# Lecture 1: Introduction

## Gaussian Markov random fields

David Bolin
Chalmers University of Technology
January 19, 2015

# Practical information

**Litterature**:
The course will mostly be based on the book
*Gaussian Markov Random Fields: Theory and Applications*
by Håvard Rue and Leonhard Held

Additional articles will be used later on.

**Homepage**:
`http://www.math.chalmers.se/~bodavid/GMRF2015/`

**Schedule**:
We will meet two times each week:
Mondays and Tuesdays (10-12)

Lectures will be in MVL:14

There will be 10 lectures and 4 computer labs

# Examination

There will be two components in the examination:

- Project assignments introduced in the computer labs
- An oral exam at the end of the course

The projects can be done individually or in pairs of two students.

The final oral exam is individual.

The grading scale comprises Fail, (U), Pass (G).

Successful completion of the course will be rewarded by 7.5 hp.

# Three relevant questions

Why take this course when we have had one course on Markov random fields and one course on Gaussian random fields this year?

Why is it a good idea to learn about Gaussian Markov random fields?

What's there to learn? Isn't all just a Gaussian?

# Outline of lectures

Lecture 1 Introduction

Lecture 2 Definitions and basic properties of GMRFs

Lecture 3 Simulation and conditioning

Lecture 4 Numerical methods for sparse matrices

Lecture 5 Intrinsic GMRFs

Lecture 6 MCMC estimation for hierarchical models

Lecture 7 Approximation techniques and INLA

Lecture 8 Stochastic PDEs and FEM

Lecture 9 SPDEs part 2

Lecture 10 Extensions and applications

# Random fields

## Random fields

A random field (or stochastic field), $X(\mathbf{s}, \omega)$, $\mathbf{s} \in \mathcal{D}$, $\omega \in \Omega$, is a random function specified by its finite-dimensional joint distributions

$$F(y_1, \ldots, y_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) = \mathsf{P}(X(\mathbf{s}_1) \leq y_1, \ldots, X(\mathbf{s}_n) \leq y_n)$$

for every finite $n$ and every collection $\mathbf{s}_1, \ldots, \mathbf{s}_n$ of locations in $\mathcal{D}$.

- The set $\mathcal{D}$ is usually a subset of $\mathbb{R}^d$.
- At every location $\mathbf{s} \in \mathcal{D}$, $X(\mathbf{s}, \omega)$ is a random variable where the event $\omega$ lies in some abstract sample space $\Omega$.
- Kolmogorov's existence theorem can be used to ensure that the random field has a valid mathematical specification.
- To simplify the notation, one often writes $X(\mathbf{s})$, removing the dependency on $\omega$ from the notation.

# Gaussian random fields

An important special case is when the random field is Gaussian.

## Gaussian random fields

A Gaussian random field $X(\mathbf{s})$ is defined by a mean function $\mu(\mathbf{s}) = \mathrm{E}(X(\mathbf{s}))$ and a covariance function $C(\mathbf{s}, \mathbf{t}) = \mathrm{Cov}(X(\mathbf{s}), X(\mathbf{t}))$. It has the property that, for every finite collection of points $\{s_1, \ldots, s_p\}$,

$$\mathbf{x} \equiv (X(s_1), \ldots, X(s_p))^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\Sigma_{ij} = C(s_i, s_j)$.

For existence of a Gaussian field with a prescribed mean and covariance it is enough to ensure that $C$ is positive definite.

# Positive definite functions

A function $C(\mathbf{s}, \mathbf{t})$ is positive definite if for any finite set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ in $\mathcal{D}$, the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} C(\mathbf{s}_1, \mathbf{s}_1) & C(\mathbf{s}_1, \mathbf{s}_2) & \cdots & C(\mathbf{s}_1, \mathbf{s}_n) \\ C(\mathbf{s}_2, \mathbf{s}_1) & C(\mathbf{s}_2, \mathbf{s}_2) & \cdots & C(\mathbf{s}_2, \mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n, \mathbf{s}_1) & C(\mathbf{s}_n, \mathbf{s}_2) & \cdots & C(\mathbf{s}_n, \mathbf{s}_n) \end{pmatrix}$$

is non-negative definite: $\mathbf{z}^\top \boldsymbol{\Sigma} \mathbf{z} \geq 0$ for all real valued vectors $\mathbf{z}$.

(Note the inconsequence in the notation here: A positive definite function requires a positive semi-definite matrix)

# Stationary random fields

A common simplifying assumption is that the random field is stationary.

## Strict stationarity

A random field $X(\mathbf{s})$ is called *strictly stationary* if for any vector $\mathbf{h}$ and for every collection $\mathbf{s}_1, \ldots, \mathbf{s}_n$ of locations in $\mathcal{D}$

$$F(y_1, \ldots, y_n; \mathbf{s}_1 + \mathbf{h}, \ldots, \mathbf{s}_n + \mathbf{h}) = F(y_1, \ldots, y_n; \mathbf{s}_1, \ldots, \mathbf{s}_n).$$

## Weak stationarity

A random field $X(\mathbf{s})$ is called *weakly stationary* if for any vector $\mathbf{h}$ and any locations $\mathbf{s}, \mathbf{t} \in \mathcal{D}$

$$\mu(\mathbf{s} + \mathbf{h}) = \mu(\mathbf{s}), \quad \text{and} \quad C(\mathbf{s} + \mathbf{h}, \mathbf{t} + \mathbf{h}) = C(\mathbf{s}, \mathbf{t}) = C(\mathbf{s} - \mathbf{t}).$$

There is no distinction between the two concepts in the Gaussian case and one then simply writes that the field is stationary.

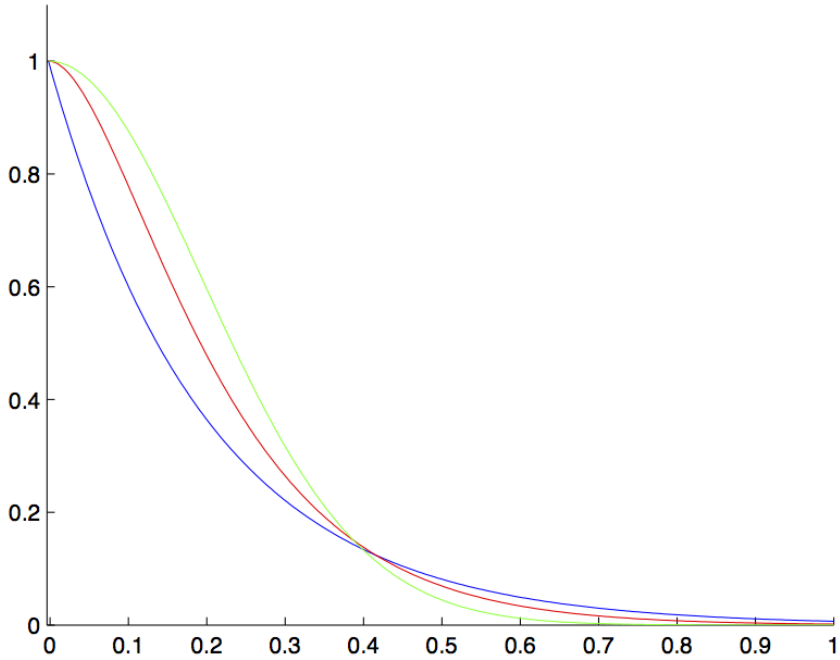
# Isotropic fields

An important subclass of the weakly stationary fields are the *isotropic* fields. These have covariance functions that depend only on distance, and not direction, between points,
i.e. $C(\mathbf{s}_1, \mathbf{s}_2) = C(\|\mathbf{s}_1 - \mathbf{s}_2\|)$.

In most practical applications of Gaussian random fields, the covariance function is chosen from a parametric family of isotropic covariance functions such as:

- Exponential covariance function.
- Gaussian covariance function.
- Matérn covariance function.

# The standard choice: Gaussian Matérn fields




The Matérn covariance function:

$$C(\mathbf{h}) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})\kappa^{2\nu}}(\kappa\|\mathbf{h}\|)^{\nu}K_{\nu}(\kappa\|\mathbf{h}\|), \quad \mathbf{h}\in\mathbb{R}^d, \nu>0,$$

Here $\nu$ is a shape parameter for the covariance function, $\kappa$ a spatial scale parameter, $\phi^2$ a variance parameter, $\Gamma$ is the gamma function, and $K_\nu$ is a modified Bessel function of the second kind.

# Spectral representations

- An alternative to covariance-based representation of Gaussian fields is to do the specification in the frequency domain.
- By Bochner's theorem, a function $C$ is a valid covariance function if and only if it can be written as

$$C(\mathbf{h}) = \int \exp(i\mathbf{h}^\top \mathbf{k}) \, \mathrm{d}\Lambda(\mathbf{k}) \tag{1}$$

  for some non-negative and symmetric measure $\Lambda$.
- Equation (1) is called the spectral representation of the covariance function, and if the measure $\Lambda$ has a Lebesgue density $S$, this is called the spectral density.
- For example, the spectral density associated with the Matérn covariance function is

$$S(\mathbf{k}) = \frac{\phi^2}{(2\pi)^d} \frac{1}{(\kappa^2 + \|\mathbf{k}\|^2)^{\nu + \frac{d}{2}}}.$$

# Variograms

- Another popular representation, first proposed by Matheron (1971), is the *(semi)variogram* $\gamma(\mathbf{h})$, that for a stationary process is defined as

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathsf{V}(X(\mathbf{s}+\mathbf{h}) - X(\mathbf{s})).$$

- One popular estimation method in geostatistics is to use so called empirical variograms
- These can be useful for non-differentiable random fields but can be missleading for differientable processes.
- We will not use variograms at all.

# Geostatistics and kriging

One of the most important problems in geostatistics is spatial reconstruction of a random field $X(\mathbf{s})$ given a finite number of observations $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ of the latent field at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ taken under measurement noise.

The most popular method for spatial reconstruction in geostatistics was developed by Georges Matheron.

Depending on the assumptions on the mean value function $\mu(\mathbf{s})$ for the latent field, linear kriging is usually divided into three cases:

Simple kriging $\mu(\mathbf{s})$ is known

Ordinary kriging $\mu(\mathbf{s}) = \mu$ and $\mu$ is unknown

Universal kriging $\mu(\mathbf{s}) = \sum_{k=1}^{m} \beta_k b_k(\mathbf{s})$ where $b_k$ are known basis functions and the parameters $\beta_k$ are unknown.

The kriging estimator of $X(\mathbf{s})$ at some location $\mathbf{s}_0$ is derived as the minimum mean squared error linear predictor.

# Hierarchical models

There is a close connection between kriging and estimation in *hierarchical models* which we use.

A hierarchical model is constructed as a hierarchy of conditional probability models that, when multiplied together, yield the joint distribution for all quantities in the model.

Typically, we have a three-stage statistical model for data $\mathbf{y}$ modelled using a latent field $\mathbf{x}$ with hyperparameters $\boldsymbol{\theta}$, structured in a hierarchical way

$$\pi(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

# The data $\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}$

We have been given some data $\mathbf{y}$.

- Normally distributed?
- Count data?
- Binary data?
- Point pattern?
- How was it collected? (Distance sampling? Capture/Recapture? Exhaustive sample? Preferential sampling?)

We place all of this information into our *likelihood* $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$.
A typical situation is when the latent field is measured under additive noise,

$$Y_i = X(\mathbf{s}_i) + \varepsilon_i \,.$$

A common assumption is that $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed with some variance $\sigma^2$, uncorrelated with the latent process.

# The latent field $\mathbf{x}|\boldsymbol{\theta}$

In our models, we will assume that the data depends on some unobserved *latent* components $\mathbf{x}$.

- Covariates
- Unstructured random effects ("white noise")
- Structured random effects (temporal dependency, spatial dependency, smoothness terms)

The dependence between the data and the latent field can be linear or non-linear, but as these are not directly observed, the modelling assumptions need to be more restrictive.

The process model can in itself be written as a hierarchical model, specified by a number of conditional sub-models.

# The hyperparameters $\boldsymbol{\theta}$

Both our likelihood and our latent field can depend on some hyperparameters $\boldsymbol{\theta}$

- Variance of observation noise
- Probability of a zero (zero-inflated models)
- Variance of the unstructured random field
- Range of a structured random effect (effective correlation distance)
- Autocorrelation parameter

For a Bayesian model, we specify these using a joint prior $\pi(\boldsymbol{\theta})$

Frequentists assume that the parameters are fixed but unknown. The model is then sometimes referred to as an empirical-Bayesian model, or empirical hierarchical model.

# Inference

Inference in hierarchical models is performed using the *posterior distribution*

$$\pi(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})\pi(\mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Kriging predictions are calculated from the marginal posterior distribution

$$\pi(\mathbf{X}|\mathbf{Y}) \propto \int \pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y})\,\mathrm{d}\boldsymbol{\theta},$$

and one typically reports the posterior mean $\mathrm{E}(\mathbf{X}|\mathbf{Y})$ as a point estimator and the posterior variance $\mathrm{V}(\mathbf{X}|\mathbf{Y})$ as a measure of the uncertainty in the predictor.

The posterior distribution for $\mathbf{X}$ and $\boldsymbol{\theta}$ generally have to be estimated using Markov Chain Monte Carlo (MCMC) methods.

# Inference II

In an empirical hierarchical model, inference is instead performed using the conditional posterior $\pi(\mathbf{X}|\mathbf{Y}, \hat{\boldsymbol{\theta}})$. Here $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ obtained using for example maximum likelihood estimation, or maximum a posteriori estimation in the Bayesian setting.

The parameter model $\pi(\boldsymbol{\theta})$ can often be chosen so that the posterior mean and variance of $\mathbf{X}$ agree with the classical kriging predictions.

Even if this is not done, we will refer to the conditional mean of the posterior distribution as the kriging predictor.

# Latent Gaussian Models

We call a Bayesian hierarchical model where $\pi(\mathbf{x}|\boldsymbol{\theta})$ is a Gaussian distribution a Latent Gaussian model (LGM):

$$
\begin{aligned}
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \\
\mathbf{x} \mid \boldsymbol{\theta} &\sim \pi(\mathbf{x} \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \\
\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta} &\sim \prod_i \pi(y_i \mid \eta_i, \boldsymbol{\theta})
\end{aligned}
$$

Note that we also assume that the observations are independent given the latent process.

This is a huge model class that is used in many seemingly unrelated areas, and which is especially useful if we let $\mathbf{x}$ be a GMRF

# Bayesian linear models

Consider the linear model $y_i = \mu + \beta_1 c_i^1 + \beta_2 c_i^2 + u_i + \epsilon_i$.

- $y_i$ is an observation
- $\mu$ is the intercept
- $\mathbf{c}^1$ and $\mathbf{c}^2$ are covariates (fixed effects) and $\beta_1$ and $\beta_2$ are the corresponding weights
- $\epsilon_i$ is i.i.d. normal observation noise.
- $\mathbf{u}$ is a random effect

To make a Bayesian model, we need to chose some priors. Classical choices:

- $\boldsymbol{\beta} = (\mu, \beta_1, \beta_2)^T \sim N(\mathbf{0}, \sigma_{\text{fix}}^2 \mathbf{I})$, where $\sigma_{\text{fix}}$ is a large number number.
- $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}})$ where the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}}$ is known.
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_{\text{n}}^2 \mathbf{I})$.

# Bayesian structured additive regression models

## GLM/GAM/GLMM/GAMM/+++

- Perhaps the most important class of statistical models
- $n$-dimensional observation vector $\mathbf{y}$, distributed according to an exponential family.
- mean $\mu_i = \mathsf{E}(\mathbf{y_i})$ linked to a linear predictor

$$\eta_i = g(\mu_i) = \alpha + \mathbf{z_i}^T \boldsymbol{\beta} + \sum_\gamma f_\gamma(c_{\gamma,i}) + \mathbf{u_i}, \quad i = 1, \ldots, n$$

where

$$\begin{aligned} \alpha &: \text{Intercept} \\ \boldsymbol{\beta} &: \text{linear effects of covariates } \mathbf{z} \\ \{f_\gamma(\cdot)\} &: \text{Non-linear/smooth effects of covariates } \mathbf{c}_\gamma \\ \mathbf{u} &: \text{Unstructured error terms} \end{aligned}$$

# Bayesian structured additive regression models cont.

Flexibility due to many different forms of the unknown functions $\{f_\gamma(\cdot)\}$

- relax linear relationship of covariates
- include random effects
- temporally and/or spatially indexed covariates

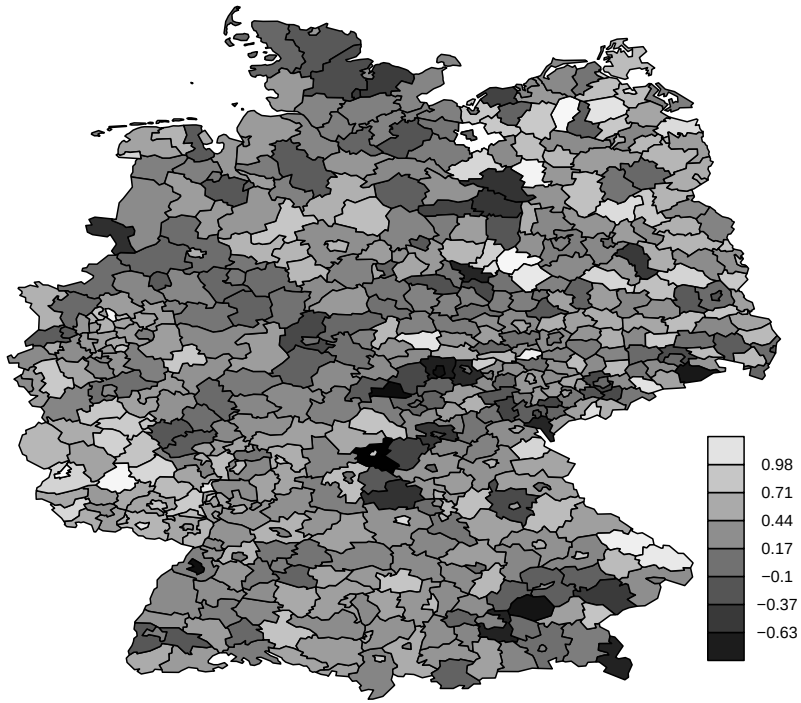
Special cases:

- Generalized linear models (GLM): $g(\mu) = \alpha + \sum_{j=1}^{m} \beta_j z_j$
- Generalized additive models (GAM): $g(\mu) = \alpha + \sum_{j=1}^{m} f_j(z_j)$

A latent Gaussian model is obtained by assigning Gaussian priors to all random terms $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \{f_\gamma(\cdot)\}, \mathbf{u}\}$ in the linear predictor.

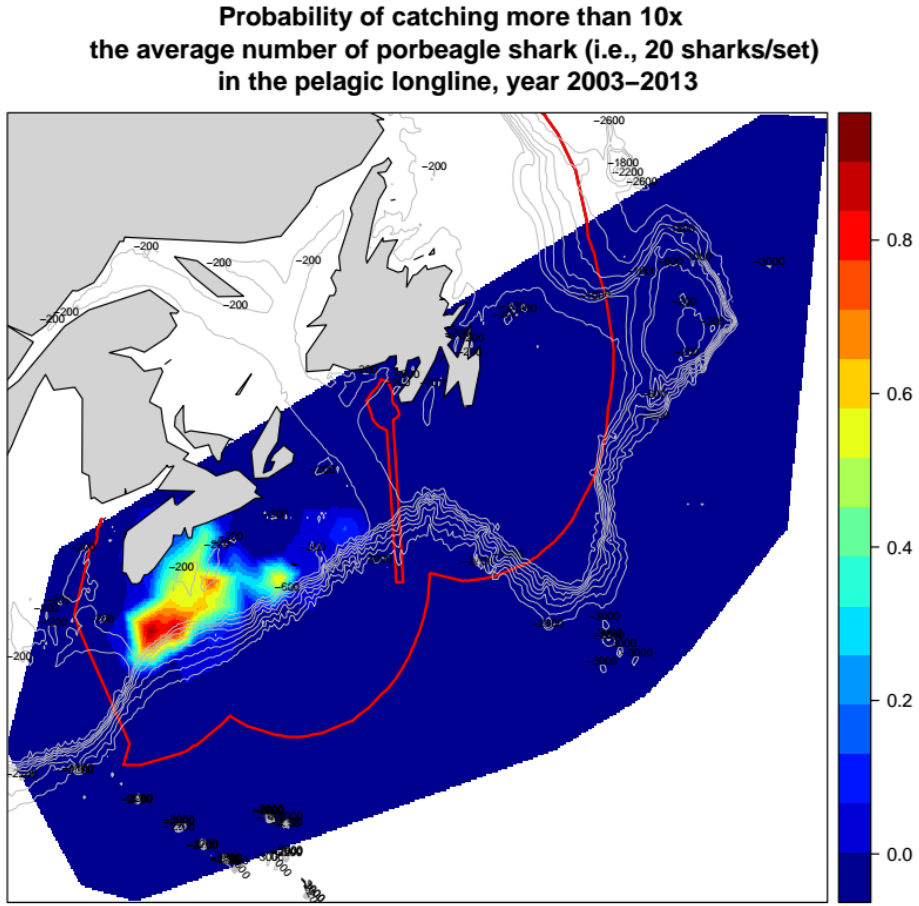
# Example: Disease mapping

- Data $y_i \sim$ Poisson$(E_i exp(\eta_i))$
- Log-relative risk
  $\eta_i = \mu + u_i + v_i + f(c_i)$
- Smooth effect of a covariate $\mathbf{c}$
- Structured component $\mathbf{u}$
- Unstructured component $\mathbf{v}$



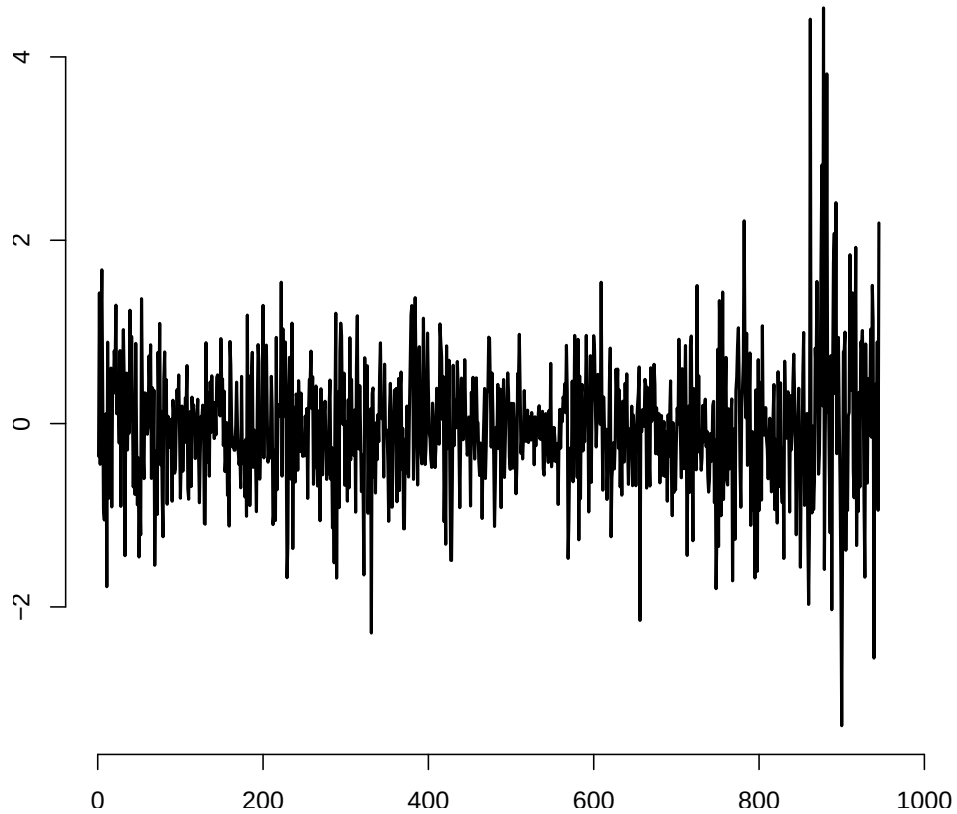
# Example: Spatial geological count data

- Spatio-temporal observations
  $y_{it} \sim \text{nBin}(E_{it} exp(\eta_{it}))$
- Log-relative risk of bycatch
  $\eta_{it} = u_{it} + \sum_j f(c_{it}^j)$
- Smooth effects of covariates $\mathbf{c}^j$
- Spatio-temporal random field $\mathbf{u}$



# Example: Stochastic volatility

- Log daily difference of the pound-dollar exchange rates
- Data $y_t|\eta_t \sim \mathcal{N}(0, \exp(\eta_t))$
- Volatility $\eta_i = \mu + u_t$
- Unknown mean $\mu$
- AR-process $\mathbf{u}$



# A more surprising example: Point processes

Spatial point processes model

- They focus on the random location at which events happen.
- They make excellent models for 'presence only' data when coupled with an appropriate observation process.
- Realistic models can be quite complicated.

# Log-Gaussian Cox processes

The homogeneous Poisson process is often too restrictive. Generalizations include:

- inhomogeneous Poisson process - inhomogeneous intensity
- Markov point process - local interactions among individuals
- Cox process - random intensity

We focus on the Cox process, the random intensity depends on a Gaussian random field $Z(s)$:

$$\Lambda(s) = \exp(Z(s))$$

If $Y$ denotes the set of observed locations, the likelihood is

$$\log(\pi(Y|\eta)) = |\Omega| - \int_\Omega \Lambda(s)\, ds + \sum_{s_i \in Y} \Lambda(s_i),$$

*This is very different to the previous examples!*

# Or is it?

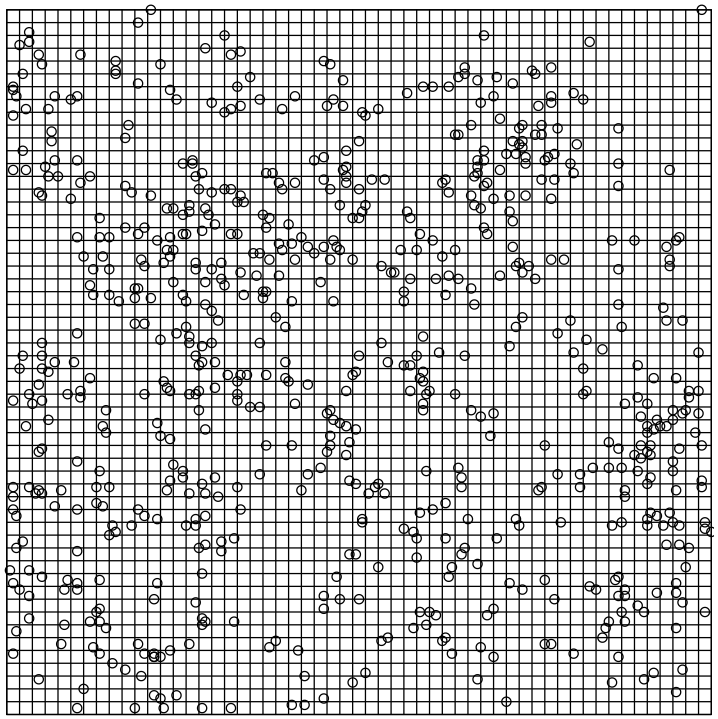
NB: *The number of points in a region $R$ is Poisson distributed with mean $\int_R \Lambda(s)\, ds$.*

- Divide the 'observation window' into rectangles.
- Let $y_i$ be the number of points in rectangle $i$.

$$y_i | x_i, \boldsymbol{\theta} \sim Po(e^{x_i}),$$

- The log-risk surface is replaced with

$$\mathbf{x} | \boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})).$$



# Back to the linear model

*Observation*: $(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta})$ are jointly Gaussian!

$$\pi(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) \propto \exp\left(-\frac{\tau_{\mathsf{n}}}{2}(\mathbf{y} - \mathbf{u} - \mathbf{X}^T\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{u} - \mathbf{X}^T\boldsymbol{\beta})\right)$$

$$= \exp\left(-\frac{\tau_{\mathsf{n}}}{2}\begin{pmatrix}\mathbf{y}^T & \mathbf{u}^T & \boldsymbol{\beta}^T\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{I} & -\mathbf{X}^T \\ -\mathbf{I} & \mathbf{I} & -\mathbf{X}^T \\ -\mathbf{X} & -\mathbf{X} & \mathbf{X}^T\mathbf{X}\end{pmatrix}\begin{pmatrix}\mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\beta}\end{pmatrix}\right)$$

It follows that

$$\pi(\mathbf{y}, \mathbf{u}, \boldsymbol{\beta}) = \pi(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta})\pi(\mathbf{u})\pi(\boldsymbol{\beta})$$

$$\propto \exp\left(-\frac{\tau_{\mathsf{n}}}{2}\begin{pmatrix}\mathbf{y}^T\mathbf{u}^T\boldsymbol{\beta}^T\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{I} & -\mathbf{X}^T \\ -\mathbf{I} & \mathbf{I} + \tau_{\mathsf{n}}^{-1}\mathbf{Q}_{\mathbf{u}} & -\mathbf{X}^T \\ -\mathbf{X} & -\mathbf{X} & \mathbf{X}^T\mathbf{X} + \frac{\tau_{\mathsf{fix}}}{\tau_{\mathsf{n}}}\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\beta}\end{pmatrix}\right)$$

# Estimation

Let $\mathbf{x} = (\mathbf{y}, \mathbf{u}, \boldsymbol{\beta})$. Estimating the parameters in the model, using MCMC or ML, we have to evaluate the log-likelihood

$$\log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{x}$$

We can easily calculate marginal distributions, for example to do kriging. Recall that if

$$\begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix} \sim N\left( \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix} \right),$$

then the conditional distribution is given by

$$\mathbf{x}_A | \mathbf{x}_B \sim \mathsf{N}(\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\mathbf{x}_B, \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA})$$

# Can we calculate these things in practice?

Evaluating the likelihood and the kriging predictor both require solving $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{x}$. Evaluating the likelihood also requires calculating $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$.

- Computations scale as $\mathcal{O}(N^3)$
- Storage scales as $\mathcal{O}(N^2)$:
  - $2500$ points for $20$ years requires $\sim 20$ Gbytes

Thus, even this very simple model is not feasible for large problems

For more complicated models that require MCMC, $N$ does not have to be particularly large for computations to be a major problem

We need to decrease the computational burden!

# "Low rank methods"

A popular approach to decrease the computational cost is to Approximate $X(\mathbf{s})$ using some basis expansion

$$X(\mathbf{s}) = \sum_{j=1}^{m} w_j \varphi_j(\mathbf{s}), \tag{2}$$

where $w_j$ are Gaussian random variables and $\{\varphi_j\}_{j=1}^m$ are some pre-defined basis functions.
This allows us to write $\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{B}\mathbf{\Sigma}_w\mathbf{B}^\top$ and basically gives us $\mathcal{O}(K^3)$ cost instead of $\mathcal{O}(N^3)$ cost: Choose $K \ll N$.
There are many ways to obtain these "low rank" approximations:

- Karhunen-Loève transforms
- Empirical orthogonal functions
- Process convolutions
- Fixed-rank kriging or predictive processes
- +++

# Key Lesson: Sparse matrices

**Definition (Sparse Matrix)**

A matrix $\mathbf{Q}$ is called *sparse* if most of its elements is zero.

- There exist very efficient numerical algorithms to deal with sparse matrices
  - Computations scale as $\mathcal{O}(N^{3/2})$.
  - Storage scales as $\mathcal{O}(N)$:
    - $2500$ points for $20$ years requires $\sim 400$ Kilobytes

Two possible options:

1. Force $\Sigma$ to be sparse.
   Forces *independence* between variables.
2. Force the *precision matrix* $Q = \Sigma^{-1}$ to be sparse.
   What does this correspond to?

# Example: AR(1) process

The simplest example of a GMRF is the AR(1) process:

$$x_t = \phi x_{t-1} + \epsilon_t, \qquad t = 1, 2, \ldots, \qquad \epsilon_t \sim \mathcal{N}(0, 1)$$

where $t$ represents time and the distribution of $x_0$ is chosen as the stationary distribution of the process: $x_0 \sim \mathcal{N}\left(0, \frac{1}{1-\phi^2}\right)$

The joint density for $\mathbf{x}$ is

$$\begin{aligned}\pi(\mathbf{x}) &= \pi(x_0)\pi(x_1|x_2)\cdots\pi(x_{n-1}|x_{n-2}) \\ &= \frac{1}{(2\pi)^{n/2}}|\mathbf{Q}|^{1/2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right).\end{aligned}$$

The matrix $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ is called the precision matrix

The covariance matrix for $\mathbf{x}$ is dense since all time points are dependent, $\Sigma_{ij} = \frac{1}{1-\phi^2}\phi^{|i-j|}$.

# Conditional independence

However, the precision matrix $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ is sparse:

$$Q = \begin{bmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi^2 & -\phi & & \\ & \ddots & \ddots & \ddots & \\ & & -\phi & 1+\phi^2 & -\phi \\ & & & -\phi & 1 \end{bmatrix}$$

What is the key property of this example that causes $\mathbf{Q}$ to be sparse?

- The key lies in the full conditionals

$$x_t | \boldsymbol{x}_{-t} \sim \mathcal{N}\left( \frac{\phi}{1-\phi^2}(x_{t-1} + x_{t+1}), \frac{\sigma^2}{1+\phi^2} \right)$$

- Each timepoint is only conditionally dependent on the two closest timepoints, which is the reason for the tridiagonal structure of $\mathbf{Q}$

# Main features of GMRFs

- Analytically tractable
- Modelling using conditional independence
- Merging GMRFs using conditioning (hierarchical models)
- Unified framework for
  - understanding
  - representation
  - computation using numerical methods for sparse matrices
- Fits nicely into the MCMC world
- Can construct faster and more reliable block-MCMC algorithms.
- Approximate Bayesian inference
- Approximate GRFs through SPDE representations

# Usage of GMRFs (I)

**Structural time-series analysis**

- Autoregressive models.
- Gaussian state-space models.
- Computational algorithms based on the Kalman filter and its variants.

**Analysis of longitudinal and survival data**

- temporal GMRF priors
- state-space approaches
- spatial GMRF priors

used to analyse longitudinal and survival data.

# Usage of GMRFs (II)

**Graphical models**

- A key model
- Estimate $\mathbf{Q}$ and its (associated) graph from data.
- Often used in a larger context.

**Semiparametric regression and splines**

- Model a smooth curve in time or a surface in space.
- Intrinsic GMRF models and random walk models
- Discretely observed integrated Wiener processes are GMRFs
- GMRFs models for coefficients in B-splines.

# Usage of GMRFs (III)

**Image analysis**

- Image restoration using the Wiener filter.
- Texture modelling and texture discrimination.
- Segmentation and object identification
- Deformable templates
- 3D reconstruction
- Restoring ultrasound images

**Spatial statistics**

- Latent GMRF model analysis of spatial binary data
- Geostatistics using GMRFs
- Analysis of data in social sciences and spatial econometrics
- Spatial and space-time epidemiology
- Environmental statistics
- Inverse problems