Computation fundamentals of discrete GMRF representations of continuous domain spatial models

Finn Lindgren

September 23, 2015
v0.2.2

Abstract

The fundamental formulas and algorithms for Bayesian spatial statistics using continuous domain GMRF models are presented. The construction of the models as approximate SPDE solutions is not covered. The results are applicable to non-spatial GMRF models as well, but the important property of observations being *local* and *non-local* is only discussed in the spatial context.

Contents

1 Basic spatial model in GMRF representation

Basic linear spatial model:

- Spatial basis expansion model $u(\boldsymbol{s}) = \sum_{k=1}^{m} \psi_k(\boldsymbol{s}) u_k$, with m basis functions $\psi_k(\boldsymbol{s})$ with local (compact) support, and GMRF coefficients $(u_1, \ldots, u_m) = \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1})$, where $\boldsymbol{Q}_u$ is derived from an SPDE construction with parameters θ_u and $\boldsymbol{\mu}_u$ is usually $\boldsymbol{0}$.
- Covariate matrix $\boldsymbol{B}$, p coefficients $\boldsymbol{b} \sim \mathrm{N}(\boldsymbol{\mu}_b, \boldsymbol{Q}_b^{-1})$. Typically, $\boldsymbol{\mu}_b = \boldsymbol{0}$, and $\boldsymbol{Q}_b = \tau_b \boldsymbol{I}_p$ with $\tau_b \approx 0$.
- n observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ at locations $\boldsymbol{s}_i$, and additive zero mean Gaussian noise e_i with $(e_1, \ldots, e_n) = \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}_\epsilon^{-1})$. Typically, $\boldsymbol{Q}_\epsilon = \tau_\epsilon \boldsymbol{I}$.

  Define $\boldsymbol{A}$ so that $A_{i,j} = \psi_j(\boldsymbol{s}_i)$. The observation model is then

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{b} + \boldsymbol{A}\boldsymbol{u} + \boldsymbol{\epsilon}.$$

Since the covariate effects and the spatial field are jointly Gaussian, it's often practical to join them into a single vector $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{b})$, and

$$\begin{aligned}
\boldsymbol{\mu}_x &= \begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_b \end{bmatrix}, \\
\boldsymbol{Q}_x &= \begin{bmatrix} \boldsymbol{Q}_u & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q}_b \end{bmatrix}, \\
\boldsymbol{A}_x &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix}, \\
\boldsymbol{y} &= \boldsymbol{A}_x \boldsymbol{x} + \boldsymbol{\epsilon}.
\end{aligned}$$

1.1 Bayesian hierarchical formulation

Written as a full Bayesian hierarchical model, the model is

$$\begin{aligned}
\theta = (\theta_u, \tau_b, \tau_\epsilon) &\sim \pi(\theta), \\
(\boldsymbol{x} \mid \theta) &\sim \mathrm{N}(\boldsymbol{\mu}_x, \boldsymbol{Q}_x^{-1}), \\
(\boldsymbol{y} \mid \theta, \boldsymbol{x}) &\sim \mathrm{N}(\boldsymbol{A}_x \boldsymbol{x}, \boldsymbol{Q}_\epsilon^{-1}).
\end{aligned}$$

It will be important to note that for local basis functions (and at most Markov dependent noise $\boldsymbol{\epsilon}$), $\boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}$ is sparse, and that

$$\boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}_x = \begin{bmatrix} \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A} & \boldsymbol{A}^\top \boldsymbol{Q}_\epsilon \boldsymbol{B} \\ \boldsymbol{B}^\top \boldsymbol{Q}_\epsilon \boldsymbol{A} & \boldsymbol{B}^\top \boldsymbol{Q}_\epsilon \boldsymbol{B} \end{bmatrix}$$

is therefore also sparse if p is small.

1.2 Joint distribution

The joint distribution for the observations and the latent variables, $(\boldsymbol{y}, \boldsymbol{x})$ is given by

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{A}_x \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q}_\epsilon & -\boldsymbol{Q}_\epsilon \boldsymbol{A}_x \\ -\boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon & \boldsymbol{Q}_x + \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}_x \end{bmatrix}^{-1} \right)$$

1.3 Conditional/posterior distribution (or Kriging)

The conditional distribution for $\boldsymbol{x}$ given $\boldsymbol{y}$ and θ is given by

$$\begin{aligned}(\boldsymbol{x} \mid \theta, \boldsymbol{y}) &\sim \mathrm{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{Q}_{x|y}^{-1}),\\ \boldsymbol{Q}_{x|y} &= \boldsymbol{Q}_x + \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}_x,\\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{Q}_{x|y}^{-1} \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \boldsymbol{A}_x \boldsymbol{\mu}_x).\end{aligned}$$

The proof is by completing the square in the exponent of the Gaussian conditional density expression

$$\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})}{\pi(\boldsymbol{y} \mid \boldsymbol{\theta})} \propto \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}).$$

Note that the elements of $\boldsymbol{\mu}_{x|y}$ are the basis function coefficients and covariate effect estimates in the universal Kriging predictor, and that we've arrived at it without going "the long way around" via the traditional covariance based Kriging equations and matrix inversion lemmas.

2 Expectations, covariances, and samples

Direct calculations and simulation (for models smaller than approximately $m = 10^6$) is typically done with sparse Cholesky decomposition. Reordering methods are required to achieve sparsity (CAMD, nested dissection, etc.). Here, assume for ease of notation that the nodes are already in a suitable order, and that both the prior and posterior precision decompositions are sparse:

$$\begin{aligned}\boldsymbol{Q}_x &= \boldsymbol{L}_x \boldsymbol{L}_x^\top,\\ \boldsymbol{Q}_{x|y} &= \boldsymbol{L}_{x|y} \boldsymbol{L}_{x|y}^\top,\end{aligned}$$

where all $\boldsymbol{L}_\cdot$ are lower triangular sparse matrices.

See Sections 2.5.2 and 4 for examples of how to handle models where only the prior precision has a sparse Cholesky decomposition. In several instances, the needed adjustments for the use of iterative equation solvers are also mentioned.

2.1 Solving with the posterior precision

Kriging requires solving a linear system

$$\begin{aligned}\boldsymbol{Q}_{x|y} \boldsymbol{z} &= \boldsymbol{w},\\ \boldsymbol{z} &= \boldsymbol{L}_{x|y}^{-\top} \left(\boldsymbol{L}_{x|y}^{-1} \boldsymbol{w} \right),\end{aligned}$$

requiring one forward solve and one backward solve. Several systems with the same posterior precision can be solved simultaneously by having a multi-column $\boldsymbol{w}$. Often, practical implementations avoid actually transposing the large sparse matrices, effectively doing

$$\boldsymbol{z} = \left\{ \left(\boldsymbol{L}_{x|y}^{-1} \boldsymbol{w} \right)^\top \boldsymbol{L}_{x|y}^{-1} \right\}^\top$$

instead.

The Kriging predictor (the posterior expectation) is obtained for $\boldsymbol{w} = \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y} - \boldsymbol{A}_x \boldsymbol{\mu}_x)$, with $\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{z}$.

2.2 Prior and posterior variances

Even though the full covariance matrix is too large to be stored, the variances and covariances of neighbouring Markov nodes can be calculated, using an algorithm related to the Cholesky decomposition. In R-INLA, this is available as $\boldsymbol{S}$=`inla.qinv(Q)`, where

$$S_{ij} = \begin{cases} \left(\boldsymbol{Q}^{-1}\right)_{ij}, & \text{when } Q_{ij} \neq 0 \text{ or } (i,j) \text{ is in the Cholesky infill, and} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This means that the posterior neighbour covariances $\mathsf{Cov}(x_i, x_j \mid \boldsymbol{\theta}, \boldsymbol{y})$ for neighbours (i,j) are obtained by calling `inla.qinv()` with the posterior precision $\boldsymbol{Q}_{x|y}$.

Provided that the desired posterior predictions $\boldsymbol{A}_x^{\text{pred}}\boldsymbol{x}$ have the same sparse connectivity structure as the prior model itself, e.g. when $\boldsymbol{A}_x^{\text{pred}} = \boldsymbol{A}_x$, the sparse symmetric $\boldsymbol{S}$ matrix is sufficient to evaluate marginal posterior predictive variances, as the diagonal of $\boldsymbol{A}_x^{\text{pred}}\boldsymbol{S}(\boldsymbol{A}_x^{\text{pred}})^\top$:

$$\mathsf{Var}\left((\boldsymbol{A}_x^{\text{pred}}\boldsymbol{x})_i\right) = \sum_{j=1}^{m+p} (\boldsymbol{A}_x^{\text{pred}})_{ij} \sum_{k=1}^{m+p} S_{jk} (\boldsymbol{A}_x^{\text{pred}})_{ik},$$

which in R would typically be implemented as `rowSums(A * (A %*% S))`.

When the desired predictions are dense relative to the model (e.g. when predicting the overall spatial average) solves with $\boldsymbol{L}_{x|y}$ are required. A small number of predictions (i.e. the number of rows of $\boldsymbol{A}_x^{\text{pred}}$ is small) can be evaluated simultaneously with

```
LAt <- solve(L, t(A))
colSums(LAt * LAt)
```

Data predictive marginal variances, for $\boldsymbol{A}_x^{\text{pred}}\boldsymbol{x} + \boldsymbol{\epsilon}^{\text{pred}}$, are obtained by adding τ_ϵ^{-1} to the variances for $\boldsymbol{A}_x^{\text{pred}}\boldsymbol{x}$.

2.3 Leave-one-out cross-validation

In the case of conditionally independent observations, expressions for expectations and variances of leave-one-out cross-validation predictive distributions can be obtained via rank one modification of the full posterior precision matrix, and are obtained "for free" by manipulating the above expressions.

Let $\boldsymbol{S}_{x|y}$ be the neighbour covariance matrix obtained from (1) applied to $\boldsymbol{Q}_{x|y}$, and define

$$\begin{aligned} E_i &= (\boldsymbol{A}_x \boldsymbol{\mu}_{x|y})_i, \\ V_i &= (\boldsymbol{A}_x)_{i\cdot} \boldsymbol{S}_{x|y} (\boldsymbol{A}_x)_{i\cdot}^\top. \end{aligned}$$

Then the predictive distribution for x_i based on the complete data is

$$((\boldsymbol{A}_x\boldsymbol{x})_i \,|\, \boldsymbol{\theta}, \boldsymbol{y}) \sim \mathrm{N}(E_i, V_i),$$

and the leave-one-out predictive distribution is

$$\left((\boldsymbol{A}_x\boldsymbol{x})_i \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}_j, j \neq i\right) \sim \mathrm{N}(E_{(i)}, V_{(i)}),$$

where

$$\begin{aligned} E_{(i)} &= y_i - \frac{y_i - E_i}{1 - q_i V_i}, \\ V_{(i)} &= \frac{V_i}{1 - q_i V_i}. \end{aligned}$$

where q_i is the conditional precision of observation i, from $\boldsymbol{Q}_\epsilon = \mathrm{diag}(q_1, \ldots, q_n)$.

2.4 Sampling from the prior

Let $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{m+p})$, then

$$\boldsymbol{x} = \boldsymbol{\mu}_x + \boldsymbol{L}_x^{-\top} \boldsymbol{w}$$

is a sample from the prior model for $\boldsymbol{x}$. Multiple samples are obtained by repeated forward solves.

2.5 Sampling from the posterior

2.5.1 Direct method

First compute $\boldsymbol{\mu}_{x|y}$ as in Section 2.1, and then let $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{m+p})$. Then

$$\boldsymbol{x} = \boldsymbol{\mu}_{x|y} + \boldsymbol{L}_{x|y}^{-\top} \boldsymbol{w}$$

is a sample from the posterior for $\boldsymbol{x}$ given $\boldsymbol{y}$. Multiple samples are obtained by repeated forward solves.

2.5.2 Least squares

In some models, the posterior precision matrix does not have a sparse Cholesky decomposition, but iterative linear algebra methods may still be used to solve systems with $\boldsymbol{Q}_{x|y}$. If the prior precision can be decomposed as $\boldsymbol{Q}_x = \boldsymbol{H}_x^\top \boldsymbol{H}_x$ for some sparse matrix $\boldsymbol{H}_x$ (not necessarily a Cholesky factor), then one approach to sampling is least squares. Given that the posterior expectation has already been calculated using an iterative solution to the equations in Section 1.3, a sample of $\boldsymbol{x}$ conditionally on $\boldsymbol{y}$ is given by the solution to

$$\boldsymbol{Q}_{x|y}(\boldsymbol{x} - \boldsymbol{\mu}_{x|y}) = \begin{bmatrix} \boldsymbol{H}_x & \boldsymbol{A}_x^\top \boldsymbol{L}_\epsilon \end{bmatrix} \boldsymbol{w},$$

where $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{m+p+n})$.

2.5.3 Sampling with conditioning by Kriging

The *conditioning by Kriging* approach to sampling is to construct samples from the prior and then *correct* them to become samples from the posterior. For GMRF models, this is less efficient than the direct approach in Section 2.5.1, except when some observations are non-local, which leads to a dense posterior precision.

In its traditional form, the samples are constructed as follows:

$$\begin{aligned}
\boldsymbol{x}^* &= \boldsymbol{\mu}_x + \boldsymbol{L}_x^{-\top} \boldsymbol{w}_x, & \boldsymbol{w}_x &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{m+p}), \\
\boldsymbol{y}^* &= \boldsymbol{y} + \boldsymbol{L}_\epsilon^{-\top} \boldsymbol{w}_y, & \boldsymbol{w}_y &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n), \\
\boldsymbol{x} &= \boldsymbol{x}^* - \boldsymbol{Q}_x^{-1} \boldsymbol{A}_x^\top \left(\boldsymbol{A}_x \boldsymbol{Q}_x^{-1} \boldsymbol{A}_x^\top + \boldsymbol{Q}_\epsilon^{-1} \right)^{-1} (\boldsymbol{A}_x \boldsymbol{x}^* - \boldsymbol{y}^*)
\end{aligned}$$

where the final expression is the covariance based Kriging equation. For anything but very small n, the inner matrix is dense. Using the Woodbury identity, the expression can be rewritten to recover the precision based equation from Section 1.3,

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{x}^* + \left(\boldsymbol{Q}_x + \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon \boldsymbol{A}_x \right)^{-1} \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y}^* - \boldsymbol{A}_x \boldsymbol{x}^*) \\
&= \boldsymbol{x}^* + \boldsymbol{Q}_{x|y}^{-1} \boldsymbol{A}_x^\top \boldsymbol{Q}_\epsilon (\boldsymbol{y}^* - \boldsymbol{A}_x \boldsymbol{x}^*),
\end{aligned}$$

which is calculated as before, with forward and backward solves on $\boldsymbol{L}_{x|y}$.

If N samples, as well as the posterior mean, are desired, conditioning by Kriging requires N solves with $\boldsymbol{L}_x$ and $\boldsymbol{L}_\epsilon$ (to obtain $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$), and $2 + 2N$ solves with $\boldsymbol{L}_{x|y}$. The direct method only requires $2 + N$ solves with $\boldsymbol{L}_{x|y}$, and no solves with $\boldsymbol{L}_x$ or $\boldsymbol{L}_\epsilon$, which means that it is nearly always preferable to conditioning by Kriging. The exception is non-local observations, which can be handled separately as soft constraints, see Section 4.

3 Likelihood evaluation

The expensive part of evaluating likelihoods is calculating precision determinants, but they come at no extra cost given the Cholesky factors:

$$\begin{aligned}
\log \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) &= -\frac{m+p}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{Q}_x) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^\top \boldsymbol{Q}_x(\boldsymbol{x}-\boldsymbol{\mu}_x), \\
\log\det(\boldsymbol{Q}_x) &= 2\sum_{j=1}^{m+p}\log(\boldsymbol{L}_x)_{jj}.
\end{aligned}$$

Corresponding expressions hold for $\pi(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x})$ and $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$, using $\boldsymbol{L}_\epsilon$ and $\boldsymbol{L}_{x|y}$.

Marginal data likelihoods require more effort.

3.1 Conditional marginal data likelihood

For any arbitrary $\boldsymbol{x}^*$,

$$\begin{aligned}
\pi(\boldsymbol{y} \mid \boldsymbol{\theta}) &= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{\theta})} = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\boldsymbol{x}=\boldsymbol{x}^*} = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\boldsymbol{x}=\boldsymbol{x}^*} \\
&= \frac{\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\boldsymbol{x}=\boldsymbol{x}^*}.
\end{aligned}$$

In practice, this is evaluated for $\boldsymbol{x}^* = \boldsymbol{\mu}_{x|y}$, which gives better numerical stability than $\boldsymbol{x}^* = \boldsymbol{0}$. Written explicitly,

$$\begin{aligned}
\log \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) = &-\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{Q}_x) + \frac{1}{2}\log\det(\boldsymbol{Q}_\epsilon) - \frac{1}{2}\log\det(\boldsymbol{Q}_{x|y}) \\
&- \frac{1}{2}(\boldsymbol{\mu}_{x|y}-\boldsymbol{\mu}_x)^\top \boldsymbol{Q}_x(\boldsymbol{\mu}_{x|y}-\boldsymbol{\mu}_x) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{A}_x\boldsymbol{\mu}_{x|y})^\top \boldsymbol{Q}_\epsilon(\boldsymbol{y}-\boldsymbol{A}_x\boldsymbol{\mu}_{x|y}),
\end{aligned}$$

where an anticipated third quadratic form, from $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$, has been eliminated due to the choice of $\boldsymbol{x}^*$. Comparing with a standard marginal Gaussian log-likelihood expression, the log-determinant and quadratic form can be identified,

$$\begin{aligned}
\log\det(\boldsymbol{Q}_y) &= \log\det(\boldsymbol{Q}_x) + \log\det(\boldsymbol{Q}_\epsilon) - \log\det(\boldsymbol{Q}_{x|y}) \\
(\boldsymbol{y}-\boldsymbol{A}_x\boldsymbol{\mu}_x)^\top \boldsymbol{Q}_y(\boldsymbol{y}-\boldsymbol{A}_x\boldsymbol{\mu}_x) &= (\boldsymbol{\mu}_{x|y}-\boldsymbol{\mu}_x)^\top \boldsymbol{Q}_x(\boldsymbol{\mu}_{x|y}-\boldsymbol{\mu}_x) + (\boldsymbol{y}-\boldsymbol{A}_x\boldsymbol{\mu}_{x|y})^\top \boldsymbol{Q}_\epsilon(\boldsymbol{y}-\boldsymbol{A}_x\boldsymbol{\mu}_{x|y}),
\end{aligned}$$

even though the marginal precision matrix itself, $\boldsymbol{Q}_y$, is dense and cannot be stored.

3.2 Posterior parameter likelihood and total marginal data likelihood

Using the result for the conditional marginal data likelihood, we can express the posterior parameter likelihood as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{y} \mid \boldsymbol{\theta})}{\pi(\boldsymbol{y})},$$

where the only unknown is the total marginal data likelihood $\pi(\boldsymbol{y})$, which can be approximated with numerical integration over $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{y}) = \int \pi(\boldsymbol{\theta})\pi(\boldsymbol{y} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

3.3 A note on non-Gaussian data and INLA

For non-Gaussian data, Laplace approximation can be used, which in essence involves replacing the non-Gaussian conditional posterior $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ with a Gaussian approximation in the expression for $\pi(\boldsymbol{y} \mid \boldsymbol{\theta})$, taking $\boldsymbol{x}^*$ as the mode of $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$.

The INLA implementation finds the mode of $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ using numerical optimisation, and further approximates the marginal posteriors for the latent variables x_k, $\pi(x_k \mid \boldsymbol{y})$, with numerical integration of

$$\pi(x_k \mid \boldsymbol{y}) = \int \pi(x_k \mid \boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta},$$

where a further laplace approximation of $\pi(x_k \mid \boldsymbol{\theta}, \boldsymbol{y})$ is made (as well as skweness corrections not discussed here),

$$\pi(x_k \mid \boldsymbol{\theta}, \boldsymbol{y}) = \left. \frac{\pi(x_k, \boldsymbol{x}_{(k)} \mid \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}_{(k)} \mid \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{x}_k)} \right|_{\boldsymbol{x}_{(k)} = \boldsymbol{x}^*_{(k)}}.$$

This combination of Laplace approximations is the "nested Laplace" part of INLA. In the case of Gaussian data, the only approximation in the method is the numerical integration itself.

4 Linear constraints

There are two main classes of linear constraints:

1. *Hard constraints*, which fall into two subclasses,
   (a) *Interacting hard constraints*, where linear combinations of nodes are constrained,
   (b) *Non-interacting hard constraints*, where each constraint acts only on a single node, in effect specifying those nodes explicitly,
2. *Soft constraints*, where linear combinations are specified with some uncertainty, equivalent to posterior conditioning on observations of these linear combinations. Observations are treated under this framework when they break the Markov structure in the posterior distribution, e.g. by affecting all the nodes in the field directly.

All these cases can be written on matrix form as

$$\boldsymbol{A}_c\boldsymbol{x} + \boldsymbol{\epsilon}_c = \boldsymbol{e}_c,$$

where $\boldsymbol{\epsilon}_c = \mathbf{0}$ in the hard constraint cases, and

$$\boldsymbol{\epsilon}_c \sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}_c^{-1}),$$

in the soft constraint case. The aim is to sample from and evaluate properties of the conditional distribution for $(\boldsymbol{x} \mid \boldsymbol{e}_c)$, when

$$\begin{aligned} \boldsymbol{x} &\sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}), \\ (\boldsymbol{e}_c \mid \boldsymbol{x}) &\sim \mathrm{N}(\boldsymbol{A}_c \boldsymbol{x}, \boldsymbol{Q}_c^{-1}), \end{aligned}$$

where $\boldsymbol{x}$ comes from a known GMRF model, typically either a prior or a posterior distribution.

Let r denote the number of constraints, i.e. the number of rows in $\boldsymbol{A}_c$ and the number of elements in $\boldsymbol{e}_c$.

4.1 Non-interacting hard constraints

Non-interacting hard constraints are easiest to handle, as the constrained nodes can be completely removed from all calculations in a preprocessing step, as they are in effect known deterministic values, and the conditional precision for the remaining nodes can be trivially extracted as a block from the joint precision matrix. Without loss of practical generality, we can assume that all scaling is done in $\boldsymbol{e}_c$, so that each row of $\boldsymbol{A}_c$ has a single non-zero entry that is equal to 1.

For ease of notation, assume that the nodes are ordered so that the unconstrained nodes U are followed by the constrained, C, so that the joint distribution is

$$\begin{bmatrix} \boldsymbol{x}_U \\ \boldsymbol{x}_C \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_U \\ \boldsymbol{\mu}_C \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q}_{UU} & \boldsymbol{Q}_{UC} \\ \boldsymbol{Q}_{CU} & \boldsymbol{Q}_{CC} \end{bmatrix}^{-1} \right).$$

The conditional distribution for the unconstrained nodes is given by

$$\begin{aligned} (\boldsymbol{x}_U \mid \boldsymbol{x}_C = \boldsymbol{e}_c) &\sim \mathrm{N}\left(\boldsymbol{\mu}_{U|C}, \boldsymbol{Q}_{UU}^{-1}\right), \\ \boldsymbol{\mu}_{U|C} &= \boldsymbol{\mu}_U - \boldsymbol{Q}_{UU}^{-1} \boldsymbol{Q}_{UC} (\boldsymbol{e}_c - \boldsymbol{\mu}_C). \end{aligned}$$

The constrained mean and constrained samples are obtained with the same forward and backward solving techniques as in Section 1.3 and Section 2.5.1, using the Cholesky decomposition of $\boldsymbol{Q}_{UU}$. The joint constrained distribution can be written as a degenerate Normal distribution,

$$\left(\begin{bmatrix} \boldsymbol{x}_U \\ \boldsymbol{x}_C \end{bmatrix} \middle| \, \boldsymbol{x}_C = \boldsymbol{e}_c \right) \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{U|C} \\ \boldsymbol{e}_c \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q}_{UU}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right).$$

4.2 Conditioning by Kriging

Conditioning on local observations preserves the Markov structure, and Markov breaking conditioning is therefore done as a final step, when computing posterior means, and sampling from the posterior as well as from the prior. Using the conditioning by Kriging introduced in Section 2.5.3, conditioning on linear constraints or on non-local observations can be written in a common framework. Let $\boldsymbol{Q}_c = \boldsymbol{L}_c \boldsymbol{L}_c^\top$ be the Cholesky decomposition of $\boldsymbol{Q}_c$ in the soft constraint case, and for notational convenience, define $\boldsymbol{Q}_c^{-1} = \mathbf{0}$ and $\boldsymbol{L}_c^{-1} = \mathbf{0}$ for hard constraints.

The method proceeds by constructing a sample from the unconstrained model, and then correcting for the constraints,

$$\begin{aligned}
\boldsymbol{x}^* &= \boldsymbol{\mu} + \boldsymbol{L}^{-\top}\boldsymbol{w}_x, & \boldsymbol{w}_x &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{m+p}),\\
\boldsymbol{e}_c^* &= \boldsymbol{e}_c + \boldsymbol{L}_c^{-\top}\boldsymbol{w}_c, & \boldsymbol{w}_c &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_r),\\
\boldsymbol{x} &= \boldsymbol{x}^* - \boldsymbol{Q}^{-1}\boldsymbol{A}_c^\top \left(\boldsymbol{A}_c\boldsymbol{Q}^{-1}\boldsymbol{A}_c^\top + \boldsymbol{Q}_c^{-1}\right)^{-1}(\boldsymbol{A}_c\boldsymbol{x}^* - \boldsymbol{e}_c^*).
\end{aligned}$$

In order to obtain the conditional mean instead of a sample, set $\boldsymbol{x}^* = \boldsymbol{\mu}$ and $\boldsymbol{e}_c^* = \boldsymbol{e}_c$. The dense but small inner $r \times r$ matrix can be evaluated using a single forward solve with $\boldsymbol{L}$,

$$\boldsymbol{A}_c\boldsymbol{Q}^{-1}\boldsymbol{A}_c^\top + \boldsymbol{Q}_c^{-1} = \left(\boldsymbol{L}^{-1}\boldsymbol{A}_c^\top\right)^\top \boldsymbol{L}^{-1}\boldsymbol{A}_c^\top + \boldsymbol{Q}_c^{-1}.$$

Defining $\widetilde{\boldsymbol{A}}_c^\top = \boldsymbol{L}^{-1}\boldsymbol{A}_c^\top$, the full expression becomes

$$\boldsymbol{x} = \boldsymbol{x}^* - \boldsymbol{L}^{-\top}\widetilde{\boldsymbol{A}}_c^\top \left(\widetilde{\boldsymbol{A}}_c\widetilde{\boldsymbol{A}}_c^\top + \boldsymbol{Q}_c^{-1}\right)^{-1}(\boldsymbol{A}_c\boldsymbol{x}^* - \boldsymbol{e}_c^*),$$

where the remaining inner $r \times r$ solve is evaluated using, e.g., Cholesky decomposition, and the result fed into a final backward solve with $\boldsymbol{L}^\top$.

In order to avoid unnecessary recalculations for multiple samples, further precomputation can be used. A fully Cholesky based approach that also simplifies later calulations is

$$\begin{aligned}
\widetilde{\boldsymbol{L}}_c\widetilde{\boldsymbol{L}}_c^\top &= \widetilde{\boldsymbol{A}}_c\widetilde{\boldsymbol{A}}_c^\top + \boldsymbol{L}_c^\top\boldsymbol{L}_c^{-1},\\
\boldsymbol{B} &= \boldsymbol{L}^{-\top}\widetilde{\boldsymbol{A}}_c^\top\widetilde{\boldsymbol{L}}_c^{-\top},\\
\boldsymbol{x} &= \boldsymbol{x}^* - \boldsymbol{B}\widetilde{\boldsymbol{L}}_c^{-1}(\boldsymbol{A}_c\boldsymbol{x}^* - \boldsymbol{e}_c^*),
\end{aligned}$$

where $\widetilde{\boldsymbol{L}}_c$ is a dense $r \times r$ Cholesky factor, and $\boldsymbol{B}$ is dense, $(m+p) \times r$. For models that require iterative solves for $\boldsymbol{Q}$ instead of Cholesky decomposition, the pre-calculations are modified slightly,

$$\begin{aligned}
\widetilde{\widetilde{\boldsymbol{A}}}_c^\top &= \boldsymbol{Q}^{-1}\boldsymbol{A}_c^\top\\
\widetilde{\boldsymbol{L}}_c\widetilde{\boldsymbol{L}}_c^\top &= \boldsymbol{A}_c\widetilde{\widetilde{\boldsymbol{A}}}_c^\top + \boldsymbol{L}_c^\top\boldsymbol{L}_c^{-1},\\
\boldsymbol{B} &= \widetilde{\widetilde{\boldsymbol{A}}}_c^\top\widetilde{\boldsymbol{L}}_c^{-\top},
\end{aligned}$$

but the final step for constructing $\boldsymbol{x}$ remains unchanged.

4.3 Conditional covariances

From the conditioning by Kriging algorithm, we obtain

$$\mathsf{Cov}(\boldsymbol{x}, \boldsymbol{x} \mid \boldsymbol{e}_c) = \boldsymbol{Q}^{-1} - \boldsymbol{B}\boldsymbol{B}^\top,$$

which can be combined with the neighbour covariance algorithm from Section 2.2 to yield the conditional covariances between nodes that are neighbours in the unconstrained model,

$$(\boldsymbol{S}_c)_{ij} = \begin{cases} \boldsymbol{S}_{ij} - \sum_k \boldsymbol{B}_{ik}\boldsymbol{B}_{jk}, & \text{when } Q_{ij} \neq 0 \text{ or } (i,j) \text{ is in the Cholesky infill, and} \\ 0 & \text{otherwise.} \end{cases}$$

4.4 Constrained likelihood evaluation

The constrained likelihood is given by

$$\pi(\boldsymbol{x} \mid \boldsymbol{e}_c) = \frac{\pi(\boldsymbol{x})\pi(\boldsymbol{e}_c \mid \boldsymbol{x})}{\pi(\boldsymbol{e}_c)},$$

where $\pi(\boldsymbol{e}_c \mid \boldsymbol{x})$ is a degenerate density in the hard constraint case.

The marginal distributions for $\boldsymbol{x}$ and $\boldsymbol{e}_c$ are

$$\begin{aligned}\boldsymbol{x} &\sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}),\\ \boldsymbol{e}_c &\sim \mathrm{N}(\boldsymbol{A}_c\boldsymbol{\mu}, \widetilde{\boldsymbol{L}}_c\widetilde{\boldsymbol{L}}_c^\top),\end{aligned}$$

and the densities can be evaluated with the usual approach, with

$$\begin{aligned}\log \pi(\boldsymbol{x}) &= -\frac{m+p}{2}\log(2\pi) + \log\det(\boldsymbol{L}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}),\\ \log \pi(\boldsymbol{e}_c) &= -\frac{r}{2}\log(2\pi) - \log\det(\widetilde{\boldsymbol{L}}_c) - \frac{1}{2}(\boldsymbol{e}_c - \boldsymbol{A}_c\boldsymbol{\mu})^\top \widetilde{\boldsymbol{L}}_c^{-\top}\widetilde{\boldsymbol{L}}_c^{-1}(\boldsymbol{e}_c - \boldsymbol{A}_c\boldsymbol{\mu}).\end{aligned}$$

In the hard constraint case, Section 2.3.3 of (Rue and Held, 2005) provides the degenerate conditional density through

$$\log \pi(\boldsymbol{e}_c \mid \boldsymbol{x}) = \log \pi(\boldsymbol{A}_c\boldsymbol{x} \mid \boldsymbol{x}) = -\frac{1}{2}\log\det(\boldsymbol{A}_c\boldsymbol{A}_c^\top)$$

which can be evaluated using the Cholesky decomposition of the $r \times r$ matrix $\boldsymbol{A}_c\boldsymbol{A}_c^\top$. In the soft constraint case,

$$\log \pi(\boldsymbol{e}_c \mid \boldsymbol{x}) = -\frac{r}{2}\log(2\pi) + \log\det(\boldsymbol{L}_c) - \frac{1}{2}(\boldsymbol{e}_c - \boldsymbol{A}_c\boldsymbol{x})^\top \boldsymbol{Q}_c(\boldsymbol{e}_c - \boldsymbol{A}_c\boldsymbol{x}).$$

References

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields; Theory and Applications*. Chapman & Hall/CRC.