Random fields and spatial priors

Janne Huttunen

UEF

November 10, 2015

- Next we will give short introduction to stochastic processes and random fields
- Motivation: stochastic processes and random fields can be used for (for example):
  - **Dynamical or nonstationary inverse problems:** unknown and other quantities are temporally varying (functions of time).
  - **Spatial priors:** prior models for distributed unknown quantities (unknowns are functions of the spatial coordinate x).
- Spatial priors are presented at the end of this presentation
- Dynamical inverse problems and different solution methods are the subject of the rest of part 2 (lectures L17->).

Stochastic process

A stochastic process is a parametrized collection of random variables: $\{X(s)\}_{s \in \mathcal{D}}$ where $\mathcal{D}$ is a set.

Usual terminology:

- **Discrete process**: $\mathcal{D} = \{0, 1, 2, \ldots\}$ (or some other discrete set)
- **(Continuous time) stochastic process**: $\mathcal{D}$ is a subset of real line $\mathbb{R}$ and s is usually time: e.g. $\{X(t)\}_{t \geq 0}$
- **Random field**: $\mathcal{D}$ is a subset of $\mathbb{R}^d$ ($d = 1, 2, \ldots$) and the parameter s is a spatial coordinate x. Example: $\{X(x)\}_{x \in S_1}$, where $S_R = \{x \in \mathbb{R}^3 : \|x\| = R\}$ is a sphere in $\mathbb{R}^3$ (typical for modelling processes on the surface of Earth e.g. in climate)
- **Space-time process**: e.g. $\{X(t, x) : t \geq 0, x \in D\}$, $D \subset \mathbb{R}^d$

Commonly the set $\mathcal{D}$ is not specified in the notation if it is known from the context. The brackets are also often omitted and the process is simply denoted by $X(s)$ or $X(x)$. Notations X_k and X_t are also common for discrete and continuous time processes.

How to think stochastic processes?

- In probability theory, random variables are defined as functions of $\omega \in \Omega$. Similarly stochastic processes can be considered as functions of s and ω: $X(\omega, s)$ or $X(s, \omega)$
- Stochastic processes and random fields can also be thought as function valued random variables:
  - Random variables: realizations are real numbers: $X(\omega) \in \mathbb{R}$ when ω is fixed
  - Random vectors: realizations are vectors: $X(\omega) \in \mathbb{R}^n$ when ω is fixed
  - Stochastic processes and random fields: when ω is fixed, $X(\omega)$ is a function of the parameter s, that is, $s \rightarrow X(\omega, s)$,

Simple example about a stochastic process

- Specify six functions

$$
\begin{aligned}
f^1(t) &= t \\
f^2(t) &= \sin(t), \\
f^3(t) &= \log(t+1), \\
f^4(t) &= t^2 - t, \\
f^5(t) &= \cos(t), \\
f^6(t) &= 1.
\end{aligned}
$$

Let $\omega \in \{1, \ldots, 6\}$ be an outcome of throwing a dice. We can specify a stochastic process by $X(\omega, t) = f^\omega(t)$.

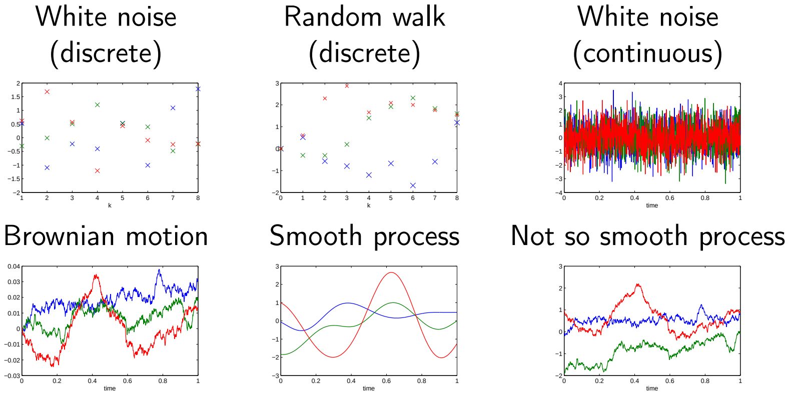


Figure: Examples of stochastic processes: red, green and blue are three different realizations of the process.

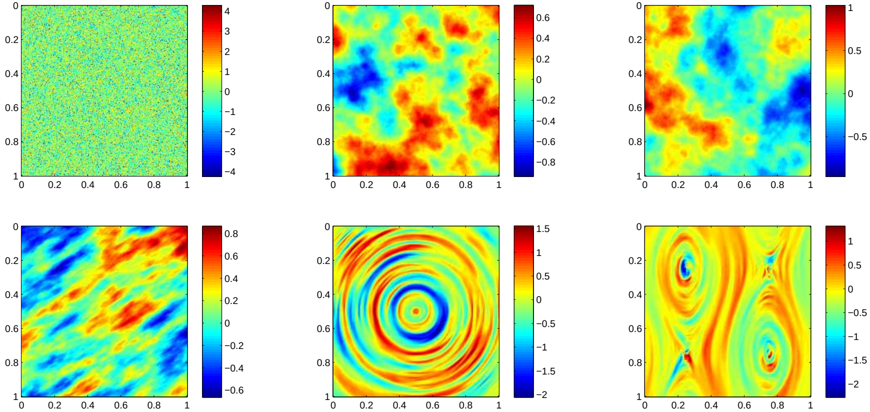


Figure: Realizations from different random fields (all Gaussian).

Basic concepts

- *The mean function*: $\mu(s) = \mathbb{E}[X(s)]$, $s \in \mathcal{D}$.

- *The covariance function*:

$$C(s, s') = \mathrm{cov}(X(s), X(s')) = \mathbb{E}\left[(X(s) - \mu(s))(X(s') - \mu(s'))\right]$$

  for $s, s' \in \mathcal{D}$.

- Note: $\mathrm{var}(X(s)) = \mathbb{E}\left[(X(s) - \mu(s))^2\right] = C(s, s)$.

- *Finite dimensional joint-distributions*: let $s_1, \ldots, s_n$ be a points in $\mathcal{D}$. The finite dimensional joint-distributions of a process $X(s)$ are given by

$$F_{s_1,\ldots,s_n}(y_1, \ldots, y_n) = \mathbb{P}(X(s_1) \leq y_1, \ldots, X(s_n) \leq y_n)$$

  for $y_1, \ldots, y_n \in \mathbb{R}$.

Stationary process

A process $X(s)$ is called (strictly) *stationary* if for every set of points $s_1, \ldots, s_n$ in $\mathcal{D}$, the finite dimensional joint-distributions are shift-invariant:

$$F_{s_1+h,\ldots,s_n+h}(y_1, \ldots, y_n) = F_{s_1,\ldots,s_n}(y_1, \ldots, y_n)$$

for all $h \in \mathcal{D}$ such that $s_i + h \in \mathcal{D}$.

Weakly stationary process

A process $X(s)$ is called *weakly stationary* if for all s, s' and h:

$$\mu(s+h) = \mu(s) \qquad C(s+h, s'+h) = C(s, s') = C(s-s').$$

- In other words: weakly stationary process has the mean function which is a constant and the covariance is a only function of $\tau = s - s'$, $C(\tau)$
- A strictly stationary process is also weakly stationary, opposite is not always true.

Isotriphic process

A process $X(s)$ is called *isotrophic* if for all s, s':

$$C(s, s') = C(\|s - s'\|).$$

- In other words, the process is isotropic if the covariance function can be expresses as a function of the distance $r = \|s - s'\|$ (no directional dependency).
- Processes that are not isotrophic (i.e. the covariance depends on the direction) are called as anisotrophic.

Gaussian processes

The process is called Gaussian if $(X(s_1), \ldots, X(s_n))$ is a Gaussian random vector for all sets of points $s_1, \ldots, s_n \in \mathcal{D}$.

- In other words, the process is Gaussian if all finite dimensional joint-distributions are Gaussian
- Gaussian processes are completely determined by the mean and covariance function
- Weakly stationary Gaussian processes are also strictly stationary
- GMRF: Gaussian Markov random field

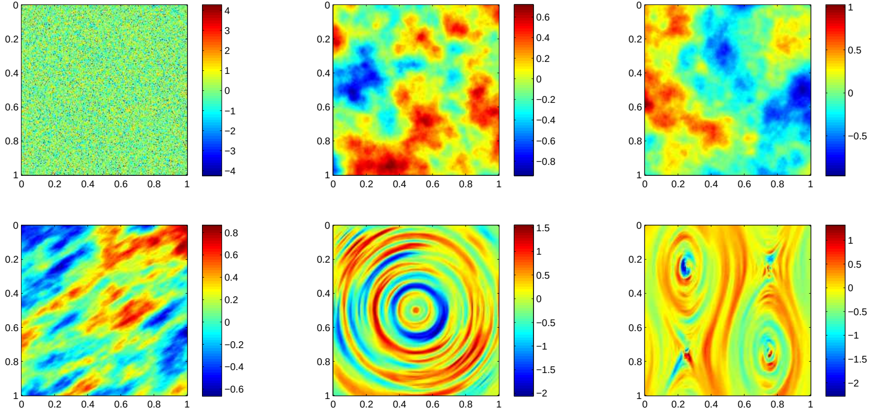


Figure: Realizations from Gaussian random fields. Top row: white noise (left) and two realizations of a same isotrophic random field (middle and right). Bottom row: a realization of an anisotrophic random field (left), and two different nonstationary random fields (middle, right).

- The following Markov property is useful with Kalman filters (dynamic inverse problems):

Markov property for a discrete process

A discrete process X_k has so called Markov property, if the conditional probability distribution of X_k given all states X_s, $s < k$, equals to the conditional probability distribution of X_k given the previous state X_{k-1}:

$$p(x_k | x_s, s < k) = p(x_k | x_{k-1})$$

In other words, if X_{k-1} is known, the knowledge of $X_{k-2}, X_{k-3}, \ldots$ does provide any additional information about the current state X_k

- Markov property can also be given for continuous processes and random fields (omitted in this course).

Practical use of Gaussian random fields

- Often unknown quantities are modelled as Gaussian variables since Gaussian distributions leads to computational efficient problems, or we just do not know any better distribution for the variable
- From now on we only consider Gaussian random fields
- When we consider Gaussian random fields, we only need to think about the mean and covariance function

The mean function

- The mean can be chosen based on the prior information related to the problem.
- Often the mean function is written a sum of functions basis ϕ_i functions:

$$\mu(s) = \sum_i \theta_i \phi_i(s)$$

  for which the coefficients θ_i are determined based on some sort of data (e.g. hyper parameters in inverse problems).
- The form of basis functions is chosen based on the application: e.g. piecewise linear functions, polynomials, sin and cos functions (wave propagation problems).

The covariance function

- In principle the form of the covariance function could be chosen to be a function of s and s' which can also include some parameters (e.g. variance and scaling parameters) that are determined based on data
- However the covariance function should satisfy some requirements implied by the definition
- Furthermore, some attention should be paid to check that the random field will have preferred continuity and smoothness properties

Requirements for covariance functions

- First of all, C has to be symmetric: $C(s, s') = C(s', s)$ for all $s, s' \in \mathcal{D}$. For stationary process: $C(\tau) = C(-\tau)$ where $\tau = s - s'$.
- Furthermore, consider a set of points $\{s_i \in \mathcal{D} \,:\, i = 1, \ldots, n\}$ and let K be a $n \times n$ matrix such that the elements are

$$K_{ij} = C(s_i, s_j), \quad i, j = 1, \ldots, n$$

- If C is the covariance function of a process X, the matrix K is the covariance matrix of the n-dimensional random vector $(X(s_1), \ldots, X(s_n))$.
- All covariance matrices should be positive semidefinite: $x^T K x \geq 0$ for all vectors $x \in \mathbb{R}^n$
- Therefore the covariance function has to be positive semidefinite: for all set of points $\{s_1, \ldots, s_n\} \subset \mathcal{D}$, the matrix K given above is positive semidefinite.

Continuity and smoothness

- The process is said to be *continuous in mean square* at s_* if $\mathbb{E}\left[|X(s_k) - X(s_*)|^2\right] \to 0$ for all sequences $s_k \to s_*$. Mean square derivatives are defined similarly using fractions $\frac{X(s_k)-X(s_*)}{s_k-s_*}$.
- A stochastic process X is continuous in mean square at s_* if and only if $C(s,s')$ is continuous at $s = s' = s_*$. For stationary X, it is sufficient to check continuity of $C(\tau)$ at $\tau = 0$.
- The derivates of C determines the smoothness of X: if $\frac{\partial^2 C(s,s')}{\partial s_i \partial s_i'}$ exists and is finite, $\frac{\partial X}{\partial s_i}$ exists (in mean square sense) and its covariance function is $\frac{\partial^2 C(s,s')}{\partial s_i \partial s_i'}$. Higher order derivatives similarly.
- Stationary X: if $\frac{\partial^{2k} C(\tau)}{\partial d^{2k}}$ exists and is finite at $\tau = 0$, the derivative $\frac{\partial^k X}{\partial s^k}$ exists (in mean square sense).

Summary (what should be remembered from the slide)

The continuity and smoothness of X are determined by the continuity and smoothness of the covariance function at $s = s'$ at $\tau = 0$.

Examples of covariance functions

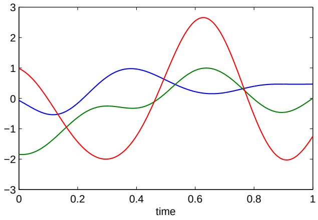
Squared exponential covariance function

Squared exponential covariance function:

$$C(\tau) = \exp\left(-\frac{\|\tau\|^2}{2\ell^2}\right)$$

where $\ell > 0$ is scaling parameter often called as *characteristic length-scale*.

- Simple form and very widely used
- $C(\tau)$ is infinitely differentiable
  $\Rightarrow X$ has mean square derivates of all orders and thus very smooth
- Such very strong smoothness properties may be unrealistic in many applications



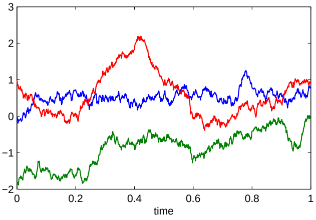
Examples of covariance functions

Exponential covariance function

Exponential covariance function is of the form

$$C(\tau) = \exp\left(-\frac{\|\tau\|}{\ell}\right)$$

- C continuous but not differentiable at $\tau = 0 \Rightarrow$ the process is continuous in mean square, but not differentiable
- May be too rough process for many applications (especially if smoothness is preferred)



Examples of covariance functions

Mátern class of covariance functions

Mátern class of covariance functions is given by

$$C_\nu(\tau) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,\|\tau\|}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}\,\|\tau\|}{\ell} \right)$$

where ν and ℓ are positive parameters and K_ν is the modified Bessel function of second order.

- The parameter ν determines the smoothness properties of the process: the process is k'th times mean square differentiable if and only if $\nu > k$.
- Furthermore, the limit $\nu \to \infty$ gives the squared exponential covariance function, $\nu = \frac{1}{2}$ gives the exponential covariance function.

Examples of covariance functions

- For other half integers $\nu = p + \frac{1}{2}$ ($p > 0$), the Matérn covariance functions are products of an exponential function and a polynomial of order p:

$$
\begin{aligned}
C_{\nu=\frac{3}{2}}(\tau) &= \left(1 + \frac{\sqrt{3}\,\|\tau\|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\,\|\tau\|}{\ell}\right) \\
C_{\nu=\frac{5}{2}}(\tau) &= \left(1 + \frac{\sqrt{5}\,\|\tau\|}{\ell} + \frac{5\,\|\tau\|^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}\,\|\tau\|}{\ell}\right)
\end{aligned}
$$

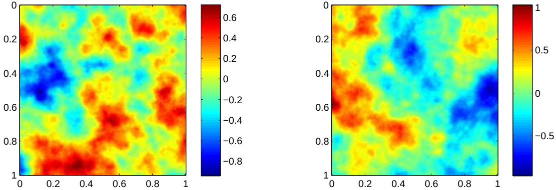


Figure: Two realizations from the Matérn class with $\nu = 3$

- The covariances functions can be also formed as combination of several covariance function:
  - The sum of two covariance functions is also a valid covariance function (the covariance function of $X_1(s) + X_2(s)$ of when X_1 and X_2 are independent)
  - The product of two covariance functions is also a valid covariance function (the covariance function of $X_1(s)X_2(s)$ when X_1 and X_2 are independent). Thus also $C(s, s')^p$ is a valid covariance function.
  - Let $a(s)$ be a deterministic function. Then the covariance function of $Y(s) = a(s)X(s)$ is $a(s)C(s, s')a(s')$ if C is the covariance function of the process $X(s)$.

Spatially varying variance

- All of the above covariance functions are stationary and isotropic, and normalized such that $C(0) = \mathrm{var}(X(s)) = 1$.
- Sometimes we may want more flexibility and, for example, choose the variance as a function of s, $\sigma(s)$. Then we can write e.g.:

$$X(s) = \mu(s) + \sigma(s)X'(s)$$

  and consider the construction of $X'(s)$ as a stationary process.
- If the covariance of $X'(s)$ is $C'(s, s')$, the covariance of X is $C(s, s') = \sigma(s)\sigma(s')C'(s, s')$ (as in the previous slide)

Anisotrophic covariance functions

- The above correlation functions can be modified for anisotrophical cases (correlation different to different directions) easily.
- We consider only stationary two–dimensional case, other dimensions are similar
- The previous isotrophic correlation functions include the term $\|\tau\|/\ell = \sqrt{\frac{\tau_x^2}{\ell^2} + \frac{\tau_y^2}{\ell^2}}$ where $\tau = (\tau_x, \tau_y)$
- To introduce different characteristic length-scales to the x and y-direction, we can replace this terms with $\sqrt{\frac{\tau_x^2}{\ell_x^2} + \frac{\tau_y^2}{\ell_y^2}}$
- For example, anisotrophic squared exponential covariance function:

$$C(\tau) = \exp\left\{-\frac{1}{2}\left(\frac{\tau_x^2}{\ell_x^2} + \frac{\tau_y^2}{\ell_y^2}\right)\right\}$$

Anisotrophic covariance functions

- Other directions can be handled using coordinate transformations
- Note that

$$\frac{\tau_x^2}{\ell_x^2} + \frac{\tau_y^2}{\ell_y^2} = \tau^{\mathrm{T}} \Lambda \tau, \quad \Lambda = \mathrm{diag}(\ell_x^{-2}, \ell_y^{-2}).$$

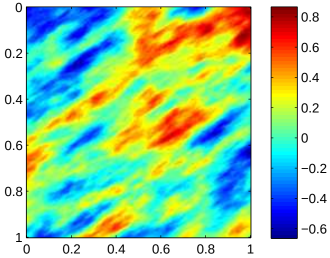
- We apply an coordinate transform matrix C and replace the term with

$$\tau^{\mathrm{T}} C \Lambda C^{\mathrm{T}} \tau$$

- E.g. C can be a rotation matrix

$$C = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

which rotates points in xy-plane counter-clockwise with an angle θ



Example: stochastic interpolation (Kriging)

- Random fields can be applied for interpolation of a function as follows.

- For example, we have an unknown function $X : [0, 1] \mapsto \mathbb{R}$.
- We have observations of X at a given set of points $x_1, \ldots, x_n \in [0, 1]$: $y_i = X(x_i)$, $i = 1, \ldots, n$.
- We want to estimate the value of X in an arbitrary point $x_0 \in [0, 1]$ (interpolation).

Example: stochastic interpolation (Kriging)

- We model X as a Gaussian random field.
- In this example, we choose $\mu(s) = 0$ and C is the Matérn covariance function with $\nu = 3/2$
- Define random variables $\mathbf{X} = X(x_0)$ and $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}} = (X(x_1), \ldots, X(x_n))^{\mathrm{T}}$.
- Since $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian random variables, the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ is Gaussian: $\mathcal{N}(\hat{\mathbf{X}}, \sigma^2_{\mathbf{X}|\mathbf{Y}})$ where

$$\begin{aligned}\hat{\mathbf{X}} &= \bar{\mathbf{X}} + \Gamma_{\mathbf{XY}}\Gamma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \bar{\mathbf{Y}}) \\ \sigma^2_{\mathbf{X}|\mathbf{Y}} &= \sigma^2_{\mathbf{X}} - \Gamma_{\mathbf{XY}}\Gamma_{\mathbf{Y}}^{-1}\Gamma_{\mathbf{XY}}^{\mathrm{T}}\end{aligned}$$

  (see the preliminaries PDF)
- The above equations gives our solution: the mean $\hat{\mathbf{X}}$ gives an estimate for $X(x_0)$ and $\sigma^2_{\mathbf{X}|\mathbf{Y}}$ is an estimate of its uncertainty (variance). For interpolation, we can vary x_0.

Example: stochastic interpolation (Kriging)

- Before we can use the above equations, we need to calculate the expectations of $\mathbf{X}$ and $\mathbf{Y}$, the variance $\sigma^2_{\mathbf{X}}$, the covariance $\Gamma_{\mathbf{Y}}$ and the cross-covariance $\Gamma_{\mathbf{XY}}$
- The expectations are given by the mean function: $\bar{\mathbf{X}} = \mu(x_0) = 0$ and $\bar{\mathbf{Y}} = (\mu(x_1), \ldots, \mu(x_n))^{\mathrm{T}} = 0$
- The variance of $\mathbf{X}$ is $\sigma^2_X = C(x_0, x_0)$
- The covariance of $\mathbf{Y}$ is the matrix Γ_Y which elements are $C(x_i, x_j)$ $(i, j = 1, \ldots, n)$
- The cross-covariance of $\mathbf{X}$ and $\mathbf{Y}$ is $\Gamma_{\mathbf{XY}} = (C(x_0, x_1), \ldots, C(x_0, x_n))$
- Note: it is easy to expand the approach for noise-corrupted measurements $y_i = X(x_i) + \epsilon_i$, where $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon I)$ independent of X. In this case $\Gamma_{\mathbf{Y}}$ in the above formulae is replaced with $\Gamma_{\mathbf{Y}} + \sigma^2_\epsilon I$.

Example: stochastic interpolation

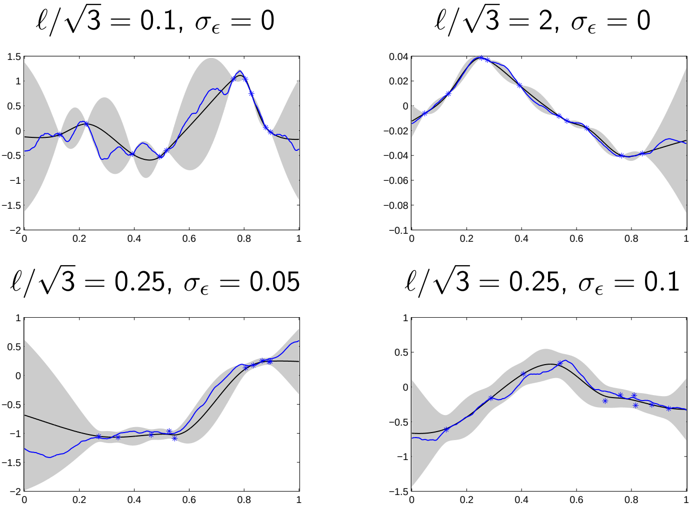


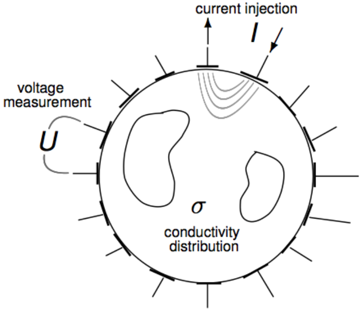
Figure: Stochastic interpolation: the blue line is the true function f and the black line is the estimate. The gray band corresponds to 2xS.D. error limits. Starts

Spatial priors

- **Spatial priors:** prior models inverse problems in which unknowns depend on the spatial coordinate x (e.g. heterogeneous variables).
  - Unknown quantities are modelled as random fields
  - The prior distribution is given by the distribution of random field.
  - If can be assumed to be Gaussian:
    $\Rightarrow$ specify the mean and covariance function

Example: Electrical impedance tomography (EIT)

- Unknown (electric) conductivity distribution $\sigma(x)$ is a heterogenous variable
- We want to determine σ (e.g. tomographic imaging)
- Electrodes on boundary
- Inject electric currents I $\rightarrow$ measure voltages U
- Problem: reconstruct σ from (I, U) information



Spatial priors

- Consider an inverse problem in which unknown $X(x)$ is a spatially varying function (distributed parameter, a heterogeneous variable)
- $X(x)$ can be modelled as a random field
- To specify a Gaussian prior: specify mean function $\mu(x)$ and covariance function $C(x, x')$
- The mean $\mu(x)$ is specified based on prior information related to the application
- For the covariance function $C(x, x')$ can be chosen to be, for example, one of the listed previously based on the prior knowlege. For example:
  - expected to be very smooth $\rightarrow$ squared exponential
  - expected to non-smooth $\rightarrow$ exponential
  - Matérn if between those

Spatial priors: practical implementation

- The inverse problem is usually discretized numerically for practical implementation (e.g. finite difference method, finite element method)
- The discretized unknown often represents the unknown $X(x)$ in a grid of points.
- Let $\{x_i, i = 1, \ldots, n\}$ be such grid points.
- Then prior can be chosen as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ where

$$\begin{aligned} \boldsymbol{\mu} &= (\mu(x_1), \ldots, \mu(x_n))^{\mathrm{T}} \\ \boldsymbol{\Gamma}(i,j) &= C(x_i, x_j), \quad i,j = 1, \ldots, n. \end{aligned}$$

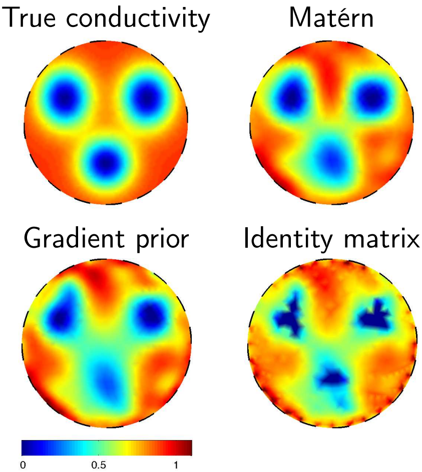
- Sometimes the expected variance of the field can also depend on the spatial variable
- We could specify a non-stationary covariance
- However, it is usually easier to work with stationary covariances and, for example, specify X as

$$X(x) = \mu(x) + \sigma(x)W(x)$$

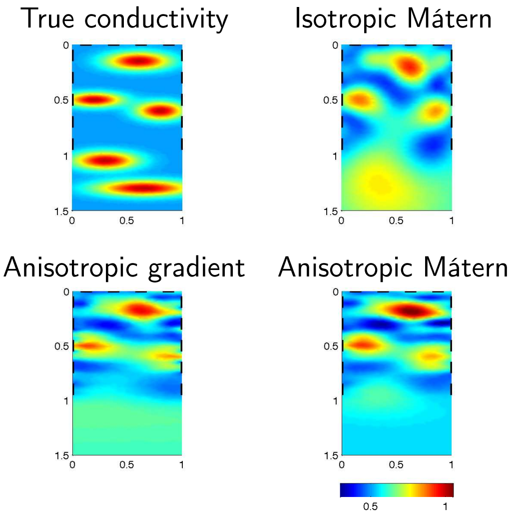
  where $\sigma(x)$ is preferred variance (also chosen based on the problem) and W is a stationary random field (zero mean)
- The stationary covariance is specified for W
- Then $C_X(x, x') = \sigma(x)C(x, x')\sigma(x')$ and

$$\mathbf{\Gamma}(i, j) \; = \; \sigma(x_i)C(x_i, x_j)\sigma(x_j)$$

Electrical impedance tomography in a circular tank



Ground prospecting with anisotropic conductivities



Hierarchical prior models (hyperparameters)

- Priors can include parameters that are not precisely known
- For example: mean, variance, length-scale ℓ
- We can model these as hierarchical prior parameters often called as hyperparameters:
  - Consider such parameters also as unknown in the inverse problems
  - Write a prior model for the hyper parameters
  - Consider both the primary unknown X and the hyper parameters as unknown and estimate it from the data

Example of hyper parameters

We consider an example:

- Assume that the mean is presented using basis functions θ_i

$$\mu(x) = \sum_{i=1}^{p} \gamma_i \phi_i(x)$$

  where γ_i are unknown.
- Assume that the variance σ^2 (assumed to be a constant) and the length-scale ℓ in the covariance function are also unknown
- We denote the vector of hyper parameters by θ:

$$\theta = (\gamma_1, \ldots, \gamma_p, \sigma^2, \ell)$$

Example of hierarchical models

- Discretization: X presented at points $x_1, \ldots, x_n$
- For discretized prior mean: $\mu_X = (\mu(x_1), \ldots, \mu(x_n))^{\mathrm{T}} = \Phi\gamma$ where

$$\Phi = \begin{pmatrix} \phi_1(x_1) & \cdots & \phi_1(x_n) \\ \vdots & \ddots & \vdots \\ \phi_p(x_1) & \cdots & \phi_p(x_n) \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{pmatrix}$$

- The prior model:

$$\pi(X, \theta) = \pi(X|\theta)\pi(\theta)$$

where

$$\pi(X|\theta) \propto e^{-\frac{1}{2}\left[(X-\Phi\gamma)^{\mathrm{T}}\Gamma_X^{-1}(\sigma^2,\ell)(X-\Phi\gamma)+\log\det(\Gamma_X(\sigma^2,\ell))\right]}$$

The log term is due to the normalization constant (which now depends on the unknown hyperparameters and has to be included).

- The hyperprior $\pi(\theta)$ is specified by using prior knowledge/beliefs of hyper parameters.
- For example: $\gamma \sim \mathcal{N}(0, \Gamma_\gamma)$ with known Γ_γ
- Inverses of the variances are often modelled using Gamma distributions:

$$\pi(\sigma^{-2}) = \mathrm{Gamma}(\alpha_\sigma, \beta_\sigma) \quad \text{or} \quad \pi(\sigma^2) = \mathrm{InvGamma}(\alpha_\sigma, \beta_\sigma)$$
$$\Rightarrow \pi(\sigma^2) \propto (\sigma^2)^{-\alpha_\sigma - 1} e^{-\beta_\sigma/\sigma^2} = e^{-\beta_\sigma/\sigma^2 - (\alpha_\sigma + 1)\log \sigma^2}$$

- The scale length parameter can be chosen to follow, for example, Gamma distribution

$$\pi(\ell) \propto \ell^{\alpha_\ell - 1} e^{-\beta_\ell \ell} = e^{-\beta_\ell \ell - (1-\alpha_\ell)\log \ell}$$

- Usually the hyper parameters are assumed to be independent:

$$\pi(\theta) = \pi(\gamma)\pi(\sigma^2)\pi(\ell)$$

- The posterior is $\pi(X,\theta|m) \propto \pi(m|X)\pi(X|\theta)\pi(\theta)$
- If we have an observation model $m = A(x) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \Gamma_\epsilon)$, the posterior for our example is

$$
\begin{aligned}
&-\log \pi(X,\theta|m) \\
&= \frac{1}{2}(m - A(x))^{\mathrm{T}}\Gamma_\epsilon^{-1}(m - A(x)) \\
&\quad + \frac{1}{2}\left(X - \Phi\gamma\right)^{\mathrm{T}} \Gamma_X^{-1}(\sigma^2,\ell)\left(X - \Phi\gamma\right) + \frac{1}{2}\log\det(\Gamma_X(\sigma^2,\ell)) \\
&\quad + \frac{1}{2}\gamma^{\mathrm{T}}\Gamma_\gamma^{-1}\gamma + \beta_\sigma/\sigma^2 + (\alpha_\sigma + 1)\log\sigma^2 + \beta_\ell\ell + (1 - \alpha_\ell)\log\ell
\end{aligned}
$$

- The above function can be minimized using optimization algorithms (e.g. Gauss-Newton) to compute MAP estimate, or use MCMC methods for CM estimates.
- Furthermore, if the posterior is simple, the hyper parameters could perhaps be integrated out to obtain $\pi(X|m)$