can then consider the process $g(\mathbf{s})$ as a correlated noise process and proceed as in section 2.2 to solve the conditional posterior of $h(\mathbf{s})$ which is a Gaussian process with mean and covariance functions

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

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

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). During lectures we go through another way to derive this result.

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

$$f(\mathbf{z}) = \mathbf{z}^\mathsf{T}\,\eta \tag{25}$$

$$= \alpha + \mathbf{x}^\mathsf{T}\,\beta = f(\mathbf{x}), \tag{26}$$

where $\alpha$ is an intercept with prior distribution $\alpha \sim N(0, \sigma_\alpha^2)$ and $\beta$ 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 $\alpha$ is a random variable with Gaussian distribution and has the same value at every $\mathbf{x}$.

Another way to define the model (26) is through a hierarchical construction

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

where $\sigma_\alpha^2$ 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 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 $\epsilon_i$ is an i.i.d. random error per measurement. It is also assumed that the prior distributions for the additive terms are mutually independent.

Let's assume $\mathbf{s}_i \subset \Re^2$, $\mathbf{x}_i \in \Re^p$ 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 $\Re^2 \times \Re^p \times \mathcal{N}$ with an additive covariance function of the form (see section 3.4.1)

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

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 observations 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 these constants.

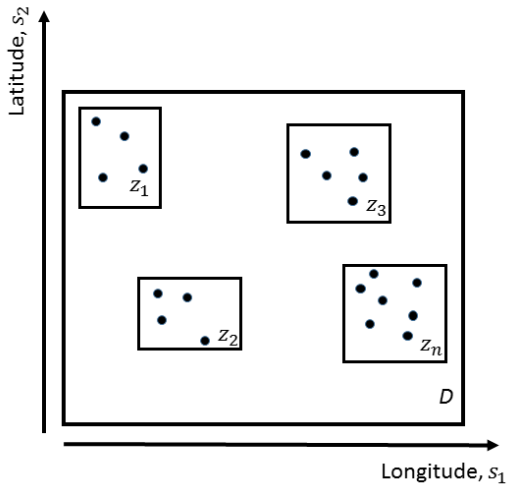



Figure 3: An illustration of an experimental setup where the spatial domain, $D$ is divided into $n$ experimental plots denoted by $z_i = 1, \ldots, n$ and at each plot we have measurements, $y_i$, (dots) with different experimental treatments, $\mathbf{x}_i$.

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.

### 3.4.1 Proof of equation (28)

Consider a function $f(\mathbf{x}, z, \mathbf{s}) = \mathbf{x}^\mathsf{T}\beta + \epsilon_z + \phi(\mathbf{s})$ with prior distributions as in the model (27). Since each of the additive terms in $f(\mathbf{x}, z, \mathbf{s})$ is Gaussian for any collection of $\{\mathbf{x}_1, z_1, \mathbf{s}_1\}, \ldots, \{\mathbf{x}_n, z_n, \mathbf{s}_n\}$ it follows that $\mathbf{f} = [f(\mathbf{x}_1, z_1, \mathbf{s}_1), \ldots, f(\mathbf{x}_n, z_n, \mathbf{s}_n)]^\mathsf{T}$ has a multivariate Gaussian distribution for any collection of input indeces and is a Gaussian process by defition. Hence, it remains to be shown that the covariance function of $f(\mathbf{x}, z, \mathbf{s})$ is (28). Let's solve this by direct calculation. Take any two index sets $\{\mathbf{x}_i, z_i, \mathbf{s}_i\}$ and $\{\mathbf{x}_j, z_j, \mathbf{s}_j\}$. Then the covariance between $f_i = f(\mathbf{x}_i, z_i, \mathbf{s}_i)$ and$f_j = f(\mathbf{x}_j, z_j, \mathbf{s}_j)$ is

$$\begin{aligned}
Cov(f(\mathbf{x}_i, z_i, \mathbf{s}_i), f(\mathbf{x}_j, z_j, \mathbf{s}_j) &= \mathrm{E}\left[(f_i - \mathrm{E}[f_i])\,(f_j - \mathrm{E}[f_j])\right] \\
&= \mathrm{E}\left[f_i f_j\right] \\
&= \mathrm{E}\left[\left(\mathbf{x}_i^\mathsf{T}\beta + \epsilon_{z_i} + \phi(\mathbf{s}_i)\right)\left(\mathbf{x}_j^\mathsf{T}\beta + \epsilon_{z_j} + \phi(\mathbf{s}_j)\right)\right] \\
&= \mathrm{E}\left[(\mathbf{x}_i^\mathsf{T}\beta)(\mathbf{x}_j^\mathsf{T}\beta) + \epsilon_{z_i}\epsilon_{z_j} + \phi(\mathbf{s}_i)\phi(\mathbf{s}_j) + \ldots\right] \\
&= k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_z^2\delta_{z_i}(z_j) + k_s(\mathbf{s}_i, \mathbf{s}_j) \qquad (30)
\end{aligned}$$

where the remainder terms on line four are pairwise multiplications of additive elements whose expectations are zero due to the assumed prior independence.

### 3.5 Spatial misalignment (change of support)

# 4 Hierarchical spatial models

This far we have considered inference with Gaussian processes when the parameters of the covariance function and obervation model (likelihood function) are fixed. Now we will extend the inference to these *hyperparameters* as well. We start with a general model definition

$$\begin{aligned}
\text{[Data | process, parameters]} \quad & \mathbf{y} \sim p(\mathbf{y}\,|f(\cdot),\gamma) && (31)\\
\text{[process | parameters]} \quad & f(\cdot)|\theta \sim \mathrm{GP}\left(m(\cdot|\theta), k(\cdot,\cdot|\theta)\right) && (32)\\
\text{[parameters]:} \quad & \theta,\gamma \sim p(\theta,\gamma) && (33)
\end{aligned}$$

where we have three hierarchical layers. The first layer is the observation process which tells the conditional distribution of observations $\mathbf{y} = [y_1, ..., y_n]^{\mathrm{T}}$ given the latent process and observation model parameters, $\gamma$. The second layer specifies the prior for the latent process conditionally to the paremeters of the covariance and mean functions, $\theta$, and the third layer specifies the prior for the hyperparameters. We have not specified the index space for the latent proces here since it can vary depending on the application. It should also be noticed that this definition does not make any assuptions of a priori conditional independence among the observations and that this formulation allows observations $y_i$ to depend on the latent function very generally, for example, at one specific location or over an entire input domain. However, the number of observations is assumed to be finite.

Next we will consider few examples of this general construction and discuss the hyperpriors for covariance function parameters. In section 5 we will consider the inference problem from practical computational point of view.

### 4.1 Hierarchical model with Gaussian observation error

Let's generalize the model with conditionally independent Gaussian observations and a Gaussian process prior in Section 2.2 to include prior distributions for its hyperparameters. Let's assume a zero mean stationary radial basis covariance function (for example the exponential or squared exponential) and that the model parameters are independent a priori so that

$$\begin{aligned}
\mathbf{y}\,|\,\mathbf{f},\sigma_\epsilon &\sim \prod_{i=1}^{n} N(y_i|f(\mathbf{s}_i),\sigma_\epsilon^2) && (34)\\
f(\mathbf{s})|l,\sigma &\sim \mathrm{GP}\left(0, k(\mathbf{s},\mathbf{s}'\,|l,\sigma^2)\right) && (35)\\
l,\sigma^2,\sigma_\epsilon^2 &\sim p(l)p(\sigma^2)p(\sigma_\epsilon^2) && (36)
\end{aligned}$$

The observation model parameter is $\gamma = \sigma_\epsilon^2$ and the process model parameters are $\theta = [l, \sigma^2, ]$. This far we have considered conditional posterior inference for the latent function $p\left(f(\tilde{\mathbf{s}})|\,\mathbf{y}, l, \sigma^2, \sigma_\epsilon^2\right)$, equations (9)-(10), and new observations $p\left(y(\tilde{\mathbf{s}})|\,\mathbf{y}, l, \sigma^2, \sigma_\epsilon^2\right)$, equations (17)-(18).

Now we want to solve the posterior distribution also for the hyperparameters

$$p(l,\sigma^2,\sigma_\epsilon^2|\,\mathbf{y}) = \frac{1}{Z}p(\mathbf{y}\,|l,\sigma^2,\sigma_\epsilon^2)p(l)p(\sigma^2)p(\sigma_\epsilon^2) \qquad (37)$$

where $Z = \int p(\mathbf{y}\,|l, \sigma^2, \sigma_\epsilon^2)p(l)p(\sigma^2)p(\sigma_\epsilon^2)dld\sigma^2 d\sigma_\epsilon^2$ is the normalizing constant. Notice that here the normalizing constant is different from the normalizing constant in equation (11) even though both denote the prior predictive distribution of the data; that is $Z = p(\mathbf{y})$. The reason is that here the model is different from the model in section 2.2 since we have included the prior for the hyperparameters to it. Hence, in this hierarchical model the normalization constant of (11) corresponds to a *marginal likelihood*

$$p(\mathbf{y}\,|l, \sigma^2, \sigma_\epsilon^2) = \int p(\mathbf{y}\,|\,\mathbf{f}, \sigma_\epsilon^2)p(\mathbf{f}\,|l, \sigma^2)d\,\mathbf{f} \tag{38}$$

$$= N(\mathbf{y}\,|0, \mathbf{K}_{\mathrm{f,f}} + \sigma^2 I). \tag{39}$$

Plugging the marginal likelihood into equation (37) we get

$$p(l, \sigma^2, \sigma_\epsilon^2|\,\mathbf{y}) \propto N(\mathbf{y}\,|0, \mathbf{K}_{\mathrm{f,f}} + \sigma^2 I)p(l)p(\sigma^2)p(\sigma_\epsilon^2). \tag{40}$$

After solving the posterior distribution for the hyperparameters we can marginalize over them to calculate the marginal predictive distributions such as

$$p(f(\tilde{\mathbf{s}})|\,\mathbf{y}) = \int p(f(\tilde{\mathbf{s}})|\,\mathbf{y}, l, \sigma^2, \sigma_\epsilon^2)p(l, \sigma^2, \sigma_\epsilon^2|\,\mathbf{y})dld\sigma^2 d\sigma_\epsilon^2 \tag{41}$$

However, the practical problem is that the equations (40) and (41) cannot be solved analytically. For this reason we need to rely on approximative methods such as Markov chain Monte Carlo (Gilks et al., 1996; Robert and Casella, 2004; Gelman et al., 2013). In this course we will not treat the theory behind Markov chains but we will concentrate on how MCMC methods can be used in Bayesian inference in practice. We will employ the STAN software package[3]. Useful references to MCMC are STAN documentation and the book by (Gelman et al., 2013). See also lecture slides on MCMC.

MCMC methods provide means to marginalize over the hyperparameters. Another option is to conduct so called *empirical Bayes inference* where the hyperparameters are optimized to their *maximum a posteriori* estimate

$$\hat{\theta}, \hat{\gamma} = \arg\max_{\theta,\gamma} p(\mathbf{y}\,|\theta, \gamma)p(\theta, \gamma). \tag{42}$$

The MAP estimates for the hyperparameters can then be used as *plug-in-values* and the posterior inference concerning the latent function condudted as in section 2. The resulting predictive distributions can then be called *conditional posterior distribution*.

## 4.2 Examples of hierarchical models with non-Gaussian observation models

In this course we will consider hierarchical models with a Binomial/Bernoulli and Poisson observation model. The former can be written as

$$\mathbf{y}\,|\,\mathbf{f}, \mathbf{N} \sim \prod_{i=1}^{n} \mathrm{Bin}(y_i|\pi(f(\mathbf{s}_i)), N_i) \tag{43}$$

$$f(\mathbf{s})|l, \sigma \sim \mathrm{GP}\left(0, k(\mathbf{s}, \mathbf{s}'\,|l, \sigma^2)\right) \tag{44}$$

$$l, \sigma^2 \sim p(l)p(\sigma^2) \tag{45}$$

[3] `http://mc-stan.org/documentation/`

where $\pi(f_i)$ is either the logistic $\pi(f_i) = 1/(1+e^{-f_i})$ or the probit, $\pi(f_i) = \Phi(f_i)$, link function and $N_i$ is the sample size at location $\mathbf{s}_i$. There are several example applications for this model and, for example, Rasmussen and Williams (2006) discus a general classification setup, where $N_i = 1$ for all $i$. In the exercises, we will consider also species distribution modeling (Gelfand et al., 2006) as an example application and derive the model from that specific perspective.

## 4.3 Prior distributions for hyperparameters

The choice of prior distributions is a central question in Bayesian statistics in general. Similarly, the choice of hyperpriors for the covariance function and observation model parameters has obtained lot of interest in the statistical literature. Here, I will review few key findings on the topic.

One very common, traditional model is a model with additive linear predictor and Gaussian process *random effect* parts, $f(\mathbf{s},\mathbf{x}) \sim GP(\mathbf{x}\beta, k(\mathbf{s},\mathbf{s}'))$. A typical choice for the prior for the (*fixed effects*) vector, $\beta$, is a zero mean Gaussian distribution with large marginal variance for each component of $\beta$ and prior independence; that is $\beta_d \sim N(0,\sigma^2_\beta)$ where, for example, $\sigma^2_\beta = 10$. This prior correponds to *vague prior* on the effects of covariates $\mathbf{x}$. As discussed in sections 3.3 and 3.4 this model can also be written as a GP with additive covariance function in which case the hyperparameters of the covariance function corresponding to the linear model are fixed.

The parameters of other covariance functions provide a more interesting problem for prior definition. In statistical literature, inference in the covariance function parameters is a natural concern but in machine learning literature they are left in less attention. An indicator of this is the usual approach to maximize the marginal likelihood which implies uniform prior for the hyperparameters (Rasmussen and Williams, 2006). However, from both practical and philosophical point of view, it is typically beneficial to give priors for the hyperparameters.

In spatial statistics literature it is well known that the length-scale and magnitude are under identifiable and the proportion $\sigma^2/l$ is more important to the predictive performance than their individual values (Diggle et al., 1998; Zhang, 2004; Diggle and Ribeiro, 2007). For example, Zhang (2004) considers Gaussian processes with Matérn covariance functions with smoothness (degrees of freedom) parameter $\nu$ and no observation error variance. Zhang (2004) shows that, under uniform prior, the ratio $\sigma^2/l^{2\nu}$ can be identified but not the individual parameters. This has direct implications to the inference concerning the hyperparameters. We cannot expect the data to be informative on both the variance and the length-scale parameters and, hence, prior information is needed to conduct sensible inference on both parameters. In the presence of observation error the identifibiality of the parameters is even bigger issue.

Typically the hyperparameters are given independent priors so that, for example, $p(\sigma^2,\sigma^2_\epsilon,l) = p(\sigma^2)p(\sigma^2_\epsilon)p(l)$. The variance parameters can be given any common variance prior, such as, the Jeffrey's log-uniform prior, $p(\sigma^2) \propto 1/\sigma^2$ or Scaled inverse $\chi^2$ (inverse-Gamma) prior (Gelman et al., 2013). In recent years, so called weakly informative priors (Gelman, 2006) for variance parameters have obtained much attention. One option that seems to have preferable properties is to give the variance parameters a prior that favors small values (that is, has a peak at zero) but has also heavy tail so that the variance parameter can increase if data is informative on it. Such priors can be con-

structed, for example, with Student-$t$ or Cauchy distributions (Gelman et al., 2013). The prior can be set also to be informative on the relative importance of the different variance components. For example, if we assume that the independent errors part should be much smaller than the process variation we can set the scale parameter of the Student-$t$ and Cauchy distributions to be smaller for $\sigma^2_\epsilon$ than for $\sigma^2$.

Typical choices for the prior for the length-scale parameter are uniform within some region or rather wide unimodal distributions with peak at the best a priori guess. However, the weakly informative priors can be used also for length-scale parameters. For example, $l \sim \text{Student} - t_+(\nu, \mu = 0, s)$ defines a prior that prefers short length-scales and, hence, short correlation ranges. By the choice $s$ we can control the width of the distribution and adjust it with respect to the size of the modelled region. With $\nu$ we can control the mass on the tail of the distribution so that with $\nu = 1$ the distribution coincides with Cauchy distribution and with $\nu \to \infty$ the distribution approaches Gaussian. Typically reasonable choices are $\nu = 1$ or $\nu = 4$. On the other hand, if we want to favor long correlation lengths we should give the prior for the inverse of the length-scale so that, for example, $1/l \sim \text{Student} - t_+(\nu, \mu = 0, s)$. This latter prior is closely related to just resently introduced penalized complexity priors (Simpson et al., 2014). In some cases one can also define (strongly) informative priors for the covariance function parameters that arise from the subject area knowledge (Hartmann et al., 2017).

The above discussion is rather practical. More fundamental reason for seriously thinking about the priors is that in Bayesian statistics leaving prior undefined (meaning uniform prior) is a prior statement as well, and sometimes it may be really awkward. Thus, it is better to spend some time thinking what the prior actually says. This is especially important with additive and other models where we have many Gaussian process components.

# 5 Inference and prediction with non-Gaussian likelihoods

Given the hierarchical model described by equations (31)-(33), our inferential interest is in the posterior distributions of the hyperparameters and the latent function, as well as in the predictive distribution of new observations. In an ideal situation, all the desired distributions could be solved analytically, but unfortunately this is not possible in general. In this section I discus few commonly used posterior approximations. I will start with methods for calculating/approximating the conditional posterior of latent variables,

$$p(\mathbf{f} \,|\, \mathbf{y}, \theta, \gamma) = \frac{p(\mathbf{y} \,|\, \mathbf{f}, \gamma) p(\mathbf{f} \,|\mathbf{X}, \theta)}{\int p(\mathbf{y} \,|\, \mathbf{f}, \gamma) p(\mathbf{f} \,|\theta) d\,\mathbf{f}}. \tag{46}$$

In sections 2.2 and 4.1 we have considered the case of Gaussian observation model, $N(\mathbf{y} \,|\, \mathbf{f}, \sigma^2_\epsilon \mathbf{I})$, where the conditional posterior can be solved analytically resulting in a multivariate Gaussian distribution. Section 5.4 treats the problem of marginalizing over the hyperparameters to obtain the marginal posterior distribution for the latent variables

$$p(\mathbf{f} \,|\, \mathbf{y}) = \int p(\mathbf{f} \,|\, \mathbf{y}, \theta, \gamma) p(\theta, \gamma | \,\mathbf{y}) d\theta d\gamma. \tag{47}$$

## 5.1 Conditional posterior of the latent function

## 5.2 Posterior mean and covariance

I will start by looking at the conditional posterior mean and variance of the latent function. If the hyperparameters are considered fixed, GP's marginalization and conditionalization properties can be exploited in prediction. Assume that we have found the conditional posterior distribution $p(\mathbf{f} \mid \mathbf{y}, \theta, \gamma)$, which, in general, is not Gaussian. We can then evaluate the posterior predictive mean simply by using the expression of the conditional mean $\mathrm{E}_{\tilde{f}\mid \mathbf{f},\theta,\gamma}[f(\tilde{\mathbf{x}})] = k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathbf{f}$ (see equation (5) and the text below it) to obtain a parametric posterior mean function

$$m_p(\tilde{\mathbf{x}}|\theta,\gamma) = \int \mathrm{E}_{\tilde{f}\mid \mathbf{f},\theta,\gamma}[f(\tilde{\mathbf{x}})] p(\mathbf{f} \mid \mathbf{y}, \theta, \gamma) d\,\mathbf{f} = k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathrm{E}_{\mathbf{f}\mid \mathbf{y},\theta,\gamma}[\mathbf{f}]. \tag{48}$$

The posterior predictive covariance between any set of latent variables, $\tilde{\mathbf{f}}$, can be evaluated with the law of total variance (see, for example, Gelman et al., 2013)

$$\mathrm{Cov}_{\tilde{\mathbf{f}}\mid \mathbf{y},\theta,\gamma}[\tilde{\mathbf{f}}] = \mathrm{E}_{\mathbf{f}\mid \mathbf{y},\theta,\gamma}\left[\mathrm{Cov}_{\tilde{\mathbf{f}}\mid \mathbf{f}}[\tilde{\mathbf{f}}]\right] + \mathrm{Cov}_{\mathbf{f}\mid \mathbf{y},\theta,\gamma}\left[\mathrm{E}_{\tilde{\mathbf{f}}\mid \mathbf{f}}[\tilde{\mathbf{f}}]\right], \tag{49}$$

where the first term simplifies to the conditional covariance in equation (5) and the second term can be written as $k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathrm{Cov}_{\mathbf{f}\mid \mathbf{y},\theta,\gamma}[\mathbf{f}]\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, k(\mathbf{X}, \tilde{\mathbf{x}}')$. Plugging these into the equation and simplifying gives us the posterior covariance function

$$k_p(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') - k(\tilde{\mathbf{x}}, \mathbf{X}) \left(\mathbf{K}_{\mathrm{f,f}}^{-1} - \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathrm{Cov}_{\mathbf{f}\mid \mathbf{y},\theta,\gamma}[\mathbf{f}]\, \mathbf{K}_{\mathrm{f,f}}^{-1}\right) k(\mathbf{X}, \tilde{\mathbf{x}}'). \tag{50}$$

Even if the exact posterior distribution $p(\tilde{f}|\mathcal{D}, \theta, \gamma)$, or in other words the posterior process, was not analytically solvable we can still evaluate its posterior mean and covariance functions easily, as long as we are able to solve the mean $\mathrm{E}_{\mathbf{f}\mid \mathcal{D},\theta,\gamma}$ and covariance $\mathrm{Cov}_{\mathbf{f}\mid \mathcal{D},\theta,\gamma}[\mathbf{f}]$. Following, for example, Csató and Opper (2002) the conditional posterior mean can be written as

$$\mathrm{E}_{\mathbf{f}\mid \mathcal{D},\theta,\gamma}[\mathbf{f}] = \mathbf{K}_{\mathrm{f,f}}\, \frac{\int d\,\mathbf{f}\, p(\mathbf{f}) \partial p(\mathbf{y} \mid \mathbf{f}) / \partial\, \mathbf{f}}{p(\mathcal{D}|\theta,\gamma)}, \tag{51}$$

and a similar result can be obtained for the covariance. The problem with the exact formulas is that the integrals in them cannot be computed exactly. The common practice to approximate the posterior distribution $p(\mathbf{f} \mid \mathbf{y}, \theta, \gamma)$ is either with Markov chain Monte Carlo (MCMC) (e.g. Neal, 1997, 1998; Diggle et al., 1998; Christensen et al., 2006; Vanhatalo and Vehtari, 2007) or by giving an analytic approximation to it (e.g. Rasmussen and Williams, 2006; Rue et al., 2009; Vanhatalo et al., 2010).

### 5.2.1 Markov chain Monte Carlo

Monte Carlo methods (Robert and Casella, 2004) are based on sampling random numbers from the desired distribution and using these samples to approximate the distribution and its properties. See Figure 4. Hence, we can sample from $p(\mathbf{f} \mid \mathbf{y}, \theta, \gamma)$ and using the samples to represent the posterior distribution of $\mathbf{f}$. In this case, the posterior marginals can be visualized with histograms and posterior statistics approximated with sample means. For example, the posterior expectation of $\mathbf{f}$ is approximated as

$$\mathrm{E}_{\mathbf{f}\mid \mathbf{y},\theta,\gamma}[\mathbf{f}] \approx \frac{1}{M} \sum_{i=1}^{M} \mathbf{f}^{(i)}, \tag{52}$$

where $\mathbf{f}^{(i)}$ is the $i$'th sample from the conditional posterior.

The problem with Monte Carlo methods is how to draw samples from arbitrary distributions. The challenge can be overcome with Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1996), where one constructs a Markov chain whose stationary distribution is the posterior distribution $p(\mathbf{f} \,|\, \mathbf{y}, \theta, \gamma)$ and uses the Markov chain samples to obtain Monte Carlo estimates. After having the posterior sample of latent variables, we can sample from the posterior predictive distribution of any set of latent variables $\tilde{\mathbf{f}}$ simply by sampling with each $\mathbf{f}^{(i)}$ one $\tilde{\mathbf{f}}^{(i)}$ from $p(\tilde{\mathbf{f}}|\, \mathbf{f}^{(i)}, \theta, \gamma)$; that is, for each $i = 1, \ldots, M$ sample

$$\tilde{\mathbf{f}}^{(i)} |\, \mathbf{f}^{(i)}, \mathbf{X}, \tilde{\mathbf{X}}, \theta \sim \mathrm{N}(\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\, \mathbf{K}_{\mathrm{f},\mathrm{f}}^{-1}\, \mathbf{f}^{(i)}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\, \mathbf{K}_{\mathrm{f},\mathrm{f}}^{-1}\, \mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}). \tag{53}$$

Similarly, we can obtain a sample of $\tilde{\mathbf{y}}$ by drawing one $\tilde{\mathbf{y}}^{(i)}$ for each $\tilde{\mathbf{f}}^{(i)}$ from the observation model $p(\mathbf{y} \,|\tilde{\mathbf{f}}, \theta, \gamma)$. We can also use the samples to approximate distributions of functions of $\mathbf{f}$. For example, the posterior distribution of the succes probability in (43) can be approximated by calculating $\pi(\tilde{f}^{(i)}) = 1/(1 + e^{-\tilde{f}^{(i)}})$ with each $i = 1, \ldots, M$.

Even though MCMC methods are theoretically appealing and the Monte Carlo estimate is proved to converge to the correct distribution as $M \to \infty$, they are often hard to implement in practice. The reason is that the time a Markov chain needs for convergence to target distribution depends on the target distribution and on the sampling algorithm. Moreover, after the convergence the sample chain might mix poorly which results in high autocorrelation and low nunber of efficient samples. Models with Gaussian process priors are notorious for their inferential challenges and there are many algorithms proposed for them.

STAN uses a tailored Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 1996, 2011) where the tuning of the sampling parameters is done in automated manner (Hoffman and Gelman, 2014). Hamiltonian Monte Carlo utilizes the gradient information of the log posterior distribution to direct the sampling to interesting regions and, hence, to speed up the convergence and improve mixing. The practical challenge with this method is how to tune the sampling parameters and in some cases this tuning might be problematic also with STAN. For example, often the latent variables are heavily dependent in their posterior distribution so that the posterior surface is narrow in some of the directions. In these situations it can help to transform the latent variables with their approximate posterior covariance (Christensen et al., 2006; Vanhatalo and Vehtari, 2007). Hence, in many cases it helps to define the model so that we sample from the posterior of

$$\mathbf{z} = \mathbf{L}^{-1}\, \mathbf{f} \tag{54}$$

where $\mathbf{L}$ is a matrix that approximates a squareroot of the posterior covariance of $\mathbf{f}$. Typicaly the posterior dependence comes mostly from the prior covariances and a simple surrogate would be the cholesky decomposition prior covariance $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{K}_{\mathrm{f},\mathrm{f}}$. In STAN this could be implemented so that

```
...
parameters {
  vector[N] z;
}
model {
  vector[N] ff;
```

```
  z ~ normal(0, 1);
  ff = L*z;

  for (n in 1:N)
    y[n] ~ ...

}
generated quantities {
  vector[N] f;
  // derived quantity (the original latent variables)
  f = L*z;
}
```

The sampling methods for the conditional posterior of the latent variables have received considerable attention in the literature. In addition to the above references, some other approaches are presented by, for example, in (Neal, 1997, 1998; Murray et al., 2010). The elliptical slice sampling method by Murray et al. (2010) has proven to be very efficient and easy to use since it does not require practically any tuning.

## 5.3 Laplace approximation

There are many analytical approximations for the conditional posterior of the latent variables. Common with them is that they all are build around the Gaussian approximation or its extensions. In this section I will consider the Laplace approximation for the conditional posterior of the latent variables. In the Laplace approximation the mean of the latent variables is approximated by the posterior mode of $\mathbf{f}$ and the covariance by the curvature of the log posterior at the mode. The approximation is constructed from the second order Taylor expansion of $\log p(\mathbf{f}\,|\mathcal{D},\theta)$ around the mode $\hat{\mathbf{f}}$, which gives a Gaussian approximation to the conditional posterior

$$p(\mathbf{f}\,|\mathcal{D},\theta,\gamma) \approx q(\mathbf{f}\,|\mathcal{D},\theta,\gamma) = \mathrm{N}(\mathbf{f}\,|\hat{\mathbf{f}},\boldsymbol{\Sigma}), \tag{55}$$

where $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}\,|\mathcal{D},\theta,\gamma)$ and $\boldsymbol{\Sigma}^{-1}$ is the Hessian of the negative log conditional posterior at the mode (Gelman et al., 2013; Rasmussen and Williams, 2006):

$$\boldsymbol{\Sigma}^{-1} = -\nabla\nabla \log p(\mathbf{f}\,|\mathcal{D},\theta,\gamma)|_{\mathbf{f}=\hat{\mathbf{f}}} = \mathbf{K}_{\mathrm{f,f}}^{-1} + \mathbf{W}. \tag{56}$$

If we assume the likelihood is factorizable as in (43), $\mathbf{W}$ is a diagonal matrix with entries $\mathbf{W}_{ii} = \nabla_{f_i}\nabla_{f_i} \log p(y|f_i,\gamma)|_{f_i=\hat{f}_i}$. This approximation is a basic building block also under the Integrated nested Laplace approximation (INLA) scheme for Gaussian Markov random field models Rue et al. (2009).

The analytic approximation constructed by Laplace approximation assumes a Gaussian form in which case it is natural to approximate the posterior predictive distribution with Gaussian as well. In this case the equations (48) and (50) give its mean and covariance. The posterior mean of $f(\tilde{\mathbf{x}})$ can be approximated from the equation (48) by replacing the posterior mean $\mathrm{E}_{\mathbf{f}\,|\,\mathbf{y},\theta}[\mathbf{f}]$ by $\hat{\mathbf{f}}$. The posterior covariance is approximated similarly by using $(\mathbf{K}_{\mathrm{f,f}}^{-1} + \mathbf{W})^{-1}$ in the place of $\mathrm{Cov}_{\mathbf{f}\,|\,\mathbf{y},\theta}[\mathbf{f}]$. Thus, after some rearrangements and

using $\mathbf{K}_{\mathrm{f,f}}^{-1}\hat{\mathbf{f}} = \nabla \log p(\mathbf{y}\,|\,\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$, the approximate posterior predictive distribution is

$$\tilde{f}|\,\mathbf{y},\theta,\sigma^2 \sim \mathrm{GP}\left(m_{\mathrm{p}}(\tilde{\mathbf{x}}), k_{\mathrm{p}}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}')\right), \tag{57}$$

where the mean and covariance are $m_{\mathrm{p}}(\tilde{\mathbf{x}}) = k(\tilde{\mathbf{x}},\mathbf{X})\nabla \log p(\mathbf{y}\,|\,\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$ and $k_{\mathrm{p}}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}') = k(\tilde{\mathbf{x}},\tilde{\mathbf{x}}') - k(\tilde{\mathbf{x}},\mathbf{X})(\mathbf{K}_{\mathrm{f,f}}+\mathbf{W}^{-1})^{-1}k(\mathbf{X},\tilde{\mathbf{x}}')$. The approximate conditional predictive density of a new observation $\tilde{y}_i$ can now be evaluated, for example, with Monte Carlo or quadrature integration over each $\tilde{f}_i$ separately

$$p(\tilde{y}_i|\,\mathbf{y},\theta,\gamma) \approx \int p(\tilde{y}_i|\tilde{f}_i,\gamma)q(\tilde{f}_i|\,\mathbf{y},\theta,\gamma)d\tilde{f}_i. \tag{58}$$

Other options for analytic approximation include, for example, expectation propagation (EP) algorithm and variational Bayes (VB) approximations (a good review is provided by Bishop, 2006) which produce Gaussian approximations but their parameterizations may be different from the parameterization of the Laplace approximation. INLA Rue et al. (2009) and extensions of the Laplace approximation, EP and VB produce approximations where the shape of the approximating distributions for each, $f(\mathbf{s}_i)$ are corrected from Gaussian to better approximate the true posterior $p(f(\mathbf{s}_i)|\,\mathbf{y},\theta,\gamma)$. The Gaussian approximation can be justified if the conditional posterior is unimodal, which it is if the likelihood is log concave, and there is enough data so that the posterior will be close to Gaussian. However, invoking the central limit theorem with GP models is not straightforward since the number of observations may grow either alongside the latent variables or per latent variable. The central limit theorem may apply in the increase alongside latent variables case as well if the effective number of latent variables remains small compared to the number of observations. The goodness of the Gaussian approximation is well discussed, for example, by Nickisch and Rasmussen (2008); Rue et al. (2009); Vanhatalo et al. (2010). A pragmatic justification for using Gaussian approximation is that many times it suffices to approximate well the mean and variance of the latent function. These, on the other hand, fully define Gaussian distribution and one can approximate the integrals over $\tilde{f}_i$ by using the Gaussian form for its conditional posterior.

## 5.4 Marginalization over hyperparameters

### 5.4.1 Maximum a posterior estimate of hyperparameters

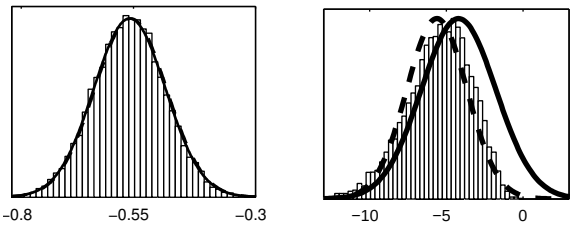
One option to approximate the integral over $p(\theta,\gamma|\,\mathbf{y})$ is to give the hyperparameters a point estimate such as the maximum a posterior (MAP) estimate

$$\{\hat{\theta},\hat{\gamma}\} = \arg\max_{\theta,\gamma} p(\theta,\gamma|\,\mathbf{y}) = \arg\max_{\theta,\gamma} \left[\log p(\mathcal{D}|\theta,\gamma) + \log p(\theta,\gamma)\right]. \tag{59}$$

In this approximation, the hyperparameter values are given a point mass one at the posterior mode, and, for example, the marginal posterior of latent variables is approximated as $p(\mathbf{f}\,|\,\mathbf{y}) \approx p(\mathbf{f}\,|\,\mathbf{y},\hat{\theta},\hat{\gamma})$ (the other posterior marginals come analogously). Alternatively the hyperparameter optimization can be interpreted a model selection over a model family indexed by continuous parameter $\vartheta = [\theta^{\mathrm{T}},\gamma^{\mathrm{T}}]^{\mathrm{T}}$ (Rasmussen and Williams, 2006).

For the MAP estimate one needs to evaluate the log marginal likelihood. In Gaussian case this is straightforward since it has an analytic solution (see equation (**??**)),

$$\log p(\mathbf{y}\,|\theta,\sigma) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\,\mathbf{K}_{\mathrm{f,f}}+\sigma^2\mathbf{I}| - \frac{1}{2}\,\mathbf{y}^{\mathrm{T}}(\mathbf{K}_{\mathrm{f,f}}+\sigma^2\mathbf{I})^{-1}\,\mathbf{y}, \tag{60}$$




(a) Disease mapping (b) classification

Figure 4: Illustration of the Laplace approximation (solid line), EP (dashed line) and MCMC (histogram) for the conditional posterior of a latent variable $p(f_i|\,\mathbf{y},\theta)$ in two applications. On the left, a disease mapping problem with Poisson observation model (Vanhatalo et al., 2010) where the Gaussian approximation works well. On the right, a classification problem with probit likelihood where the posterior is skewed and the Gaussian approximation is clearly a compromise but stillpractically useful. It should also be noted that EP approximates the mean and variance better than the Laplace approximation in this case also.

The log marginal likelihood, and thus also the log posterior, is differentiable with respect to the hyperparameters, which allows a gradient based optimization which can be done e.g. with STAN.

If the observation model is not Gaussian the marginal likelihood needs to be approximated. The Laplace approximation to the marginal likelihood is constructed, for example, by writing

$$p(\mathbf{y}\,|\theta,\gamma) = \int p(\mathbf{y}|\,\mathbf{f},\gamma)p(\mathbf{f}\,|\theta)d\,\mathbf{f} = \int \exp(g(\mathbf{f}))d\,\mathbf{f}, \tag{61}$$

and making a second order Taylor expansion of $g(\mathbf{f})$ around $\hat{\mathbf{f}}$. This gives a Gaussian integral over $\mathbf{f}$ multiplied by a constant, and results in the approximation

$$\log p(\mathbf{y}\,|\theta,\gamma) \approx \log q(\mathbf{y}\,|\theta,\gamma) \propto \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\hat{\mathbf{f}}^{\mathrm{T}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\hat{\mathbf{f}} - \frac{1}{2}\log|\mathbf{B}|, \tag{62}$$

where $|\mathbf{B}| = |I + \mathbf{W}^{1/2}\,\mathbf{K}_{\mathrm{f,f}}\,\mathbf{W}^{1/2}|$. This is the same approximation as the Gaussian approximation by Rue et al. (2009) derived from $p(\mathbf{y},\mathbf{f}\,.\theta,\gamma)/q(\mathbf{f}\,|\mathcal{D},\theta,\gamma)|_{\mathbf{f}=\hat{\mathbf{f}}}$, where the denominator is the Laplace approximation in equation (57) (see also Tierney and Kadane, 1986). The gradients of the approximate log marginal likelihood (62) can be computed analytically, which enables the use of gradient based optimization with Laplace approximation (Rasmussen and Williams, 2006). However, in this course we will not consider this.

The advantage of MAP estimate is that it is relatively easy and fast to evaluate. Optimization algorithms need typically at maximum few tens of optimization steps to find the mode whereas MCMC requires typically hundreds of iteratons to reach convergence and enough samples to approximate the posterior. The drawback, however, is that MAP underestimates the uncertainty in hyperparameters.

### 5.4.2 Monte Carlo integration

Monte Carlo integration is one of the standard choices to conduct the inference for the hyperparameters. A full Monte Carlo estimate for marginal posterior of the latent variables,

$p(\mathbf{f} \,|\, \mathbf{y})$ is obtained by running MCMC for all the parameters in the model, $\mathbf{f}, \theta, \gamma$. That is, we sample both the hyperparameters and the latent variables and estimate the needed posterior statistics by sample estimates or by histograms (Neal, 1997; Diggle et al., 1998). Sampling both, the hyperparameters and latent variables, is usually awfully slow since there is a strong correlation between them. This slows the convergence and mixing of the Markov chain (Vanhatalo and Vehtari, 2007; Vanhatalo et al., 2010). Sampling from the (approximate) marginal, $p(\theta|\, \mathbf{y})$, is a much easier task since the parameter space is smaller and correlations are not so high. Tuning the sampler parameters is also the harder the more parameters are sampled.

Traditional approach to sample from the joint posterior $p(\mathbf{f}, \theta, \gamma|\, \mathbf{y})$ is to conduct the sampling in Gibbs style so that we sample latent variables from $p(\mathbf{f}^{(i)} \,|\, \mathbf{y}, \theta^{(i-1)}, \gamma^{(i-1)})$ and after that the hyperparameters from $p(\theta^{(i)}, \gamma^{(i)}|\, \mathbf{y}, \mathbf{f}^{(i)})$. This approach allows different samplers for the hyperparameters and latent variables which may be beneficial in some cases (Murray et al., 2010; Jarno Vanhatalo, 2013). However, in some cases improvement might be obtained by sampling directly from the joint posterior (Girolami and Calderhead, 2011). STAN utilizes the former approach.

### 5.4.3 Other options for marginalizing over hyperparameters

Section 5.4.1 treated methods to calculate exactly (the Gaussian case) or approximately (Laplace approximation) the marginal posterior $p(\theta, \gamma|\, \mathbf{y})$ up to the normalization constant. There the unnormalized posterior was used for optimizing the hyperparameters but it can also be used for exploring the posterior for purposes of numerical integration with a finite sum, such as

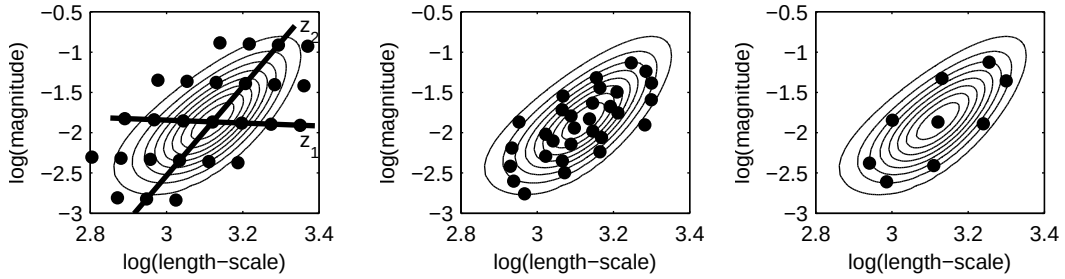
$$p(\mathbf{f} \,|\mathcal{D}) \approx \sum_{i=1}^{M} p(\mathbf{f} \,|\, \mathbf{y}, \vartheta_i) p(\vartheta_i|\, \mathbf{y}) \Delta_i. \tag{63}$$

Here $\vartheta = [\theta^{\mathrm{T}}, \gamma^{\mathrm{T}}]^{\mathrm{T}}$ and $\Delta_i$ denotes the area weight appointed to an evaluation point $\vartheta_i$. Thus, the latent variable posterior is a mixture of Gaussians. The other marginal posteriors are approximated similarly with mixture distributions.

(Markov chain) Monte Carlo is one example of the numerical integration of type (63) where the weigths are $\Delta_i = 1/M$. Other option could be, for example, grid integration where the evaluation points are set into a regular grid. The construction of the grid is started from the posterior mode $\hat{\vartheta}$, and continued so that the bulk of the posterior mass is included in the integration. If the grid points are set evenly, the area weights $\Delta_i$ are equal. In practice, the construction of the grid is aided by the information about the Hessian of $\log p(\vartheta|\mathcal{D})$ at the mode, which would be the inverse covariance matrix for $\vartheta$ if the density were Gaussian. This approximate covariance is used to select the exploration directions and step sizes as illustrated in Figure 5(a) and discussed by Rue et al. (2009).

The numerical integration using the grid search is feasible only for a small number of hyperparameters since the number of grid points grows exponentially with the dimension of the hyperparameter space $d$.

In order to decrease the number of grid points Rue et al. (2009) suggest a central composite design (CCD) for choosing the representative points from the posterior of the hyperparameters when the dimensionality $d$ is moderate or high. In this setting, the integration is considered as a quadratic design problem in a $d$ dimensional space with the aim in finding points that allow for estimating the curvature of the posterior distribution around the mode. The design used by Rue et al. (2009) is the fractional factorial design




(a) Grid based integration. (b) Monte Carlo integration (c) Central composite design

Figure 5: Illustration of the grid based, the Monte Carlo and the central composite design integration over the logarithm of the hyperparameters. The contour shows the posterior density $q(\log(\vartheta)|\mathcal{D})$ and the integration points are marked with dots. The left figure shows also the vectors $\mathbf{z}$ along which the points are searched in the grid integration and central composite desing. The integration is conducted over $q(\log(\vartheta)|\mathcal{D})$ rather than $q(\vartheta|\mathcal{D})$ since the former is closer to Gaussian. (From Vanhatalo et al. (2010).)

augmented with a center point and a group of $2d$ star points. In this setting, the design points are all on the surface of a $d$-dimensional sphere and the star points consist of $2d$ points along each axis. This is illustrated in Figure 5(c). The number of the design points grows very moderately and, for example, for $d = 6$ one needs only 45 points. The CCD integration can be summarized with the equation (63) where the weights are evaluated as described by Rue et al. (2009) and Vanhatalo et al. (2010).

The CCD integration speeds up the integration considerably. The accuracy is between the MAP estimate and the full integration with grid search or Monte Carlo. For example, Rue et al. (2009), **?** and Vanhatalo et al. (2010) report good results with this integration scheme with moderate dimensional parameter space ($< 10$).

## 5.5 Summary of the inference methods

The methods treated in this chapter can be arranged in an increasing order of accuracy and computational time. The choice of the method is then a compromise between these two attributes. The inference is the fastest when using MAP estimate for the hyperparameters and Gaussian function for the conditional posterior. With a Gaussian observation model, the Gaussian conditional distribution is exact and the only source of imprecision is the point estimate for the hyperparameters. If the observation model is other than Gaussian, the conditional distribution is an approximation, whose quality depends on, how close to Gaussian the real conditional posterior is, and how well the mean and variance are approximated. The form of the real posterior depends on many things for which reason the Gaussian approximation has to be assessed independently for every data. Methods for assessing the Gaussian approximation are discussed, for example, by Rue et al. (2009) and Vanhatalo et al. (2010). Tierney and Kadane (1986) provide asymptotic results for the accuracy of the Laplace approximation and Nickisch and Rasmussen (2008) give extensive comparison between different Gaussian approximations in classification problems. The Laplace approximation is faster than EP but EP approximates better the posterior mean

and variance. For example, in classification, this is crucial since the posterior of the latent variables is rather far from normal, as illustrated in Figure 4(b) (see also **?**Nickisch and Rasmussen, 2008). However, in many cases Laplace approximation gives, at a practical level, as good results as full MCMC or EP Vanhatalo et al. (2010) (see also Figure 4(a)). At the expense of computational time, the approximation to the marginal posterior of a latent variable could be improved by evaluating correction terms for the EP approximation (Paquet et al., 2009; Cseke and Heskes, 2010) or by improving the Laplace approximation to marginals (Tierney and Kadane, 1986; Rue et al., 2009).

A golden standard for the posterior distributions can be obtained by an extensive MCMC - given it converges and mixes well. If a MAP estimate for the hyperparameters is considered adequate but we want to sample the posterior of the latent variables we can use Laplace approximation to locate MAP after which the sampling of the latent variables can be performed efficiently with STAN aided by the variable transformation (Christensen et al., 2006). Even if the Laplace approximation and EP lacked in accuracy for the conditional posterior they may approximate the marginal likelihood well. The accuracy of the Laplace approximation depends on the effective number of latent variables and it is usually more accurate for data sets with many observations per input location (Rue et al., 2009). EP has been shown to approximate the marginal likelihood rather accurately in many problems (**?**Nickisch and Rasmussen, 2008), whereas the Laplace approximation gives somewhat less accurate approximations to the marginal likelihood. This suggests that EP's approximation to the marginal likelihood is more reliable. However, since the predictive inference on $f(\mathbf{s})$ is rather insensitive to small changes in the hyperparameters around their MAP Laplace approximation and other analytic approximations to the marginal likelihood are often justified. The identifiability of the hyperparameters is well treated, for example, by Diggle et al. (1998), Zhang (2004) and Diggle and Ribeiro (2007).

# 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.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science +Business Media, LLC.

Christensen, O. F., Roberts, G. O., and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15:1–17.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Inc.

Csató, L. and Opper, M. (2002). Sparse online Gaussian processes. *Neural Computation*, 14(3):641–669.

Cseke, B. and Heskes, T. (2010). Improving posterior marginal approximations in latent Gaussian models. *JMLR Workshop and Conference Proceedings*, 9:121–128.

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

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(3):299–350.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

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

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P., editors (2010). *Handbook of Spatial Statistics (eds.)*. Chapman & Hall/CRC.

Gelfand, A. E., Jr, J. A. S., Wu, S., Latimer, A., Lewis, P. O., Rebelo, A. G., and Holder, M. (2006). Explaining species distribution patterns through hierarchical modelling. *Bayesian Analysis*, 1(1):41–92.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC, third edition.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73(2):123–214.

Hartmann, M., Hosack, G. R., Hillary, R. M., and Vanhatalo, J. (2017). Gaussian processs framework for temporal dependence and discrepancy functions in ricker-type population growth models. *Annals of Applied Statistics*, in press.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Jarno Vanhatalo, Jaakko Riihimäki, J. H. P. J. V. T. A. V. (2013). Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179.

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.

Murray, I., Adams, R. P., and MacKay, D. J. (2010). Elliptical slice sampling. *JMLR Workshop and Conference proceedings: The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 9:541–548.

Neal, R. (1998). Regression and classification using Gaussian process priors. In Bernardo, J. M., Berger, J. O., David, A. P., and Smith, A. P. M., editors, *Bayesian Statistics 6*, pages 475–501. Oxford University Press.

Neal, R. (2011). Mcmc using hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 116–162. Chapman and Hall/CRC.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer.

Neal, R. M. (1997). Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, Dept. of statistics and Dept. of Computer Science, University of Toronto.

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078.

Paquet, U., Winther, O., and Opper, M. (2009). Perturbation corrections in approximate inference: Mixture modelling applications. *Journal of Machine Learning Research*, 10:1263–1304.

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

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, second edition.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal statistical Society B*, 71(2):1–35.

Simpson, D. P., Rue, H., Martins, T. G., Riebler, A., and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *ArXiv e-prints: 1403.4630*.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607.

Vanhatalo, J. and Vehtari, A. (2007). Sparse log gaussian processes via mcmc for spatial epidemiology. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1:73–89.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.