The Numerical Analysts Guide to Approximate Bayesian Inference

Matthew Thomas

University of Bath

8^{th} April 2016

Bayesian Inference

Bayes' Theorem

This is Bayes' theorem:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$

- For some it is just a theorem,
- For others, it is a way of life!

Bayesian Inference

It allows us to specify a model for some data $\boldsymbol{y}$ in terms of some parameters $\boldsymbol{\theta}$ in a 'Likelihood' function:

$$p(\boldsymbol{y}|\boldsymbol{\theta})$$

and any a-priori knowledge about the model parameters in a Prior probability distribution

$$p(\boldsymbol{\theta})$$

Given these two components we can infer information about the model using Bayes theorem

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

and obtain the posterior, which tells us about the uncertainty around the parameter vector $\boldsymbol{\theta}$ after observing the data $\boldsymbol{y}$

Bayesian Inference

The denominator $p(\boldsymbol{y})$ is the marginal distribution of the observation $\boldsymbol{y}$ which is unknown and consider this a normalisation constant so we often take proportionality wrt $\boldsymbol{\theta}$

$$\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}$$
$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

To obtain a proper posterior distribution we must find $p(\boldsymbol{y})$ the which is of the form

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

This integral is often analytically intractable and thus we must use other techniques to be able to find the posterior

Laplace Approximation

Laplace approximation is an technique to obtain inference by approximating a posterior distribution by a Gaussian distribution.

To make a Laplace approximation we take a Taylor expansion of the density $\log p(\boldsymbol{\theta}|\boldsymbol{y})$ around its mode $\hat{\boldsymbol{\theta}}$, i.e.

$$\log p(\boldsymbol{\theta}|\boldsymbol{y}) = \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \log p(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + h.o.t.$$

Then noting as the mode is the maximum of $p(\boldsymbol{\theta}|\boldsymbol{y})$ then the second term vanishes, i.e.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$$

Laplace approximation

Letting

$$H(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = -\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log p(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

and throw away higher order terms then we see that

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T H(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)$$

which is the kernel of a Gaussian distribution and therefore,

$$\boldsymbol{\theta}|\boldsymbol{y} \sim N(\hat{\boldsymbol{\theta}}, H(\hat{\boldsymbol{\theta}}|\boldsymbol{y})^{-1})$$

- In general, you will need to numerically find the mode (often using Newton optimisation)
- Approximation is accurate *if* the posterior is approximately Gaussian.

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are based on sampling, and are extensively used in Bayesian inference. We aim to sample from the posterior

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

to estimate such as mean and variance.

Some advantages and disadvantages:

- Very flexible with well-known algorithms
- Software available (JAGS, WinBUGS, etc.)
- Not the most efficient (particularly in large-scale problems)
- Issues regarding implementation and convergence of the chain

In many cases, its far easier and efficient to use approximations for Bayesian inference.

Integrated Nested Laplace Approximations

Integrated Nested Laplace Approximations (INLA) is a recent development in approximate Bayesian Inference.

It was introduced by Rue, Martino and Chopin (2009) as an alternative to methods such as MCMC for a general set of statistical models called latent Gaussian models.

Posterior distributions are approximated using a series of Laplace approximations meaning we do not need to sample from the posterior.

It has been shown to be accurate in all but extreme cases and reduced computational burden compared to MCMC.

Software suite called `R-INLA` suite allows implementation in `R`

Hierarchical Models

Latent Gaussian Models

Suppose we have observation vector $\boldsymbol{y}$ that arises from some distribution. We are often interested in estimating the mean $\boldsymbol{\mu}$ which is related to the linear predictor,

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{k=1}^{q} f_k(u_{ki}) + \epsilon_i, \quad i = 1, \ldots, n$$

where

- β_0 is an intercept term
- β_j is the linear effect of covariates x_{ji}
- ϵ_i is the iid noise term (i.e. $\epsilon_i \sim N(0, \sigma_\epsilon^2)$)
- $f_k(\cdot)$ is non-linear function of covariate u_{ji}. We often represent this function as $f_k(s) = \sum_m \gamma_{km} \psi_{km}(s)$ where $\psi_{km}(\cdot)$ are the basis functions and γ_{km} are the weights.

Latent Gaussian Models

By letting,

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T \quad \boldsymbol{z} = (\beta_0, \ldots, \beta_p, \{\gamma_{km}\})^T$$

we can write this as a linear system

$$\boldsymbol{\eta} = A\boldsymbol{z}$$

A model is then classed as a latent Gaussian model if we assign a Gaussian distribution to the vector $\boldsymbol{z}$ i.e.

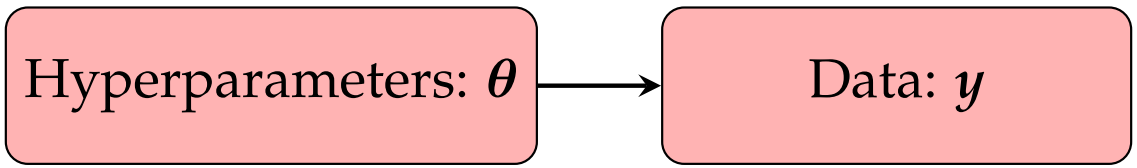
$$\boldsymbol{z} \sim N(\boldsymbol{\mu}, \Sigma)$$

where $\boldsymbol{\mu}$ is the mean vector and Σ is a positive-definite covariance matrix.

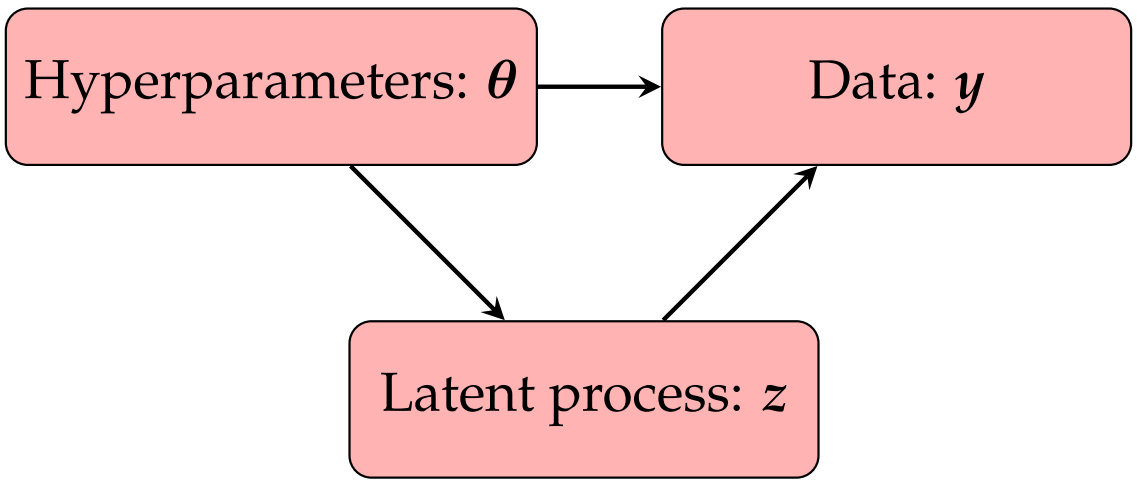
We then define hyperparameters $\boldsymbol{\theta}$ to account for scale of dependency and variability.

This offers a very flexible framework so that we can work with a big range of models.

We're moving from a standard Bayesian model:



to a hierarchical one:



Bayesian Hierarchical Models

This formulation in general is called a Bayesian Hierarchical model, which is the inherent structure of many models.

This is a flexible model that has wide application in statistical modelling.

Bayesian hierarchical models are commonly written in the form:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta} &\sim p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta}) \\ \boldsymbol{z}|\boldsymbol{\theta} &\sim p(\boldsymbol{z}|\boldsymbol{\theta}) \\ \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}) \end{aligned}$$

Bayesian Hierarchical Models

- The observation level $\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}$ - Data $\boldsymbol{y}$, are assumed to arise from the underlying latent (Gaussian) process $\boldsymbol{z}$, which is unobservable, although may be measured with error. For example consider,

$$\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta} \sim N(A\boldsymbol{z}, \sigma_\epsilon^2 I)$$

- The underlying process level $\boldsymbol{z}|\boldsymbol{\theta}$ - The latent process z assumed to drive the observable data and Represents the true value of the quantity of interest. For example consider,

$$\boldsymbol{z}|\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \sigma_z^2 \Sigma)$$

Bayesian Hierarchical Models

- The prior level $\boldsymbol{\theta}$ - This level describes known prior information about the model parameters $\boldsymbol{\theta}$ and controls the scale and variability of the data and the latent process. For example consider,

$$\boldsymbol{\theta} = (\sigma_\epsilon, \sigma_z)^T \sim p(\boldsymbol{\theta})$$

Inference on all model parameters in a hierarchical model such as these can be done as follows.

INFERENCE

We can write the posterior of the model parameters in a similar way as before

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

We are interested in the marginal effects of all the latent process parameters and the hyperparameters

$$p(\theta_i|\boldsymbol{y}) = \iint p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y})\, d\boldsymbol{z}\, d\boldsymbol{\theta}_{-i}, \quad p(z_i|\boldsymbol{y}) = \iint p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y})\, d\boldsymbol{z}_{-i}\, d\boldsymbol{\theta}$$

Typically $dim(\boldsymbol{z}) = 10^2 - 10^6$ and $dim(\boldsymbol{\theta}) \leq 10$ so these are, high-dimensional integrals, so we simplify

Inference

Using the fact that $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) = p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})$ then we see that

$$p(\theta_i|\boldsymbol{y}) = \iint p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y})\, d\boldsymbol{z}\, d\boldsymbol{\theta}_{-i} = \int p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}_{-i}$$

$$p(z_i|\boldsymbol{y}) = \iint p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y})\, d\boldsymbol{z}_{-i}\, d\boldsymbol{\theta} = \int p(z_i|\boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

So instead of having to find $p(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y})$ and do very high dimensional integrals we are just required to find the distributions $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(z_i|\boldsymbol{\theta}, \boldsymbol{y})$ and able to do lower dimensional numerical integration.

GMRFs

Gaussian Random Fields

A random vector $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ is called a Gaussian Markov Field with respect to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in a domain $D \subset \mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ, if and only if its density has the form

$$p(\boldsymbol{z}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)$$

There are some issues with working with this parameterisation in practice, especially when n is large.

- Covariance matrix has $\mathcal{O}(n^2)$ elements.
- Computation often rises $\mathcal{O}(n^3)$ (determinants, inverse, etc.)

Gaussian Random Fields

In some cases it may be better to parameterise the distribution by the precision matrix $Q = \Sigma^{-1}$ (inverse of the covariance),

$$p(\boldsymbol{z}) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T Q(\boldsymbol{z}-\boldsymbol{\mu})\right)$$

By using this parameterisation we have reduced some of the expected computation.

In particular, the non-zero pattern in the precision matrix tells us a lot about the conditional distributional structure.

The Markov property states that if $Q_{ij} = 0$ if and only if z_i and z_j are conditionally independent given all other elements $\boldsymbol{z}_{-ij}$

Gaussian Markov Random Field

A random vector $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ is called a Gaussian Markov Random Field (GMRF) with respect to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in a domain $D \subset \mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and positive definite precision matrix Q, if and only if

$$p(\boldsymbol{z}) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T Q(\boldsymbol{z}-\boldsymbol{\mu})\right)$$

$$Q_{ij} = 0 \iff (i,j) \notin \mathcal{E}$$

To reduce computation, we assume that the latent variables follow a GMRF.

The conditional independence allows for computing with *sparse* matrices. This helps until n gets really large, but we'll return to this later.

Example: AR(1) Process

Suppose we have a Autoregressive process of order 1,

$$z_0 \sim N(0, (1-\alpha)^{-1}\sigma^2)$$
$$\{z_t | z_s, s < t\} \sim N(\alpha z_{t-1}, \sigma^2); t = 1, \ldots, T$$

where $\alpha \in (0, 1)$ and $\sigma > 0$.

$$\Sigma_{ij} = \frac{\alpha^{|i-j|}}{1-\alpha}\sigma^2, \quad Q = \frac{1}{\sigma^2}\begin{bmatrix} 1 & -\alpha & & & \\ -\alpha & 1+\alpha^2 & -\alpha & & \\ & \ddots & \ddots & \ddots & \\ & & -\alpha & 1+\alpha^2 & -\alpha \\ & & & -\alpha & 1 \end{bmatrix}$$

so Σ is dense and Q is sparse.

Integrated Nested Laplace Approximation

Integrated Nested Laplace Approximation

As mentioned earlier, we have to find

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \text{ and } p(z_i|\boldsymbol{\theta}, \boldsymbol{y})$$

to be able to find the marginal posterior distributions

$$p(\theta_i|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}_{-i}, \quad p(z_i|\boldsymbol{y}) = \int p(z_i|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

This is done in three steps.

1. Find an approximation to the distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$
2. Find an approximation to the marginal distributions $p(z_i|\boldsymbol{\theta}, \boldsymbol{y})$
3. Numerically integrate to get marginal distributions

Univariate Example

To demonstrate best how INLA works in practice, we will work with the following hierarchical model.

Suppose that observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are assumed to be independent and identically distributed with

$$y_i | z, \theta \sim N(z, \theta^{-1})$$

Next, suppose we place a Gaussian prior distribution on the latent variable z

$$z | \theta \sim N(\mu_0, \tau_0^{-1})$$

with $\mu_0 \in \mathbb{R}$ and $\tau_0 > 0$ known. The hyperparameter θ is given a Gamma prior

$$\theta \sim Gamma(\alpha, \beta)$$

with $\alpha, \beta > 0$ known.

INLA: Step 1

The first task is to find the joint posterior of the hyperparameters $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$. By applying Bayes theorem and taking proportionality with respect to $\boldsymbol{\theta}$

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \frac{p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})} \\ &\approx \left.\frac{p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p_G(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{z}=\hat{\boldsymbol{z}}(\boldsymbol{\theta})} = \hat{p}(\boldsymbol{\theta}|\boldsymbol{y}) \end{aligned}$$

where $\hat{\boldsymbol{z}}(\boldsymbol{\theta})$ is the mode of $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$ for a given θ.

INLA: Step 1

To find the Gaussian approximation $p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$, we apply Bayes theorem and taking proportionality with respect to $\boldsymbol{z}$

$$\begin{aligned} p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta}) \\ &\approx p_G(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta}) \end{aligned}$$

As the approximation (right hand side) is proportional to the product of two Gaussians, we know that the posterior will be Gaussian

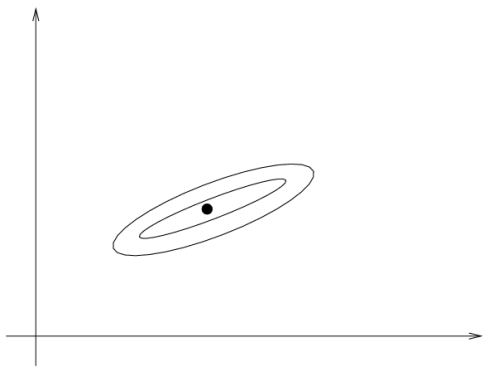
$$p_G(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y}) \sim N(\boldsymbol{\mu}_{\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y}}, Q_{\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y}})$$

This approximation turns out to be accurate in the vast majority of cases due to the underlying process being Gaussian.

INLA: Step 1

We now aim to explore $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ to find 'good' integration points $\{\theta^{(j)}\}$ for the numerical integration later.

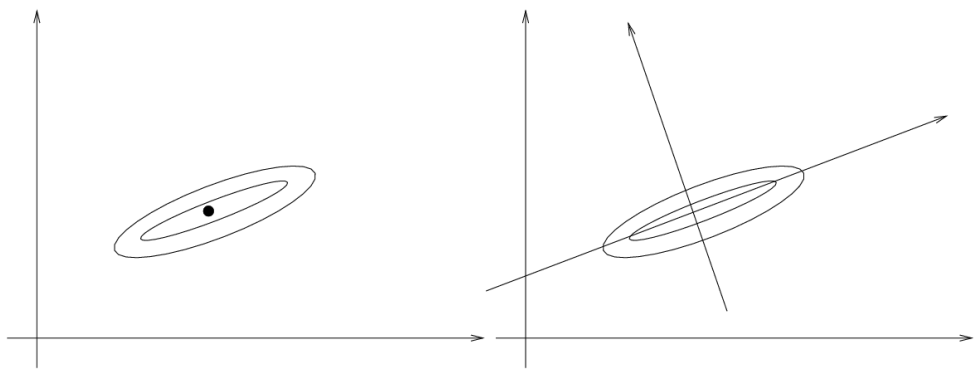
1. Find the mode of $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ (using a Newton optimisation)



INLA: Step 1

We now aim to explore $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ to find 'good' integration points $\{\theta^{(j)}\}$ for the numerical integration later.

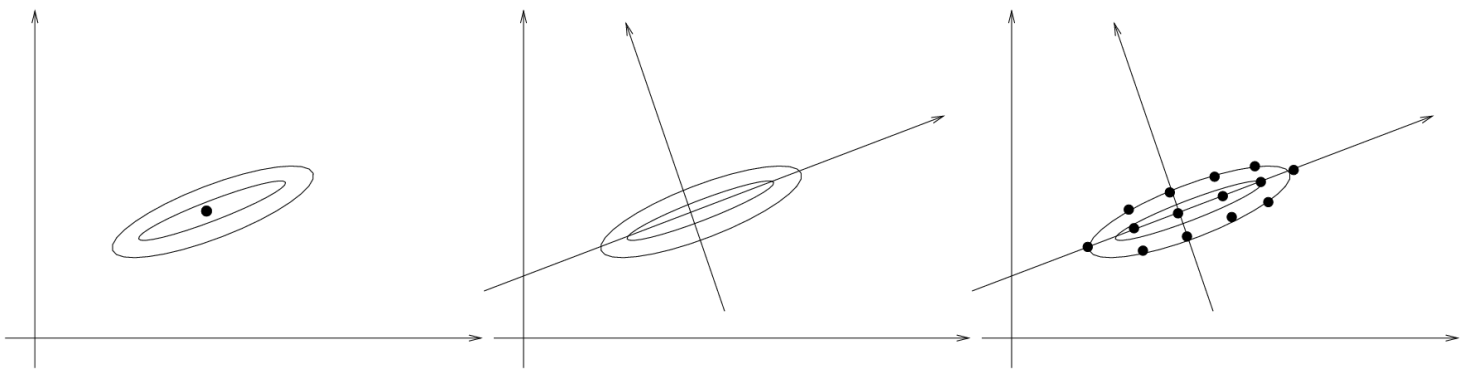
1. Find the mode of $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ (using a Newton optimisation)
2. Compute the Hessian at the mode to be able to define new variables for search direction.



INLA: Step 1

We now aim to explore $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ to find 'good' integration points $\{\theta^{(j)}\}$ for the numerical integration later.

1. Find the mode of $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ (using a Newton optimisation)
2. Compute the Hessian at the mode to be able to define new variables for search direction.
3. Perform the grid search



Univariate Example

Firstly, we derive an expression for the posterior of the latent parameter z given θ and $\boldsymbol{y}$,

$$\begin{aligned} p(z|\theta,\boldsymbol{y}) &\propto p(\boldsymbol{y}|z,\theta)p(z|\theta) = \prod_{i=1}^{n} p(y_i|z,\theta)p(z|\theta) \\ &= \frac{\sqrt{\tau_0\theta}}{2\pi}\exp\left(-\frac{\theta}{2}\sum_{i=1}^{n}(y_i-z)^2\right)\exp\left(-\frac{\tau_0}{2}(z-\mu_0)^2\right) \\ &\propto \exp\left(-\frac{(n\theta+\tau_0)}{2}\left(z^2-2\frac{\theta\sum_i y_i+\mu_0\tau_0}{n\theta+\tau_0}z\right)\right) \end{aligned}$$

which we recognise as the kernel of the Gaussian

$$z|\theta,\boldsymbol{y} \sim N\left(\frac{\theta\sum_i y_i+\mu_0\tau_0}{n\theta+\tau_0}, \frac{1}{n\theta+\tau_0}\right)$$

Univariate Example

We now approximate the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\begin{aligned}\hat{p}(\theta|\boldsymbol{y}) &\approx \left.\frac{p(\boldsymbol{y}|z,\theta)p(z|\theta)p(\theta)}{p(z|\theta,\boldsymbol{y})}\right|_{z=\hat{z}(\boldsymbol{\theta})} \\ &= \sqrt{\frac{2\pi}{n\theta+\tau_0}}\left(\prod_{i=1}^{n} p(y_i|z,\theta)\right)p(z|\theta)p(\theta)\end{aligned}$$

Because of conjugacy this has exact form

$$\theta|\boldsymbol{y} \sim Gamma\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{z})^2\right)$$

However, for demonstration, we will explore this distribution and set up an 'approximation' for numerical integration later.

Univariate Example

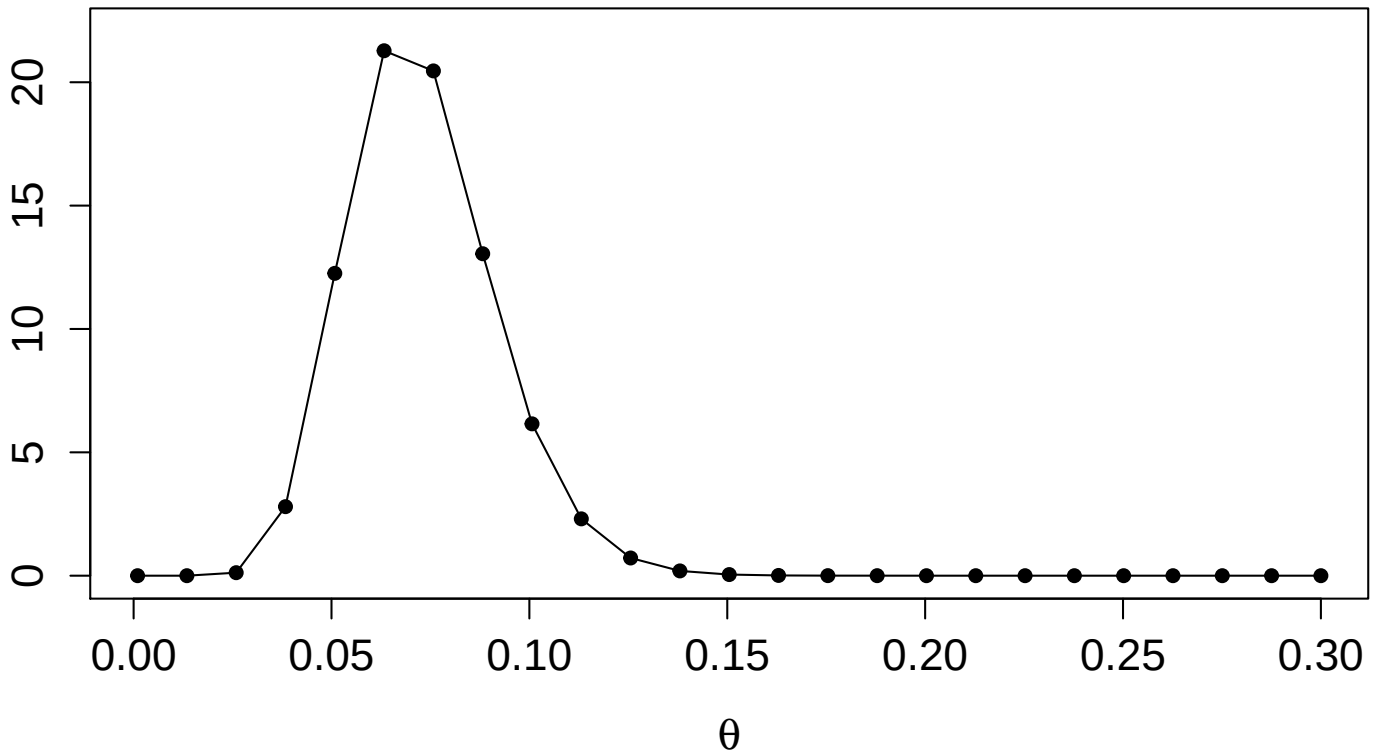


Figure: Approximate posterior distribution $p(\theta|\boldsymbol{y})$ with black dots denoting the set of values $\{\theta\}$ for $\mu_0 = 0$, $\tau_0 = 0.25$, $\alpha = 1.6$ and $\beta = 0.4$

INLA: Step 2

The next task is to find an approximation to the marginal distributions $p(z_i|\boldsymbol{\theta}, \boldsymbol{y})$. Again, we apply Bayes theorem and taking proportionality with respect to z_i

$$\begin{aligned} p(z_i|\boldsymbol{\theta}, \boldsymbol{y}) &\propto \frac{p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})}{p(\boldsymbol{z}_{-i}|z_i, \boldsymbol{\theta}, \boldsymbol{y})} \\ &\approx \left.\frac{p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})}{p_G(\boldsymbol{z}_{-i}|z_i, \boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{z}_{-i}=\boldsymbol{z}_{-i}(z_i,\boldsymbol{\theta})} = \tilde{p}(z_i|\boldsymbol{\theta}, \boldsymbol{y}) \end{aligned}$$

where $\boldsymbol{z}_{-i}(z_i, \boldsymbol{\theta})$ is the mode of $p(\boldsymbol{z}_{-i}|z_i, \boldsymbol{\theta}, \boldsymbol{y})$ for a given z_i and θ.

INLA: Step 2

Similarly to before,

$$p(\boldsymbol{z}_{-i}|z_i, \boldsymbol{\theta}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})$$

which will give us a slightly different configuration than if we were to find it from $p_G(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$

The distribution is then normalised using numerical quadrature

$$\hat{\hat{p}}(z_i|\boldsymbol{\theta}, \boldsymbol{y}) = c\hat{p}(z_i|\boldsymbol{\theta}, \boldsymbol{y})$$

where

$$c = \int \hat{p}(z_i|\boldsymbol{\theta}, \boldsymbol{y})\, dz_i$$

INLA: Step 3

Once we have obtained the approximations

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{y}) \text{ and } \hat{\hat{p}}(z_i|\boldsymbol{\theta}, \boldsymbol{y})$$

for $p(\boldsymbol{\theta}|\boldsymbol{y})$ and $p(z_i|\boldsymbol{\theta}, \boldsymbol{y})$ respectively, the marginal posterior distributions can be approximated.

To find the marginal distribution of $\hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$, we use the grid $\{\boldsymbol{\theta}^{(j)}\}$ explored earlier. We use these points to construct an interpolant to $\log \hat{p}(\boldsymbol{\theta}|\boldsymbol{y})$ and compute marginal distributions using numerical integration of this object.

INLA: Step 3

To find the marginal distribution of $\hat{\hat{p}}(z_i|\boldsymbol{y})$, we use the grid $\{\boldsymbol{\theta}^{(j)}\}$ to numerically integrate as follows

$$\hat{\hat{p}}(z_i|\boldsymbol{y}) \approx \sum_j \hat{\hat{p}}(z_i, |\boldsymbol{\theta}^{(j)}, \boldsymbol{y})\hat{p}(\boldsymbol{\theta}^{(j)}|\boldsymbol{y})\Delta_j$$

where Δ_j are the integration weights.

Accuracy of the marginals depends on the density of the configuration $\{\theta^{(j)}\}$.

Univariate Example

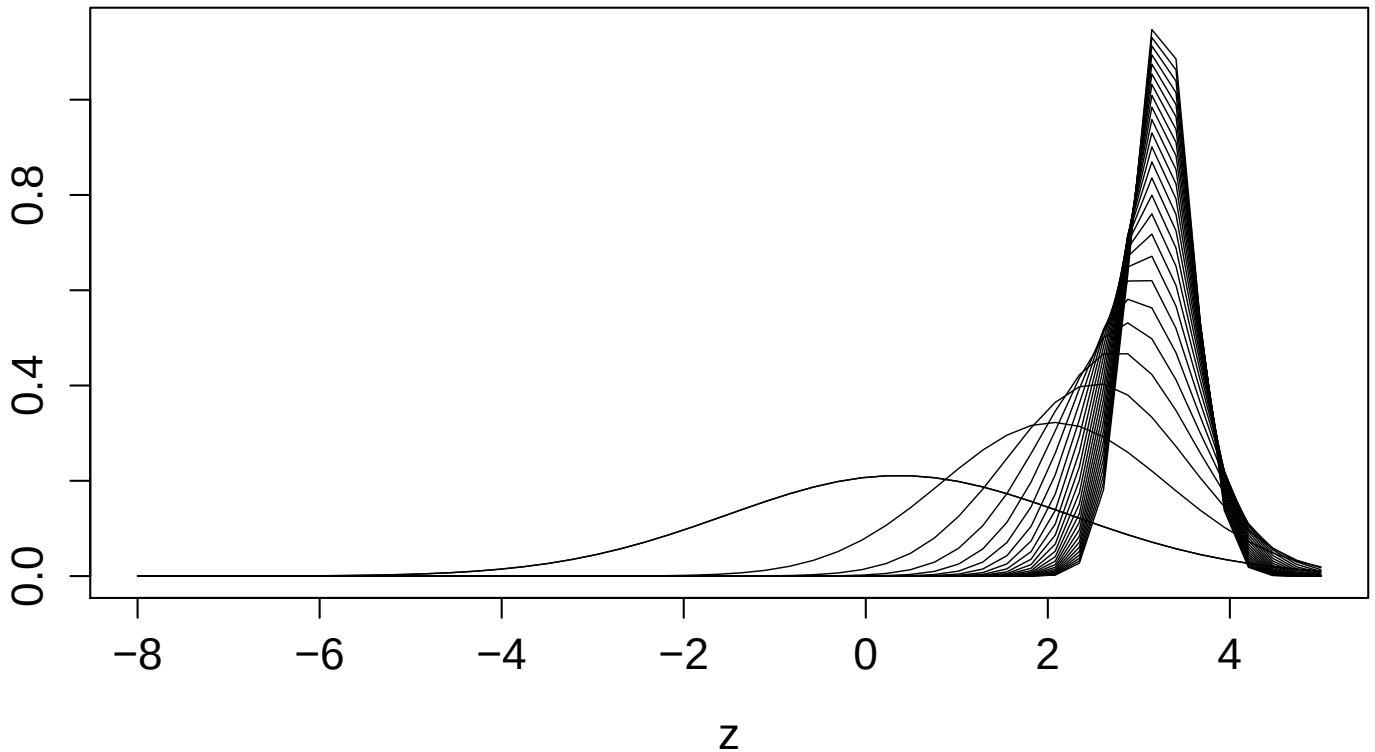


Figure: Approximation of the full conditional distribution $z|\theta^{(j)}, \boldsymbol{y}$ for each value of θ in the set $\{\theta^{(j)}\}$

Univariate Example

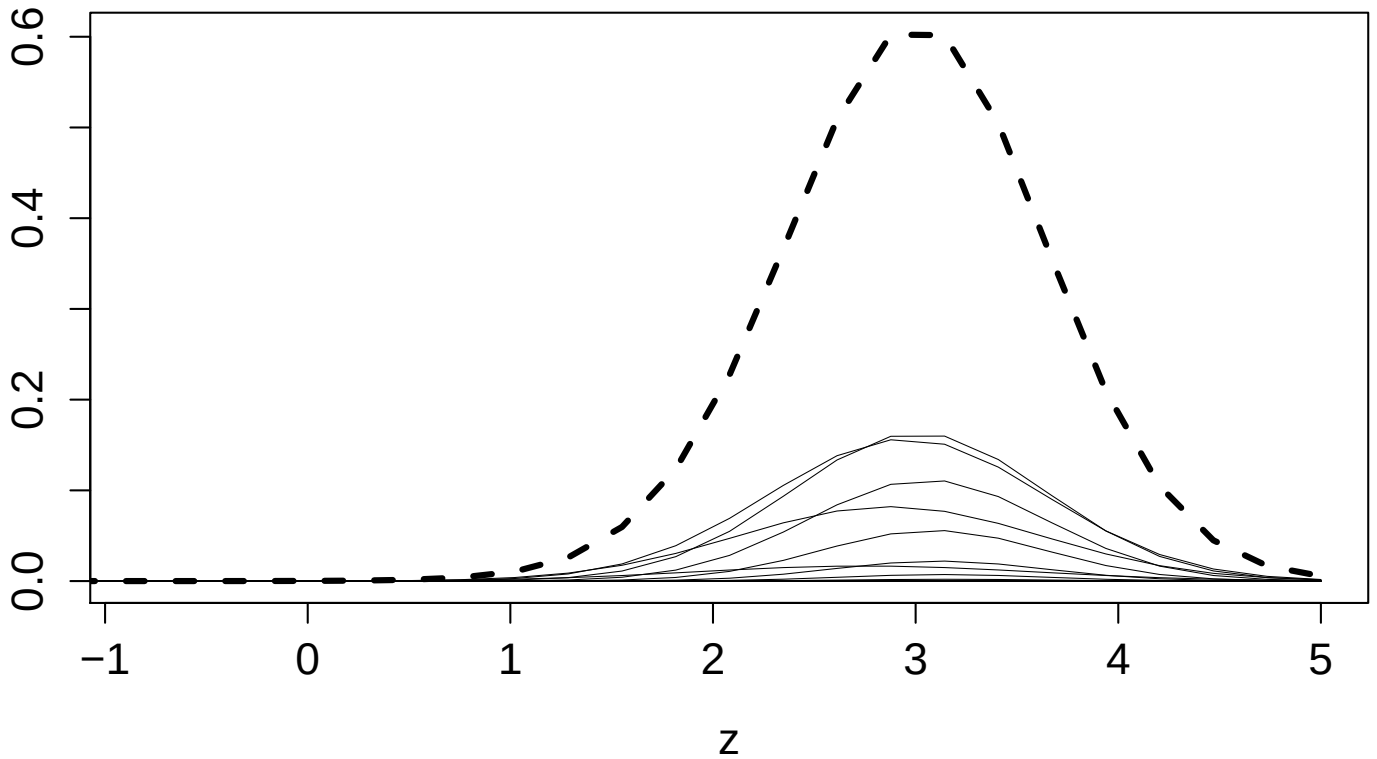


Figure: The dashed curve denotes the approximated posterior distribution of $z|\boldsymbol{y}$ computed as the finite sum of the weighted approximated joint posterior distributions - given by $p_G(z_i, |\boldsymbol{\theta}^{(j)}, \boldsymbol{y})\hat{p}(\boldsymbol{\theta}^{(j)})\Delta_j$ which are depicted as solid lines

Important Observation

If we observe

$$p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta}) \sim N(A\boldsymbol{z},\sigma_\epsilon^2 I)$$

then the approximation is exact and the problem reduces to a numerical linear algebra problem.

There are some caveats on this claim on the choice of hyperprior. If its Gaussian, or conjugate and proper then this is the case. If not, then some approximations will need to be made.

Future Work and Conclusion

What now?

So we have a way of performing Bayesian inference on a wide range of latent Gaussian models, and a suite of programs that will implement it. Does this mean we can all go home?

NO!

There are many numerical methods used in INLA that make it quite efficient. However, there are many things that need to be done to enable INLA to work efficiently on big data.

In particular, there is one problem that we cannot avoid...

What now?

We need to calculate determinants of very large matrices!

WHY WE NEED THE DETERMINANT

As discussed earlier, GMRFs are pivotal to the INLA approach to obtain efficiency.

The density function of the GMRF $\boldsymbol{z} \sim N(\boldsymbol{\mu}, Q^{-1})$ is

$$p(\boldsymbol{z}) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T Q(\boldsymbol{z}-\boldsymbol{\mu})\right)$$

The precision matrix Q is sparse so will have $\mathcal{O}(n)$ non-zero entries.

WHY WE NEED THE DETERMINANT

But what happens when I need to place a Gaussian process on 1.4 million points?

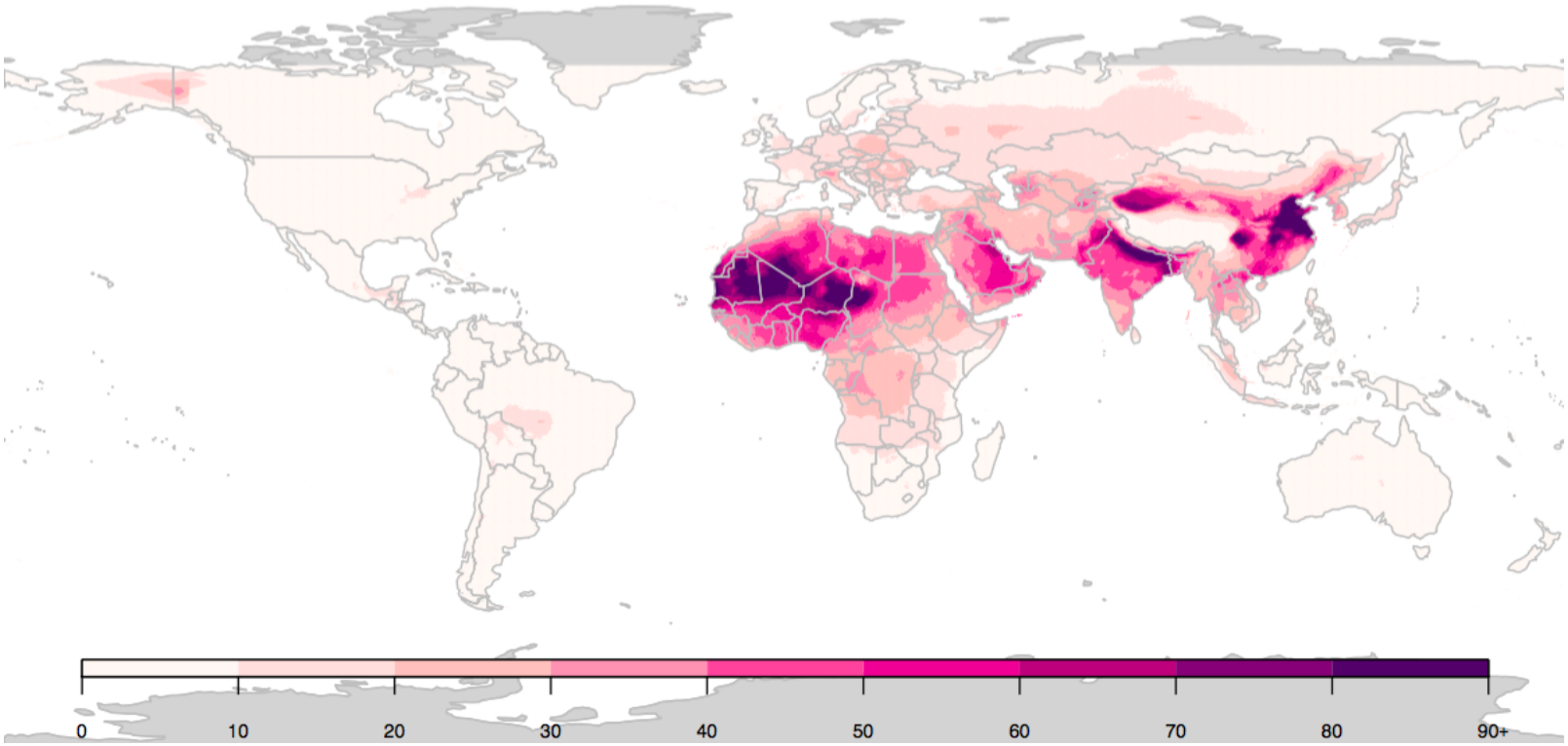


Figure: Satellite remote sensing estimates of $PM_{2.5}$ in μgm^{-3} for 2014 on a 10km$\times$10km grid used in GBD2015

WHY WE NEED THE DETERMINANT

There are a few large computations done with the precision matrix to enable inference,

- Solving linear systems (Can be done very efficiently)
- Evaluating the density at a given point (Requires the determinant of Q which is more difficult)

WHY WE NEED THE DETERMINANT

At the moment to enable these computations a Cholesky factorisation is used

$$Q = LL^T$$

Thankfully, the Markov property lends itself to allow for sparse Cholesky factor L.

This decomposition enables both a solve of a linear system

$$Qx = LL^T\boldsymbol{x} = \boldsymbol{z} \implies \boldsymbol{x} = L^{-T}L^{-1}\boldsymbol{z}$$

using back-substitution twice and calculation of the log-determinant

$$\log|Q| = 2\sum_{i=1}^{n}\log(L_{ii})$$

Can we do better than the Cholesky factorisation?

Conclusion

In general, INLA is a technique that

- allows inference from a large class of models in a wide range of applications.
- is more computationally attractive than say, MCMC.

However, the emergence of big data has meant that some models are computationally infeasible in the current implementation, so in the future, we have to be smarter about the numerics we use.

This is where I come in...!

REFERENCES

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319-392.

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.

Thank you for listening.

Any Questions?