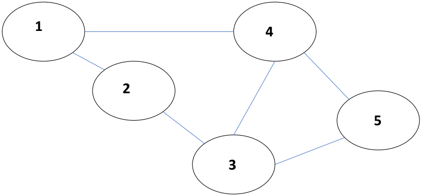
Conditional independence via $\boldsymbol{Q}$

All other variables than y_i are denoted $\boldsymbol{y}_{-i}$.
Neighborhood of node i is denoted $\mathcal{N}_i$.
Markov assumption:

$$p(y_i|\boldsymbol{y}_{-i}) = p(y_i|y_j; j \in \mathcal{N}_i)$$

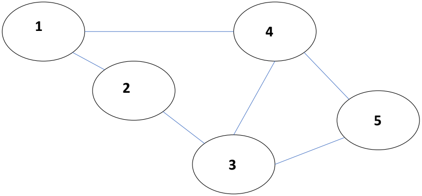
The neighborhood structure is given by the non-zero entries in $\boldsymbol{Q}$. This modeling approach very popular for graphical models.

Conditional independence via $\boldsymbol{Q}$



The neighborhood structure is given by the non-zero entries in $\boldsymbol{Q}$.
The Cholesky factor is a matrix square root defined by $\boldsymbol{L}\boldsymbol{L}^t = \boldsymbol{Q}$.

Conditional (ordered) independence via $\boldsymbol{L}$



The Cholesky factor is a matrix square root defined by $\boldsymbol{L}\boldsymbol{L}^t = \boldsymbol{Q}$.
It defines the conditional independence in order $p(y_i | y_j ; j = i + 1, \ldots, n)$.
There exists algorithms for finding the optimal order of calculation (minimum fill-in) to maintain a sparse matrix.

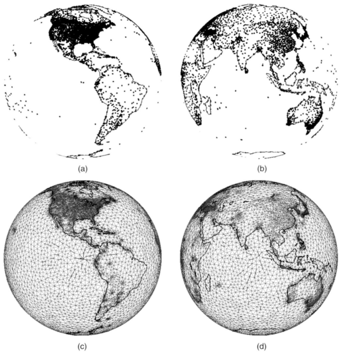
GMRF result for continuous spatial processes

- There is an explicit link between a Matern covariance function and an Stochastic partial differential equation. (Whittle)
- This differential equation can be solved on a mesh for test functions giving a GMRF with sparse precision matrix. (Lindgren et al.)

$$(\kappa^2 - \Delta)^{\alpha/2} x(\boldsymbol{s}) = \boldsymbol{z}(\boldsymbol{s}) \tag{1}$$

$z(\boldsymbol{s})$ is an independent (white noise) Gaussian process. The spatial process $x(\boldsymbol{s})$ is a Gaussian process with Matern covariance. The parameters α and κ relates to the covariance and smoothness in the Matern process.

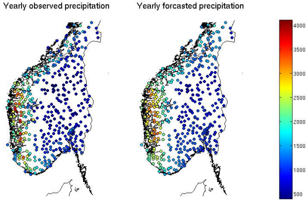
Mesh illustration



The result means that GPs can be computed quickly $O(n^{3/2})$ for large lattices, while still maintaining properties of the Matern process.

Gaussian processes and applications

Large spatial (spatio-temporal) datasets of positive variables or counts data.
GP is a building block.



Hierarchical model

Conditionally independent data Y_i, given x_i, $i = 1, \ldots, n$.
Latent variable $\boldsymbol{x} = (x_1, \ldots, x_n)$.

$$p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

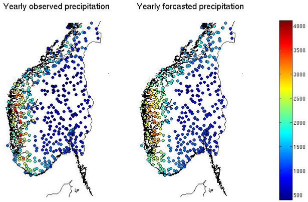
Mean is $E(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Positive-definite variance-covariance matrix is

$$Var(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \Sigma_{1,1} & \ldots & \Sigma_{1,n} \\ \ldots & \ldots & \ldots \\ \Sigma_{n,1} & \ldots & \Sigma_{n,n} \end{bmatrix},$$

$\Sigma_{i,i} = \sigma_i^2 = \mathrm{Var}(x_i)$, $\Sigma_{i,j} = \mathrm{Cov}(x_i, x_j)$, $\mathrm{Corr}(x_i, x_j) = \Sigma_{i,j}/(\sigma_i\sigma_j)$.

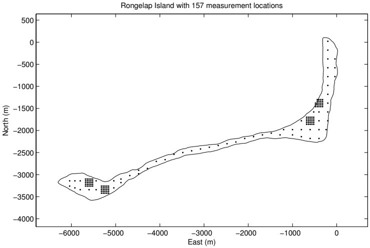
Examples of spatial latent Gaussian models

Rainfall data are not Gaussian, but the correlation in model parameters can be integrated by a latent GP.



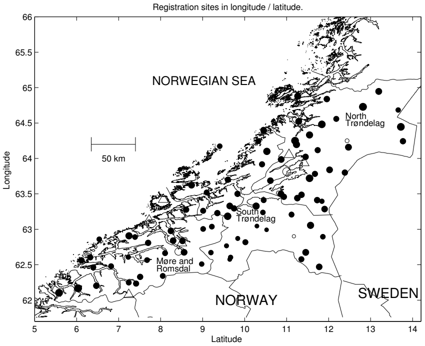
Examples of spatial latent Gaussian models

Radioactivity counts: Poisson. The log intensity can be modeled as a GP. This is a simple approach for getting multivariate distribution functions for count data.



Example of spatial latent Gaussian models

Number of days with rain for $k = 92$ sites in September-October 2006. The logit probability can be modeled as a GP, getting multivariate distribution functions for count data.



Statistical model

Consider the following hierarchical model

1. Observed data $\boldsymbol{y} = (y_1, \dots, y_n)$ where

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} p(y_i \mid x_i)$$

   Often exponential family: Normal, Poisson, binomial, etc. $\log p(y_i|x_i) = \frac{y_i x_i - b(x_i)}{a(\phi)} + c(\phi, y_i)$. $b(x)$ canonical link.

2. Latent Gaussian process $\boldsymbol{x} = (x_1, \dots, x_n)$

$$p(\boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = N[\boldsymbol{H}\boldsymbol{\beta},\ \boldsymbol{\Sigma}(\boldsymbol{\theta})]$$

3. Prior for hyperparameters $p(\boldsymbol{\theta})$, $p(\boldsymbol{\beta})$ **if Bayesian**

Mixed models - Normal linear case

Common model

- $y_i = \boldsymbol{H}_i \boldsymbol{\beta} + v_i + \epsilon_i = x_i + \epsilon_i$,
- $x_i = \boldsymbol{H}_i \boldsymbol{\beta} + v_i$, v_i is a structured effect.
- x_i Gaussian random effect having a structured covariance model with parameter $\boldsymbol{\theta}$. (Could be $\boldsymbol{U}_{ij}\boldsymbol{x}$ for group or individual i.)
- $\epsilon_i \sim N(0, \tau^2)$, iid effect.
- y_i is observation. (Could be y_{ij}, individual or group i, replicate j. Could be only at some locations, not all.)
- $\boldsymbol{\beta}$ fixed effect. Prior $p(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$.
- ϵ_i is random (unstructured) measurement noise. $\epsilon_i \sim N(0, \tau^2)$.

Mixed models - marginaliztion

Can integrate out $\boldsymbol{\beta}$.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \int p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta} = N[\boldsymbol{H}\boldsymbol{\mu}_{\beta}, \boldsymbol{H}\boldsymbol{\Sigma}_{\beta}\boldsymbol{H}' + \boldsymbol{\Sigma}(\boldsymbol{\theta})]$$

Mixed Gaussian models - full posterior of $\boldsymbol{x}$

Model for latent process: $p(\boldsymbol{x}|\theta) = N(\boldsymbol{\mu}, \boldsymbol{Q})$ (precision formulation, assuming β known),
Model for data, given latent variable: $p(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{P})$. $\boldsymbol{P}$ is diagonal (precision of measurement).

$$p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto p(\boldsymbol{x}|\theta)p(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})$$

$\boldsymbol{\Sigma}^{-1}_{x|y,\boldsymbol{\theta}} = \boldsymbol{Q} + \boldsymbol{A}'\boldsymbol{P}\boldsymbol{A}$, $\boldsymbol{\Sigma}^{-1}_{x|y,\boldsymbol{\theta}}\boldsymbol{\mu}_{x|y,\boldsymbol{\theta}} = \boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{A}'\boldsymbol{P}\boldsymbol{y}$.
(algebraically equivalent with covariance forms given in earlier lectures)

Mixed models - Gaussian approximate posterior of $\boldsymbol{x}$

Likelihood model for data, given latent field, is Poisson, binomial, or similar.
With non-Gaussian data one can optimize the posterior and fit a quadratic form at the mode. This gives a Gaussian approximation to the full posterior of $\boldsymbol{x}$.
Model for data, given latent variable: $p(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{P})$. $\boldsymbol{P}$ is diagonal (precision of measurement).

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{x}) \approx N(\hat{\boldsymbol{\mu}}_{x|y,\theta}, \hat{\boldsymbol{\Sigma}}_{x|y,\theta})$$

$\boldsymbol{\mu}_{x|y,\theta} = \text{argmax} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. $\boldsymbol{\Sigma}^{-1}_{x|y,\theta}$ fit from the curvature at the mode. More later.

Mixed models - Inference

Common situation that has been hard to infer effectively:

- Frequentist, $\hat{\boldsymbol{\theta}}$: Laplace approximations or estimating equations.
- Bayesian $p(\boldsymbol{\theta}|\boldsymbol{y})$: Markov chain Monte Carlo or INLA.
- Inference not enough, wish to do model criticism, outlier detection, design, etc. Such goals require fast computational tools!

Project A: Bayesian optimization

Gaussian processes are commonly used in optimization of complex functions.

Usually the function $Y(\boldsymbol{a})$ is very expensive to evaluate.

Goal

$$\hat{\boldsymbol{a}} = \text{argmax}\, Y(\boldsymbol{a})$$

Example: $\boldsymbol{a} = (a_1, a_2)$ is decision alternative, $Y(\boldsymbol{a})$ is profit.

Project A: Bayesian optimization

Expected improvement:

$$\begin{aligned}\text{EI} &= E(\max\{0, Y(\boldsymbol{a}) - Y^*\} | \boldsymbol{Y}_B) \\ &= (\hat{\mu}(\boldsymbol{a}) - Y^*)\Phi\left[\frac{\hat{\mu}(\boldsymbol{a}) - Y^*}{\hat{\sigma}(\boldsymbol{a})}\right] + \hat{\sigma}(\boldsymbol{a})\phi\left[\frac{\hat{\mu}(\boldsymbol{a}) - Y^*}{\hat{\sigma}(\boldsymbol{a})}\right]\end{aligned}$$

$$Y^* = \max \boldsymbol{Y}_B$$

$\hat{\mu}(\boldsymbol{a})$ and $\hat{\sigma}(\boldsymbol{a})$ are posterior mean and standard deviation, given $\boldsymbol{Y}_B$.
Φ and ϕ is cdf and pdf of standard Gaussian distribution.

Project A: Bayesian optimization

Sequential optimization using expected improvement.
Repeat the following for some iterations:

- Use EI to find next best point, given current data.
- Evaluate next point.
- Augment B set with this observation.

Project A: Spatial regression model

Model: $Y(\boldsymbol{a}) = \beta + w(\boldsymbol{a}) + \epsilon(\boldsymbol{a})$.

1. $Y(\boldsymbol{a})$ response variable at alternative $\boldsymbol{a} = (a_1, a_2)$.
2. β trend.
3. $w(\boldsymbol{a})$ structured GP.
4. $\epsilon(\boldsymbol{a})$ unstructured (independent) Gaussian measurement noise.

Use MLE to specify parameters β and $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)$ from evaluation data at $n_B = 100$ random locations: $\boldsymbol{Y}_B = (Y(\boldsymbol{a}_1), \ldots, Y(\boldsymbol{a}_{n_B}))'$.

Project A: GMRF tests for a few examples

- Calculate GMRF structure. Cholesky matrix.
- Compare computational costs /gains of sparse matrices.