Inference for Gaussian Markov random fields and the INLA approach

Johan Lindström[1]

[1]Centre for Mathematical Sciences
Lund University

Lund 2016-02-11

LUND
UNIVERSITY

Whom I "borrowed" slides from

David Bolin's PhD course on GMRFs from 2015.
`www.math.chalmers.se/~bodavid/GMRF2015/`

Various presentations from the INLA group.
`www.r-inla.org/`

Daniel Simpson's Helsinki presentation.
`www.math.ntnu.no/~daniesi/slides.pdf`

Spatial interpolation

Given observations at some locations, $Y(\boldsymbol{s}_i),\ i = 1 \dots n$
we want to make statements about the value at unobserved location(s), $X(\boldsymbol{s})$.

The general model formulation is

$$Y(\boldsymbol{s}_i)|\boldsymbol{X}, \vartheta \sim F(X(\boldsymbol{s}_i), \vartheta)$$
$$\boldsymbol{X}|\vartheta \sim \text{Spatial field}(\vartheta)$$

often we assume $\boldsymbol{X}$ to be a Gaussian field.

Bayesian hierarchical modelling using GMRF

Data model, $\mathrm{p}(\boldsymbol{y}|\boldsymbol{\eta}, \vartheta)$: Describing how **observations** arise assuming a **known latent field** $\boldsymbol{\eta}$.

Latent model, $\mathrm{p}(\boldsymbol{\eta}|\vartheta)$: Describing how the **latent field** behaves.

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\beta \qquad \boldsymbol{x} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\vartheta)\right)$$

Parameters, $\mathrm{p}(\vartheta)$: **Prior knowledge** of the parameters.

Bayesian hierarchical modelling using GMRF

In general these models can be seen as GLMs with

Data model The data is assumed to be **conditionally independent**

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{\eta}, \vartheta) = \prod_i \mathrm{p}(y_i|\eta_i, \vartheta)$$

Latent model The latent model is often constructed as a linear combination of **spatial effects** and **regression mean**

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\beta = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ \beta\end{bmatrix}$$

$$\boldsymbol{x} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\vartheta)\right) \qquad \beta \sim \mathrm{N}\left(\boldsymbol{0}, q_\beta^{-1} \cdot \boldsymbol{I}\right)$$

$$\begin{bmatrix}\boldsymbol{x} \\ \beta\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}(\vartheta) & \boldsymbol{0} \\ \boldsymbol{0} & q_\beta \cdot \boldsymbol{I}\end{bmatrix}^{-1}\right)$$

Here $q_\beta \cdot \boldsymbol{I}$ represents a vauge prior on β.

Inference

Given a Bayesian hierarchical model we are interested in

Posteriors for the parameters

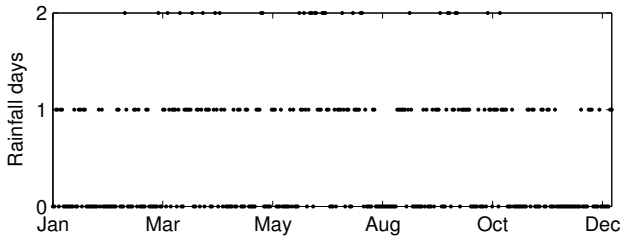
$$\mathrm{p}(\vartheta|\boldsymbol{y}) \propto \mathrm{p}(\boldsymbol{y}|\vartheta)\,\mathrm{p}(\vartheta)$$

Posteriors for the latent field

$$\mathrm{p}(\boldsymbol{\eta}|\boldsymbol{y}) \propto \int \mathrm{p}(\boldsymbol{\eta}|\boldsymbol{y}, \vartheta)\,\mathrm{p}(\vartheta|\boldsymbol{y})\,\mathrm{d}\vartheta$$

Example: Tokyo rainfall (Kitagawa, 1987)

- During 1983 and 1984 each day with with more than 1 mm rainfall in Tokyo was recorded.
- We want to estimate the probability of rainfall, p_t, for a given calendar day, $t = 1, \ldots, 366$.



Model for the Tokyo rainfall

The Tokyo data can be modelled using binomial-observations of a latent Gaussian process:

$$y_t | x_t \sim \text{Bin}(N_t, p_t) \qquad p_t = \frac{e^{x_t}}{1 + e^{x_t}}$$
$$\boldsymbol{x} | \tau \sim \text{N}\left(\boldsymbol{0}, \tau^{-1} \boldsymbol{Q}^{-1}\right)$$
$$\tau \sim \Gamma(a, b) \qquad a = 1 \quad b = 10^{-5}$$

To account for the seasonality, the latent process is modelled as a **cyclic random walk**.

For this model, $\boldsymbol{\eta} = \boldsymbol{x}$ and we use $\boldsymbol{x}$ for the rest of the example.

Latent field

The latent field is taken as the solution, on $t = 1, 2, \ldots, 366$, to the SPDE

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} x(t) = \frac{1}{\sqrt{\tau}} \mathcal{W}(t).$$

with cyclical boundaries, $x(1) \approx x(T)$, and $\kappa^2 = 0$.

The resulting $\boldsymbol{G}$ and $\boldsymbol{C}$ matrices are

$$\boldsymbol{G} = \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 & -1 \\ -1 & 2 & -1 & 0 & \ldots & \\ 0 & -1 & 2 & -1 & 0 & \ldots \\ & & & \ddots & & \\ -1 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix} \qquad \boldsymbol{C} = \boldsymbol{I}$$

giving $\boldsymbol{Q}_{\alpha=1} = \boldsymbol{G}$ (crw1) and $\boldsymbol{Q}_{\alpha=2} = \boldsymbol{G}\boldsymbol{G}$ (crw2).

Naïve Inference

Due to the markov structure of the model we have conditional posteriors

$$\mathrm{p}(x_t|x_{-t}, y_t) \propto (1 + e^{x_t})^{-N_t} \exp\left(y_t x_t - \frac{\tau Q_{tt}}{2}\left(x_t + \frac{\sum_{s \neq t} Q_{ts} x_s}{Q_{tt}}\right)^2\right)$$

and

$$\tau|\boldsymbol{x} \sim \Gamma\left(\frac{N_x - 1}{2} + a, \frac{\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}}{2} + b\right)$$

The posterior $\mathrm{p}(x_t|x_{-t}, y_t)$ is **log-concave** and can be sampled exactly using **adaptive rejection sampling** (Gilks and Wild, 1992).

Given these conditionals, the posterior $p(\boldsymbol{x}, \tau|\boldsymbol{y})$ can be estimated using Gibbs-sampling.

Naïve Inference — Auxiliary variables

As an alternative to the **adaptive rejection sampling** the MCMC estimation can be simplified through the use of **auxiliary variables**

For **student's t** we have:

If $X \sim \mathrm{N}(0, 1)$ and $Z \sim \mathrm{I}\Gamma(\nu/2, \nu/2)$ then

$$X\sqrt{Z} \sim \mathrm{t}(\nu)$$

Auxiliary variables for Binomials

For binary models with probit link (Albert and Chib, 1993) the model

$$y_t \sim \mathrm{Bin}(1, p_t) \qquad p_t = \Phi(x_t) \quad \text{probit link}$$

is equivalent to

$$w_t = x_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, 1)$$

$$y_t = \begin{cases} 1 & \text{if } w_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

and the problem reduces to sampling from a normal, $\boldsymbol{x}|\boldsymbol{w} \sim \mathrm{N}$ and a truncated normal $w_t|x_t, y_t \sim \mathrm{N}_{\text{trunc.}}$.

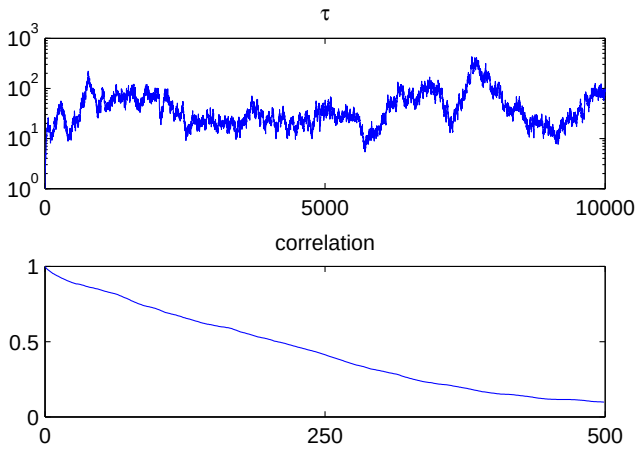
Auxiliary variables for logit links also exist (Held and Holmes, 2006).

Results — Single site Gibbs

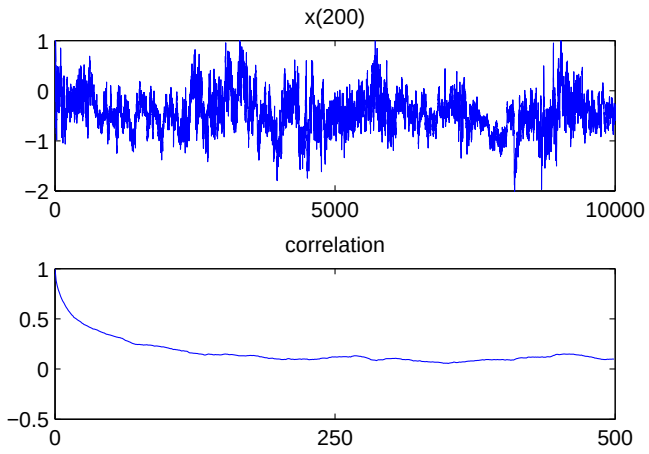
The single site Gibbs-sampler is easy to implement but has several drawbacks:

- **Very long run time** (hard to vectorise code)
- **Very slow mixing** of the chains.
- Mixing is poor between τ and the latent field $\boldsymbol{x}$
- And mixing is poor within the latent field $\boldsymbol{x}$

τ — Single site Gibbs



x_{200} — Single site Gibbs



Comments

For inference in latent fields it is well known (Knorr-Held and Rue, 2002) that **blocking improves mixing**. Using the auxiliary variables allows us to block the updates as

1. $\boldsymbol{x}|\boldsymbol{w} \sim \mathsf{N}$ can be sampled jointly
2. $w_t|x_t, y_t \sim \mathsf{N}_{\text{trunc.}}$ are conditionally independent and can be sampled jointly.
3. $\tau|\boldsymbol{x} \sim \Gamma\left(\frac{N_x-1}{2} + a, \frac{\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}}{2} + b\right)$

This will improve mixing but still leaves poor mixing between τ and $\boldsymbol{x}$.

MCMC is "right" at the limit (infinite run time). In practice, **fast approximate solutions** are better than exact slow solutions.

Inference for parameters and latent field

Given a model with **conditionally independent** observations

$$\mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}, \vartheta\right) = \prod_i \mathrm{p}\left(y_i|x_i, \vartheta\right)$$

from a latent GMRF $\boldsymbol{x} \sim \mathsf{N}\left(0, \boldsymbol{Q}^{-1}\vartheta\right)$ we want:

Posteriors for the parameters

$$\mathrm{p}\left(\vartheta|\boldsymbol{y}\right) \propto \mathrm{p}\left(\boldsymbol{y}|\vartheta\right)\mathrm{p}\left(\vartheta\right) = \int \mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}, \vartheta\right)\mathrm{p}\left(\boldsymbol{x}|\vartheta\right)\,\mathrm{d}\boldsymbol{x}\,\mathrm{p}\left(\vartheta\right)$$

Maximum posterior estimates

$$\arg\max_{\vartheta} \mathrm{p}(\vartheta|\boldsymbol{y})$$

Posteriors for the latent field

$$\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y}\right) \propto \int \mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y}, \vartheta\right)\mathrm{p}\left(\vartheta|\boldsymbol{y}\right)\,\mathrm{d}\vartheta$$

Computing the posterior — A "trick"

To avoid explicitly computing the integral

$$\int \mathrm{p}\left(\boldsymbol{y}, \boldsymbol{x}|\vartheta\right)\,\mathrm{d}\boldsymbol{x} = \int \mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}, \vartheta\right)\mathrm{p}\left(\boldsymbol{x}|\vartheta\right)\,\mathrm{d}\boldsymbol{x}$$

we note that conditional distributions provide the equality

$$\mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}, \vartheta\right)\mathrm{p}\left(\boldsymbol{x}|\vartheta\right) = \mathrm{p}\left(\boldsymbol{y}, \boldsymbol{x}|\vartheta\right) = \mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y}, \vartheta\right)\mathrm{p}\left(\boldsymbol{y}|\vartheta\right)$$

This gives

$$\mathrm{p}\left(\vartheta|\boldsymbol{y}\right) \propto \underbrace{\frac{\mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x}, \vartheta\right)\mathrm{p}\left(\boldsymbol{x}|\vartheta\right)}{\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y}, \vartheta\right)}}_{\mathrm{p}(\boldsymbol{y}|\vartheta)} \cdot \mathrm{p}\left(\vartheta\right) \quad \text{for any } \boldsymbol{x}.$$

With $\mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y}, \vartheta\right)$ being the posterior of $\boldsymbol{x}$ given $\boldsymbol{y}$.

Can we compute $\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta)$?

For Gaussian observations we have

$$\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x},\vartheta \sim \mathsf{N}\left(\boldsymbol{A}\boldsymbol{x},\boldsymbol{Q}_{\varepsilon}^{-1}(\vartheta)\right),$$
$$\boldsymbol{x}|\vartheta \sim \mathsf{N}\left(\boldsymbol{0},\boldsymbol{Q}^{-1}(\vartheta)\right),$$

with posteriors

$$\boldsymbol{x}|\boldsymbol{y},\vartheta \sim \mathsf{N}\left(\mu_{\boldsymbol{x}|\boldsymbol{y}}(\vartheta),\boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}}^{-1}(\vartheta)\right),$$

where

$$\mu_{\boldsymbol{x}|\boldsymbol{y}}(\vartheta) = \boldsymbol{Q}_{x|y}^{-1}(\vartheta)\boldsymbol{A}^{\top}\boldsymbol{Q}_{\varepsilon}(\vartheta)\boldsymbol{y},$$
$$\boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}}(\vartheta) = \boldsymbol{Q}(\vartheta) + \boldsymbol{A}^{\top}\boldsymbol{Q}_{\varepsilon}(\vartheta)\boldsymbol{A}.$$

Parameter estimation — Gaussian case

We now have:

$$\mathrm{p}(\vartheta|\boldsymbol{y}) \propto \mathrm{p}(\boldsymbol{y} \mid \vartheta)\,\mathrm{p}(\vartheta) = \frac{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x},\vartheta)\,\mathrm{p}(\boldsymbol{x}|\vartheta)}{\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta)} \cdot \mathrm{p}(\vartheta) \quad \text{for any } \boldsymbol{x}.$$

Pick $\boldsymbol{x} = \mathsf{E}(\boldsymbol{x} \mid \boldsymbol{y},\vartheta) = \mu_{\boldsymbol{x}|\boldsymbol{y}}$

$$\mathrm{p}(\boldsymbol{y}|\vartheta) \propto \frac{|\boldsymbol{Q}|^{1/2}\,|\boldsymbol{Q}_{\varepsilon}|^{1/2}}{\left|\boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}}\right|^{1/2}} \exp\left(-\frac{1}{2}\left[\mu_{\boldsymbol{x}|\boldsymbol{y}}^{\top}\boldsymbol{Q}\mu_{\boldsymbol{x}|\boldsymbol{y}} + \left(\boldsymbol{y}-\boldsymbol{A}\mu_{\boldsymbol{x}|\boldsymbol{y}}\right)^{\top}\boldsymbol{Q}_{\varepsilon}\left(\boldsymbol{y}-\boldsymbol{A}\mu_{\boldsymbol{x}|\boldsymbol{y}}\right)\right]\right).$$

Estimate of ϑ

$$\widehat{\vartheta} = \arg\max_{\vartheta} \mathrm{p}(\vartheta|\boldsymbol{y}) = \arg\max_{\vartheta} \mathrm{p}(\boldsymbol{y}|\vartheta)\mathrm{p}(\vartheta)$$

Non-Gaussian Observations

If the observations $\mathrm{p}(y_i|x_i,\vartheta)$ are **non-gaussian** the posterior

$$\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta)$$

does not have a closed form expression.

Note that the ordinary Bayesian solution of taking

$$\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta) \propto \mathrm{p}(\boldsymbol{y}|\boldsymbol{x},\vartheta)\,\mathrm{p}(\boldsymbol{x}|\vartheta)$$

is **insufficient** since we need the **explicit dependence** on ϑ!

We need an **approximation** of $\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta)$.

Taylor Series

Given a univariate function $f(x)$ the **Taylor series** around $x^{(0)}$ is given by

$$f(x) \approx f(x^{(0)}) + f'(x^{(0)})\left(x - x^{(0)}\right) + \frac{f''(x^{(0)})}{2!}\left(x - x^{(0)}\right)^2 + \dots$$

Taylor Series — Multivariate

Given a function $f(\boldsymbol{x})$ where $\boldsymbol{x}$ is a column vector, then

$$\begin{aligned} f(\boldsymbol{x}) \approx & f(\boldsymbol{x}^{(0)}) + \left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)^\top \nabla f^{(0)} \\ & + \frac{1}{2!}\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)^\top \boldsymbol{H}^{(0)}\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right) + \dots \end{aligned}$$

Here $\nabla f^{(0)}$ is the **gradient** of $f(\boldsymbol{x})$ evaluated at $\boldsymbol{x}^{(0)}$ and $\boldsymbol{H}^{(0)}$ is the **Hessian (second derivative)** matrix of $f(\boldsymbol{x})$.

The classical Laplace approximation

To compute and approximate integrals of the form

$$\int \exp\left(ng(x)\right)\,\mathrm{d}x$$

Pierre-Simon Laplace suggested a Taylor expansion of $g(x)$ around it's maximum, x_0, (i.e. $g'(x_0) = 0$ and $g''(x_0) < 0$)

$$g(x) \approx g(x_0) + \frac{1}{2}g''(x_0)(x - x_0)^2 + \cdots$$

Giving

$$\begin{aligned} \int \exp\left(ng(x)\right)\,\mathrm{d}x &\approx \exp\left(ng(x_0)\right)\int \exp\left(-\frac{n\left|g''(x_0)\right|}{2}(x - x_0)^2\right)\,\mathrm{d}x \\ &= \exp\left(ng(x_0)\right)\sqrt{\frac{2\pi}{n\left|g''(x_0)\right|}} \end{aligned}$$

Laplace approximation — Use in Statistics

Instead of approximating the integral $\int \exp\left(ng(x)\right)\,\mathrm{d}x$ we could use similar ideas to approximate densities of the form

$$\mathrm{p}(x) \propto \exp\left(g(x)\right)$$

Given the Taylor expansion of $g(x)$

$$\begin{aligned} g(x) &\approx g(x_0) + g'(x_0)(x - x_0) + \frac{1}{2}g''(x_0)(x - x_0)^2 \\ &= \left(g'(x_0) - x_0 g''(x_0)\right)x + \frac{g''(x_0)}{2}x^2 + \text{const.} \end{aligned}$$

and assuming $g''(x_0) < 0$ we have

$$X \overset{\text{approx.}}{\sim} \mathrm{N}\left(\frac{g'(x_0) - x_0 g''(x_0)}{-g''(x_0)}, \frac{-1}{g''(x_0)}\right)$$

Example — Gamma observations

Assuming a simple model with Gamma-distributed observations of one pixel

$$x \sim \mathrm{N}\left(0, q^{-1}\right) \qquad y|x \sim \Gamma\left(b, \frac{e^x}{b}\right)$$

with density functions

$$\mathrm{p}(x) = \sqrt{\frac{q}{2\pi}} \mathrm{e}^{-\frac{q}{2}x^2} \qquad \mathrm{p}(y|x) = \frac{y^{b-1}}{\Gamma(b)} \left(\frac{b}{e^x}\right)^b \mathrm{e}^{-\frac{b \cdot y}{e^x}}$$

Example — Posterior

For the posterior we now have

$$\begin{aligned}\log \mathrm{p}(x|y) &= \log \mathrm{p}(y|x) + \log \mathrm{p}(x) - \log p(y) \\ &= -b \cdot x - b \cdot y \cdot \mathrm{e}^{-x} - \frac{q \cdot x^2}{2} + \text{const.}\end{aligned}$$

with derivatives

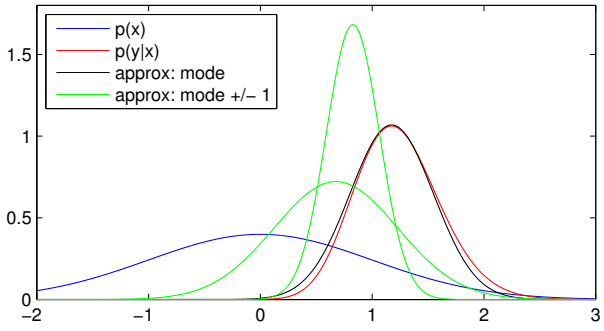
$$\begin{aligned}\frac{\partial \log \mathrm{p}(x|y)}{\partial x} &= -b + by\mathrm{e}^{-x} - qx \\ \frac{\partial^2 \log \mathrm{p}(x|y)}{\partial x^2} &= -by\mathrm{e}^{-x} - q\end{aligned}$$

giving

$$x|y \overset{\text{approx.}}{\sim} \mathrm{N}\left(\frac{by\mathrm{e}^{-x_0}(1+x_0) - b}{q + by\mathrm{e}^{-x_0}}, \frac{1}{q + by\mathrm{e}^{-x_0}}\right)$$

Example — Effect of $x^{(0)}$

Posterior given $q = 1$, $b = 5$ and $y = 4$ (or $\log y = 1.39$).



Comparing the effect of different choices for $x^{(0)}$; the mode of $\log p(y|x)$ provides the best approximation.

Parameter estimation using the Laplace approximation

Maximum posterior estimates of the parameters are given by maximising

$$\mathrm{p}(\vartheta|\boldsymbol{y}) \propto \frac{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x},\vartheta)\,\mathrm{p}(\boldsymbol{x}|\vartheta)}{\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta)} \cdot \mathrm{p}(\vartheta) \quad \text{for any } \boldsymbol{x}.$$

Replacing $\mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta)$ with the Laplace approximation $\mathrm{p}_G(\boldsymbol{x}|\boldsymbol{y},\vartheta)$ we obtain the approximate posterior

$$\widetilde{\mathrm{p}}(\vartheta|\boldsymbol{y}) \propto \frac{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x},\vartheta)\,\mathrm{p}(\boldsymbol{x}|\vartheta)}{\mathrm{p}_G(\boldsymbol{x}|\boldsymbol{y},\vartheta)} \cdot \mathrm{p}(\vartheta)$$

Approximating the posterior

For (non-Gaussian) observations we have that

$$\begin{aligned}\log \mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta) &= \log \mathrm{p}(\boldsymbol{y}|\boldsymbol{x},\vartheta) + \log \mathrm{p}(\boldsymbol{x}|\vartheta) + \text{const.} \\ &= \sum_i \log \mathrm{p}(y_i|x_i,\vartheta) + \log \mathrm{p}(\boldsymbol{x}|\vartheta) + \text{const.} \\ &= \sum_i \log \mathrm{p}(y_i|x_i,\vartheta) - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \text{const.}\end{aligned}$$

Our plan is to construct a Gaussian approximation of the posterior $p(\boldsymbol{x}|\boldsymbol{y},\vartheta)$ by computation of a **second order Laplace (Taylor) approximation** of

$$f_i(x_i) = \log \mathrm{p}(y_i|x_i,\vartheta) \qquad \text{around } \boldsymbol{x}^{(0)}$$

Laplace approximation

Given the log-posterior

$$\begin{aligned}\log \mathrm{p}(\boldsymbol{x}|\boldsymbol{y},\vartheta) &= \sum_i f_i(x_i) - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \text{const.} \\ f_i(x_i) &= \log \mathrm{p}(y_i|x_i,\vartheta)\end{aligned}$$

a Taylor expansion of $f_i(x_i)$ around $\boldsymbol{x}^{(0)}$ gives

$$f_i(x_i) \approx f_i(x_i^{(0)}) + f_i'(x_i^{(0)})\left(x_i - x_i^{(0)}\right) + \frac{1}{2}f_i''(x_i^{(0)})\left(x_i - x_i^{(0)}\right)^2$$

Or in vector form

$$\sum_i f_i(x_i) \approx \sum_i f_i(x_i^{(0)}) + \left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)^\top \nabla f^{(0)} + \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)^\top \boldsymbol{H}_f^{(0)}\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)$$

where $\nabla f^{(0)} = \left[f_i'(x_i^{(0)})\right]_i$ and $\boldsymbol{H}_f^{(0)} = \mathrm{diag}(f_i''(x_i^{(0)}))$.

Laplace approximation

Collecting terms and accounting for the Gaussian-prior on $\boldsymbol{x}$ we get

$$\begin{aligned}\log \mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right) &= \sum_i f_i(x_i) - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \text{const.} \\ &\approx \boldsymbol{x}^\top\left(\nabla f^{(0)} - \boldsymbol{H}_f^{(0)}\boldsymbol{x}^{(0)}\right) - \frac{1}{2}\boldsymbol{x}^\top\left(\boldsymbol{Q} - \boldsymbol{H}_f^{(0)}\right)\boldsymbol{x} + \text{const.}\end{aligned}$$

A Gaussian approximation $\mathrm{p}_G\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right)$ is obtained with

$$\begin{aligned}\mathsf{E}_{\boldsymbol{x}^{(0)}}\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right) &= \boldsymbol{Q}_{x|y}^{-1}\left(\nabla f^{(0)} - \boldsymbol{H}_f^{(0)}\boldsymbol{x}^{(0)}\right) \\ \mathsf{V}_{\boldsymbol{x}^{(0)}}\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right) &= \boldsymbol{Q}_{x|y}^{-1} = \left(\boldsymbol{Q} - \boldsymbol{H}_f^{(0)}\right)^{-1}\end{aligned}$$

Gaussian approximation — Suitable $\boldsymbol{x}^{(0)}$

With the Gaussian approximation of the denominator we have

$$\widetilde{\mathrm{p}}\left(\vartheta|\boldsymbol{y}\right) \propto \frac{\mathrm{p}\left(\boldsymbol{y}|\boldsymbol{x},\vartheta\right)\mathrm{p}\left(\boldsymbol{x}|\vartheta\right)}{\mathrm{p}_G\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right)} \cdot \mathrm{p}\left(\vartheta\right) \quad \text{for any } \boldsymbol{x}.$$

We want a good approximation for the **most likely values** of the posterior distribution; expand around the **mode**

$$\begin{aligned}\boldsymbol{x}^{(0)} &= \arg\max_{\boldsymbol{x}} \mathrm{p}\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right) \\ &= \arg\max_{\boldsymbol{x}}\Big(\underbrace{\log \mathrm{p}\left(\boldsymbol{y} \mid \boldsymbol{x},\vartheta\right)}_{f(\boldsymbol{x})} + \log \mathrm{p}\left(\boldsymbol{x} \mid \vartheta\right)\Big) \\ &= \arg\max_{\boldsymbol{x}}\Big(\sum_i f_i(x_i) - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}\Big).\end{aligned}$$

Gaussian approximation — Suitable $\boldsymbol{x}^{(0)}$

Since the derivative of

$$\sum_i f_i(x_i) - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}$$

has to be zero at the mode, $\boldsymbol{x}^{(0)}$, we have

$$\nabla f^{(0)} - \boldsymbol{Q}\boldsymbol{x}^{(0)} = 0 \qquad \Longleftrightarrow \qquad \nabla f^{(0)} = \boldsymbol{Q}\boldsymbol{x}^{(0)}$$

Inserting this into the conditional expectation for the Gaussian approximation we have

$$\begin{aligned}\mathsf{E}_{\boldsymbol{x}^{(0)}}\left(\boldsymbol{x}|\boldsymbol{y},\vartheta\right) &= \boldsymbol{Q}_{x|y}^{-1}\left(\nabla f^{(0)} - \boldsymbol{H}_f^{(0)}\boldsymbol{x}^{(0)}\right) \\ &= \boldsymbol{Q}_{x|y}^{-1}\left(\boldsymbol{Q}\boldsymbol{x}^{(0)} - \boldsymbol{H}_f^{(0)}\boldsymbol{x}^{(0)}\right) \\ &= \boldsymbol{Q}_{x|y}^{-1}\left(\boldsymbol{Q} - \boldsymbol{H}_f^{(0)}\right)\boldsymbol{x}^{(0)} = \boldsymbol{x}^{(0)}\end{aligned}$$

Parameter Estimation — Approximate posterior

To evaluate $\mathrm{p}(\vartheta|\boldsymbol{y})$ using the Taylor (Laplace) approximation do:

1. For a given ϑ find the mode
$$\boldsymbol{x}^{(0)} = \arg\max_{\boldsymbol{x}} \mathrm{p}(\boldsymbol{x}|\boldsymbol{y}, \vartheta)$$
2. Taylor expansion of $f(\boldsymbol{x})$ around $\boldsymbol{x}^{(0)}$
3. The approximation of $\mathrm{p}(\vartheta|\boldsymbol{y})$ is
$$\widetilde{\mathrm{p}}(\vartheta|\boldsymbol{y}) \propto \frac{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}^{(0)}, \vartheta)\,\mathrm{p}(\boldsymbol{x}^{(0)}|\vartheta)}{\mathrm{p}_G(\boldsymbol{x}^{(0)}|\boldsymbol{y}, \vartheta)} \cdot \mathrm{p}(\vartheta)$$

The (approximate) estimate of ϑ is

$$\vartheta_{\mathsf{MAP}} \approx \arg\max_{\vartheta} \widetilde{\mathrm{p}}(\vartheta|\boldsymbol{y})$$

Parameter Estimation — Details

- Technically we have
$$\widetilde{\mathrm{p}}(\vartheta|\boldsymbol{y}) \propto \frac{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}, \vartheta)\,\mathrm{p}(\boldsymbol{x}|\vartheta)}{\mathrm{p}_G(\boldsymbol{x}|\boldsymbol{y}, \vartheta)} \cdot \mathrm{p}(\vartheta) \quad \text{for any } \boldsymbol{x}.$$
so the $\boldsymbol{x}$ at which we evaluate and the $\boldsymbol{x}^{(0)}$ for the Taylor expansion can be different.
- The approximation $\mathrm{p}_G(\boldsymbol{x}|\boldsymbol{y}, \vartheta)$ is best at $\boldsymbol{x}^{(0)}$ making this a good choice to evaluate.
- Since the expectation and mode coincides the approximate posterior then simplifies to
$$\widetilde{\mathrm{p}}(\vartheta|\boldsymbol{y}) \propto \frac{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}^{(0)}, \vartheta)\,\mathrm{p}(\boldsymbol{x}^{(0)}|\vartheta)}{\left|\boldsymbol{Q}_{x|y}\right|^{1/2}} \cdot \mathrm{p}(\vartheta)$$

INLA (Rue et al., 2009)

The above estimate of ϑ_{MAP} describes the first part of INLA (integrated nested Laplace approximations). Recall that we also want

- $\mathrm{p}(\vartheta|\boldsymbol{y})$
- $\mathrm{p}(\boldsymbol{x}|\boldsymbol{y})$

These can not be computed, however INLA provides a way of computing the marginals

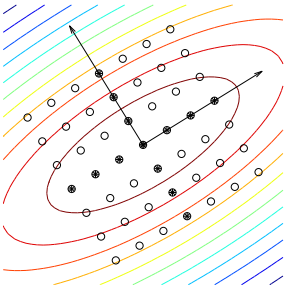
- $\mathrm{p}(\vartheta_i|\boldsymbol{y})$
- $\mathrm{p}(x_i|\boldsymbol{y})$

which is often sufficient for inference.

Posteriors for ϑ_i

Given the mode $\boldsymbol{\vartheta}_{\text{MAP}}$ we compute the posterior $\text{p}(\vartheta_i|\boldsymbol{y})$ by

1. Compute the Hessian of $\widetilde{\text{p}}(\boldsymbol{\vartheta}|\boldsymbol{y})$ at the mode.
2. Use the Hessian to construct an integration grid.
3. Do numerical integration.



Posteriors for $\boldsymbol{x}_i$

We can write the posteriors of $\text{p}(x_i|\boldsymbol{y})$ as

$$\begin{aligned}\text{p}(x_i|\boldsymbol{y}) &= \int \text{p}(x_i|\boldsymbol{y},\boldsymbol{\vartheta})\text{p}(\boldsymbol{\vartheta}|\boldsymbol{y})\,\text{d}\boldsymbol{\vartheta} \\ &\approx \int \text{p}_{\text{G}}(x_i|\boldsymbol{y},\boldsymbol{\vartheta})\widetilde{\text{p}}(\boldsymbol{\vartheta}|\boldsymbol{y})\,\text{d}\boldsymbol{\vartheta} \quad \approx \sum_k \text{p}_{\text{G}}(x_i|\boldsymbol{y},\boldsymbol{\vartheta}_k)\widetilde{\text{p}}(\boldsymbol{\vartheta}_k|\boldsymbol{y})\Delta_k\end{aligned}$$

Δ_k are integration weights due to the point configuration.

- The part $\widetilde{\text{p}}(\boldsymbol{\vartheta}_k|\boldsymbol{y})$ can be computed by evaluating the approximate posterior for each configuration $\boldsymbol{\vartheta}_k$.
- The Gaussian approximation $\text{p}_{\text{G}}(x_i|\boldsymbol{y},\boldsymbol{\vartheta})$ can be obtained from the approximation $\text{p}_{\text{G}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\vartheta})$ as

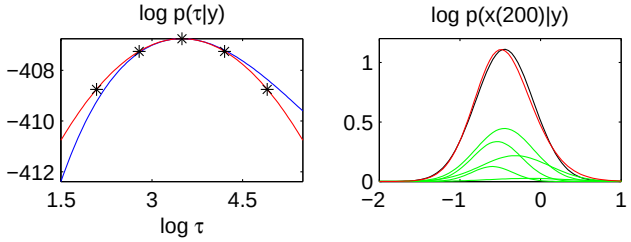
$$x_i|\boldsymbol{y},\boldsymbol{\vartheta} \overset{\text{approx.}}{\sim} \text{N}\left(\mu_i^{(0)}, (\boldsymbol{Q}_{x|y}^{-1})_{ii}\right)$$

- The multifrontal approach (Campbell and Davis, 1995) can be used to compute relevant elements of $\boldsymbol{Q}_{x|y}^{-1}$.

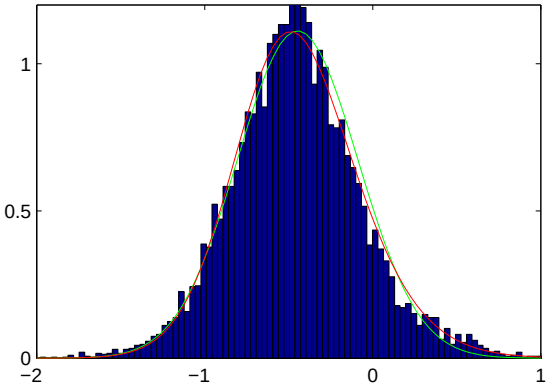
Example: Tokyo rainfall

For the Tokyo rainfall the approximate posterior, $\widetilde{\text{p}}(\tau|\boldsymbol{y})$, is evaluated and an integration grid using 5 points taken at

$$\tau_k = \tau_{\text{MAP}} \pm \frac{k}{\sqrt{-\widetilde{\text{p}}''(\tau|\boldsymbol{y})}} \qquad k = -2,-1,0,1,2$$



Tokyo rainfall — Comparing to MCMC



green: Approximation using one Gaussian.
red: Approximation using five Gaussians.

INLA Details

For the integration two options could be used:

Grid More accurate, but time-consuming.

Central Composite Design (CCD) Faster than grid, often sufficient for computations of $p(x_i|y)$ (less accurate for $p(\theta_i|y)$).

When constructing the integration grid, the Hessian can be adjusted to account for different variances on each side of the mode.

Computations of $p(x_i|y)$ can be improved by using a **skew-Normal** approximation instead of $p_G(x_i|y,\theta)$.

Implemented in the **R-INLA** package.

INLA Assumptions

Latent GMRF:
Sparse of Q (and $Q_{x|y}$) allowing for fast computations.

Conditionally independent observations:
Sparsity in Q translates into sparsity for $Q_{x|y}$ and allows for simple Laplace approximation (no cross-derivatives).

Few parameters (≤ 10)**:**
Needed for the numerical integration.

Using Laplace Approximation to improve MCMC

As an alternative to INLA the Laplace approximation can be used to improve the MCMC sampling and achieve blocking. I will briefly present a few ideas (and potential issues)

Laplace Approximation — Only $\boldsymbol{x}$

A MCMC algorithm using proposals for $\boldsymbol{x}$ based on the Laplace approximation is given by

1. Find the mode $\widehat{\boldsymbol{x}}(\vartheta)$.
2. Draw a proposal, $\boldsymbol{x}^*$ from $\mathsf{N}\left(\widehat{\boldsymbol{x}}(\vartheta), \boldsymbol{Q}^{-1}_{\boldsymbol{x}|\boldsymbol{y}}(\vartheta)\right)$
3. Accept the new sample with probability

$$\alpha(\boldsymbol{x}^*; \boldsymbol{x}^i) = \frac{\mathrm{p}(\boldsymbol{x}^*|\tau, \boldsymbol{y})}{\mathrm{p}(\boldsymbol{x}^i|\tau, \boldsymbol{y})} \cdot \frac{\mathrm{q}(\boldsymbol{x}^i|\tau)}{\mathrm{q}(\boldsymbol{x}^*|\tau)}$$

   where $\boldsymbol{x}^i$ is the current MCMC-sample and

$$\mathrm{q}(\boldsymbol{x}^*|\tau) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}^* - \widehat{\boldsymbol{x}}(\tau))^\top \boldsymbol{Q}_{\boldsymbol{x}|\boldsymbol{y}}(\tau)(\boldsymbol{x}^* - \widehat{\boldsymbol{x}}(\tau))\right)$$

4. Sample the parameters ϑ using, e.g. random-walk (RW) or conjugate-priors.
5. Repeat.

Laplace Approximation — Both $\boldsymbol{x}$ and ϑ

To updated both $\boldsymbol{x}$ and ϑ jointly a suitable strategy is a RW proposal for ϑ followed by sampling from the Laplace approximation obtained for the new ϑ^*. This gives

1. Sample a new ϑ^* using a RW
2. Sample a proposal for $\boldsymbol{x}$ from the Laplace approximation given the new ϑ^* (this includes computing the mode $\widehat{\boldsymbol{x}}(\vartheta^*)$)

$$\boldsymbol{x}^*|\vartheta^* \sim \mathsf{N}\left(\widehat{\boldsymbol{x}}(\vartheta^*), \boldsymbol{Q}^{-1}_{\boldsymbol{x}|\boldsymbol{y}}(\vartheta^*)\right)$$

   The resulting proposal density can be decomposed as

$$\mathrm{q}(\boldsymbol{x}^*, \vartheta^*|\boldsymbol{x}^i, \tau^j) = \mathrm{q}(\boldsymbol{x}^*|\vartheta^*)\mathrm{q}(\vartheta^*|\vartheta^i)$$

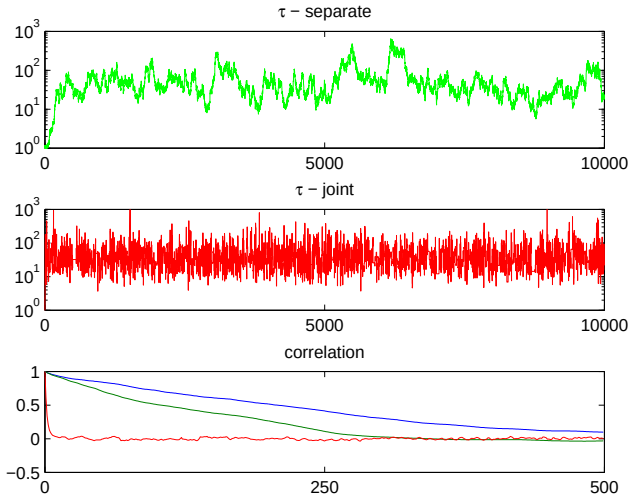
3. Accept the new sample with probability

$$\alpha(\boldsymbol{x}^*, \vartheta^*; \boldsymbol{x}^i, \vartheta^i) = \frac{\mathrm{p}(\boldsymbol{x}^*, \vartheta^*|\boldsymbol{y})}{\mathrm{p}(\boldsymbol{x}^i, \vartheta^i|\boldsymbol{y})} \cdot \frac{\mathrm{q}(\boldsymbol{x}^i|\vartheta^i)\mathrm{q}(\vartheta^i|\vartheta^*)}{\mathrm{q}(\boldsymbol{x}^*|\vartheta^*)\mathrm{q}(\vartheta^*|\vartheta^i)}$$

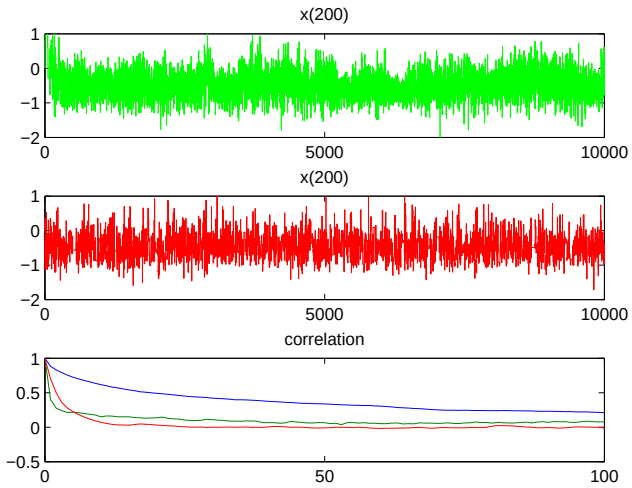
4. Repeat.

A simple adaptive MCMC (Haario et al., 2001; Andrieu and Thoms, 2008) can used to be adjust the proposal variance for the RW.

τ — Laplace Approximation MCMC



x_{200} — Laplace Approximation MCMC



Using Laplace Approximation to improve MCMC

Although popular (especially before INLA) the use of Laplace Approximations in the proposal for MCMC has several drawbacks:

- Potentially expensive mode-finding in each iteration.
- The acceptance rate decreases with increasing size of the latent field, for large applications $\alpha(\boldsymbol{x}^*, \boldsymbol{\vartheta}^*; \boldsymbol{x}^i, \boldsymbol{\vartheta}^i) \ll 0.1$ are common.
- The joint proposal is dependent on good proposals for $\boldsymbol{\vartheta}^*$.
- The separate proposals still have issues with poor mixing between $\boldsymbol{\vartheta}$ and $\boldsymbol{x}$.

Since we're computing derivatives anyway it's often better to use **Metropolis-Adjusted Langevin Algorithm (MALA)** (Girolami and Calderhead, 2011)

Questions?

Bibliography I

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *J. Am. Stat. Assoc.*, 88, 669–679.

Andrieu, C. and Thoms, J. (2008), "A tutorial on adaptive MCMC," *Stat. Comput.*, 18, 343–373.

Campbell, Y. E. and Davis, T. A. (1995), "Computing the sparse inverse subset: an inverse multifrontal approach," Tech. rep., University of Florida, rEP-1995-183.

Gilks, W. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *J. R. Stat. Soc. C*, 41, 337–348.

Girolami, M. and Calderhead, B. (2011), "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *J. R. Stat. Soc. B*, 73, 123–214.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–224.

Held, L. and Holmes, C. C. (2006), "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Anal.*, 1, 145–168.

Bibliography II

Kitagawa, G. (1987), "Non-Gaussian State — Space Modeling of Nonstationary Time Series," *J. Am. Stat. Assoc.*, 82, 1032–1041.

Knorr-Held, L. and Rue, H. (2002), "On Block Updating in Markov Random Field Models for Disease Mapping," *Scand. J. Stat.*, 29, 597–614.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for hierarchical Gaussian Markov random field models," *J. R. Stat. Soc. B*, 71, 1–35.