Lecture 11: Conditional Dependency Graphs and More INLA

Review: Bayesian Hierarchical models

Hierarchical models are an extremely useful tool in Bayesian model building.

Three parts:

- Observation model $\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1$: Encodes information about observed data.
- The latent model $\boldsymbol{x}|\boldsymbol{\theta}_2$: The unobserved process.
- Hyperpriors for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2)$: Models for all of the parameters in the observation and latent processes.

Note: here we indicate the observed data by $\boldsymbol{y}$ while $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are parameters

Bayesian Hierarchical models

Unless otherwise specified or implied:

- Conditional independence is assumed
- Prior parameters, $\boldsymbol{\theta}$, are independent except when conditioning of the responses, $\boldsymbol{y}$

Hierarchical Bayesian models - Tokyo rainfall example

Tokyo rainfall example from exercise 2

- y_t number of times daily rainfall in Tokyo ≥ 1 mm, $t = 1, \ldots, 366$
- τ_t logit probability of exceeding 1 mm $t = 1, \ldots, 366$
- n_t number of trials, $t = 1, \ldots, 366$
- $\pi(\tau_t) = \frac{1}{1+\exp(-\tau_t)}$

Hierarchical Bayesian models - Tokyo rainfall example

Tokyo rainfall example from exercise 2

- y_t number of times daily rainfall in Tokyo ≥ 1 mm, $t = 1, \ldots, 366$
- τ_t logit probability of exceeding 1 mm $t = 1, \ldots, 366$
- n_t number of trials, $t = 1, \ldots, 366$
- $\pi(\tau_t) = \frac{1}{1+\exp(-\tau_t)}$

Model:

$$y_t \mid \tau_t \sim \text{Bin}(n_t, \pi(\tau_t))$$

Hierarchical Bayesian models - Tokyo rainfall example

Tokyo rainfall example from exercise 2

- y_t number of times daily rainfall in Tokyo ≥ 1 mm, $t = 1, \ldots, 366$
- τ_t logit probability of exceeding 1 mm $t = 1, \ldots, 366$
- n_t number of trials, $t = 1, \ldots, 366$
- $\pi(\tau_t) = \frac{1}{1+\exp(-\tau_t)}$

Model:

$$y_t \mid \tau_t \sim \text{Bin}(n_t, \pi(\tau_t))$$

Prior for τ_t:

$$\tau_t = \tau_{t-1} + u_t, \quad u_t \overset{iid}{\sim} N(0, \sigma_u^2), \quad t = 2, \ldots, 366.$$

Hierarchical Bayesian models - Tokyo rainfall example

Tokyo rainfall example from exercise 2

- y_t number of times daily rainfall in Tokyo ≥ 1 mm, $t = 1, \ldots, 366$
- τ_t logit probability of exceeding 1 mm $t = 1, \ldots, 366$
- n_t number of trials, $t = 1, \ldots, 366$
- $\pi(\tau_t) = \frac{1}{1+\exp(-\tau_t)}$

Model:

$$y_t \mid \tau_t \sim \text{Bin}(n_t, \pi(\tau_t))$$

Prior for τ_t:

$$\tau_t = \tau_{t-1} + u_t, \quad u_t \overset{iid}{\sim} N(0, \sigma_u^2), \quad t = 2, \ldots, 366.$$

Hyper-prior on σ_u^2:

$$\sigma_u^2 \sim \text{Inv-Gamma}(\alpha, \beta)$$

Hierarchical Bayesian models - Tokyo rainfall example

Tokyo rainfall example from exercise 2

- y_t number of times daily rainfall in Tokyo ≥ 1 mm, $t = 1, \ldots, 366$
- τ_t logit probability of exceeding 1 mm $t = 1, \ldots, 366$
- n_t number of trials, $t = 1, \ldots, 366$
- $\pi(\tau_t) = \frac{1}{1+\exp(-\tau_t)}$

Model:

$$y_t \mid \tau_t \sim \text{Bin}(n_t, \pi(\tau_t))$$

Prior for τ_t:

$$\tau_t = \tau_{t-1} + u_t, \quad u_t \overset{iid}{\sim} N(0, \sigma_u^2), \quad t = 2, \ldots, 366.$$

Hyper-prior on σ_u^2:

$$\sigma_u^2 \sim \text{Inv-Gamma}(\alpha, \beta)$$

Use conditional dependency graphs to visualize the conditional independence structure!

 See notes

Review: INLA

What is it? A numerical method to do fast approximate Bayesian inference

Why? We do not want to wait for the MCMC to converge.

Where can it be applied? The (wide) class of Latent Gaussian Models (a subclass of Bayesian hierarchical models)

How does it work? Uses GMRF and sparse matrix computations, Laplace approximation, numerical integration

How do we use it Already implemented in the `R-INLA` library

Review: Ingredients of INLA

- Latent Gaussian Models
  - Class of models where INLA can be applied
- Gaussian Markov Random Fields
  - Sparse matrix computations
- Laplace Approximation
  - Method of approximating posterior

Latent Gaussian Models: A Unified Framework

Observations: $\boldsymbol{y}$

Latent field: $\boldsymbol{x}$

Hyperparameters: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

Latent Gaussian Models: A Unified Framework

Observations: $\boldsymbol{y}$ Assumed conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}_1$

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1 \sim \prod_i \pi(y_i|x_i, , \boldsymbol{\theta}).$$

Latent field: $\boldsymbol{x}$

Hyperparameters: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

Latent Gaussian Models: A Unified Framework

Observations: $\boldsymbol{y}$ Assumed conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}_1$

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_1 \sim \prod_i \pi(y_i|x_i,,\boldsymbol{\theta}).$$

Latent field: $\boldsymbol{x}$ Assumed to be a GMRF with sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_2)$

$$\boldsymbol{x}|\boldsymbol{\theta}_1 \sim \mathcal{N}(0, \boldsymbol{Q}(\boldsymbol{\theta}_2)^{-1})$$

The latent field $\boldsymbol{x}$ can be large $(10^1 - 10^6)$

Hyperparameters: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$

Latent Gaussian Models: A Unified Framework

Observations: $\boldsymbol{y}$ Assumed conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}_1$

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_1 \sim \prod_i \pi(y_i|x_i,,\boldsymbol{\theta}).$$

Latent field: $\boldsymbol{x}$ Assumed to be a GMRF with sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_2)$

$$\boldsymbol{x}|\boldsymbol{\theta}_1 \sim \mathcal{N}(0, \boldsymbol{Q}(\boldsymbol{\theta}_2)^{-1})$$

The latent field $\boldsymbol{x}$ can be large $(10^1 - 10^6)$

Hyperparameters: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ Precision parameters of the Gaussian field and parameters of the likelihood

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$$

The vector $\boldsymbol{\theta}$ is usually small (1-10)

Latent Gaussian models

A very general way of specifying the problem is by modelling the mean for the i-th unit by means of an additive linear predictor, defined on a suitable scale (e.g. logistic for binomial data)

$$\eta_i = \alpha + \sum_{l=1}^{L} f_l(u_{li}) + \sum_{k=1}^{K} \beta_k z_{ki} + \epsilon_i$$

where

- α is the intercept
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$ quantify the effect of $\boldsymbol{x} = (x_1, \dots, x_K)$ on the response
- $\boldsymbol{f} = (f_1, \dots, f_L)$ is a set of functions defined in terms of some covariates $\boldsymbol{z} = (z_1, \dots, z_K)$

And assume

$$\boldsymbol{x} = (\alpha, \boldsymbol{\beta}, \boldsymbol{f}) \sim \mathcal{N}(0, \boldsymbol{Q}(\theta)^{-1})$$

Quantities of interest:

The posterior distribution is:

$$\pi(\theta, \mathrm{x}|\mathrm{y}) \propto \pi(\mathrm{y}|\theta, \mathrm{x})\pi(\mathrm{x}|\theta)\pi(\theta)$$

We want to approximate the posterior **marginals**

$$\pi(\theta_i|\mathrm{y}) = \int \pi(\theta|\mathrm{y}) d\theta_{-\mathrm{i}}$$

and

$$\pi(x_i|\mathrm{y}) = \int \pi(\mathrm{x_i}|\theta, \mathrm{y})\pi(\theta|\mathrm{y}) d\theta$$

INLA strategy:

- approximate $\pi(\theta|\mathrm{y})$ and $\pi(x_i|\theta, \mathrm{y})$
- solve the integrals numerically

Approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

- From $\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \times \pi(\boldsymbol{\theta}|\boldsymbol{y}) \times \pi(\boldsymbol{y})$ it follows that

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \text{ for all } \boldsymbol{x}.$$

Approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

- From $\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \times \pi(\boldsymbol{\theta}|\boldsymbol{y}) \times \pi(\boldsymbol{y})$ it follows that

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \text{ for all } \boldsymbol{x}.$$

- INLA approximates $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ using

$$\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})}.$$

which is also known as Laplace approximation.

Approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

- From $\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \times \pi(\boldsymbol{\theta}|\boldsymbol{y}) \times \pi(\boldsymbol{y})$ it follows that

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \text{ for all } \boldsymbol{x}.$$

- INLA approximates $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ using

$$\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})}.$$

which is also known as Laplace approximation.

- Here $\widetilde{\pi}_G$ is the Gaussian (GMRF) approximation to $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$.

The GMRF approximation

Let $\boldsymbol{x}$ denote a GMRF with precision matrix $\boldsymbol{Q}$ and mean $\boldsymbol{\mu}$.

Approximate

$$\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \sum_{i=1}^{n} \log \pi(y_i|x_i)\right)$$

by using a second-order Taylor expansion of $\log \pi(y_i|x_i)$ around $\boldsymbol{\mu}_0$, say.

Recall

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 = a + bx - \frac{1}{2}cx^2$$

with $b = f'(x_0) - f''(x_0)x_0$ and $c = -f''(x_0)$.

The GMRF approximation (II)

Thus,

$$\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \sum_{i=1}^{n}(a_i + b_i x_i - 0.5 c_i x_i^2)\right)$$

$$\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right)$$

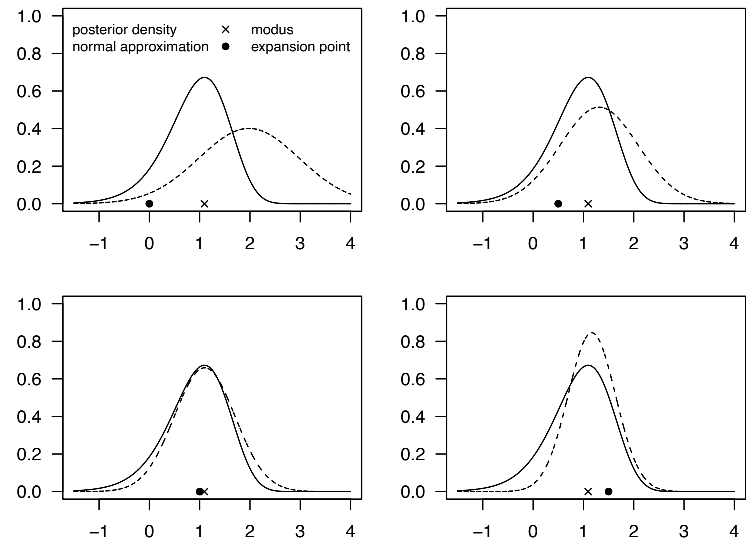
to get a Gaussian approximation with precision matrix $\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c})$ and mean given by the solution of $(\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$. The canonical parameterization is

$$\mathcal{N}_C(\mathrm{b}, \boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))$$

which corresponds to

$$\mathcal{N}((\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))^{-1}\mathrm{b}, (\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))^{-1}).$$

The GMRF approximation



Exploring $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is "numerically explored" to find suitable support points $\boldsymbol{\theta}_k$.

Main use: Select good evaluation points $\boldsymbol{\theta}_k$ for the numerical integration when approximating $\tilde{\pi}(x_i|\boldsymbol{y})$

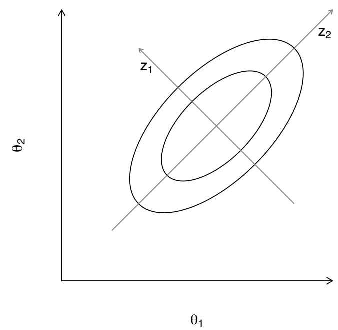
- Locate the mode

Exploring $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is "numerically explored" to find suitable support points $\boldsymbol{\theta}_k$.

Main use: Select good evaluation points $\boldsymbol{\theta}_k$ for the numerical integration when approximating $\tilde{\pi}(x_i|\boldsymbol{y})$

- Locate the mode
- Compute the Hessian to construct principal components

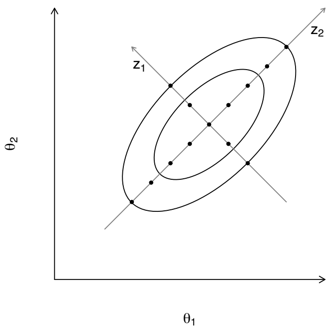


Exploring $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is "numerically explored" to find suitable support points $\boldsymbol{\theta}_k$.

Main use: Select good evaluation points $\boldsymbol{\theta}_k$ for the numerical integration when approximating $\tilde{\pi}(x_i|\boldsymbol{y})$

- Locate the mode
- Compute the Hessian to construct principal components
- Grid-search to locate bulk of the probability mass

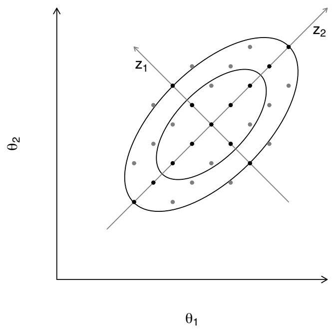


Exploring $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is "numerically explored" to find suitable support points $\boldsymbol{\theta}_k$.

Main use: Select good evaluation points $\boldsymbol{\theta}_k$ for the numerical integration when approximating $\tilde{\pi}(x_i|\boldsymbol{y})$

- Locate the mode
- Compute the Hessian to construct principal components
- Grid-search to locate bulk of the probability mass



All points found have equal area weight Δ_k.

Approximating $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

For approximating the first component $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ we can use

- a Gaussian approximation, easily extractable from $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. However, errors in location and/or lack of skewness possible.

Approximating $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

For approximating the first component $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ we can use

- a Gaussian approximation, easily extractable from $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. However, errors in location and/or lack of skewness possible.
- a Laplace approximation

$$\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{\text{GG}}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^{\star}_{-i}(x_i,\boldsymbol{\theta})}.$$

The approximation is very accurate but very expensive.

Approximating $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

For approximating the first component $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ we can use

- a Gaussian approximation, easily extractable from $\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. However, errors in location and/or lack of skewness possible.
- a Laplace approximation

$$\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{\mathrm{GG}}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^{\star}_{-i}(x_i,\boldsymbol{\theta})} .$$

  The approximation is very accurate but very expensive.
- a simplified Laplace approximation based on fitting a skew-normal distribution to a series expansion of $\tilde{\pi}_{LA}$.

INLA: Overview

Step I Approximate $\pi(\boldsymbol{\theta}|y)$ using the Laplace approximation and select good evaluation points $\boldsymbol{\theta}_k$.

Step II For each $\boldsymbol{\theta}_k$ and i approximate $\pi(x_i|\boldsymbol{\theta}_k, \boldsymbol{y})$ using the Laplace or simplified Laplace approximation for selected values of x_i .

Step III For each i, sum out $\boldsymbol{\theta}_k$

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k, \boldsymbol{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y}) \times \Delta_k.$$

Build a log spline corrected Gaussian to represent $\tilde{\pi}(x_i|\boldsymbol{y})$.

INLA features

INLA fully incorporates posterior uncertainty with respect to hyperparameters $\Rightarrow$ tool for full Bayesian inference

- Marginal posterior densities of all (hyper-)parameters
- Posterior mean, median, quantiles, std. deviation, etc.

The approach can be used for predictions, model assessment, . . .