Case Study: Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations

BIOSTAT830: Graphical Models

December 08, 2016

Introduction - INLA

- Inference for latent Gaussian Markov random field (GMRF) models, avoiding MCMC simulations
- Fast Bayesian inference using accurate, multiple types of approximations to
  - $pr(\boldsymbol{\theta} \mid \boldsymbol{y})$: marginal density for the model parameters
  - $pr(x_i \mid \boldsymbol{y})$: marginal posterior densities for one (or more) latent variables.
- Can be used for model criticisms:
  1. Fast cross-validation
  2. Bayes factors and deviation information criterion (DIC) can be efficiently calculated for model comparisoins
- Software *inla* available from R; very easy to use

Supported Models

- Hierarchical GMRF of the form:

$$\text{likelihood}: y_j \mid \eta_j, \boldsymbol{\theta}_1 \sim pr(y_j \mid \eta_j, \boldsymbol{\theta}_1), j \in J,$$
$$\text{linear predictor}: \eta_i = \text{Offset}_i + \sum_{k=0}^{n_f-1} w_{ki} f_k(c_{ki}) + \boldsymbol{z}'_i \beta + \epsilon_i,$$
$$i = 0, \ldots, n_\eta - 1.$$

- $J \subset \{0, 1, \ldots, n_\eta - 1\}$, i.e., not all latent $\boldsymbol{\eta}$ are observed through data $\boldsymbol{y}$
- $pr(y_j \mid \eta_j, \boldsymbol{\theta}_1)$: likelihood of data; known link function
- $\boldsymbol{\epsilon} = (\epsilon_0, \ldots, \epsilon_{n_\eta - 1})' \mid \lambda_\eta \sim \mathcal{N}(\mathbf{0}, \lambda_\eta \boldsymbol{I})$; λ_η denotes precision
- $\boldsymbol{\eta} = \{\eta_i\}$: a vector of linear predictors, e.g., in GLM
- $\boldsymbol{w}_k$: in the k-th nonlinear effect, the known weights, one for each observed data point

Supported Models (continued)

- Hierarchical GMRF of the form:

$$\text{likelihood}: y_j \mid \eta_j, \boldsymbol{\theta}_1 \sim pr(y_j \mid \eta_j, \boldsymbol{\theta}_1), j \in J,$$

$$\text{linear predictor}: \eta_i = \text{Offset}_i + \sum_{k=0}^{n_f - 1} w_{ki} f_k(c_{ki}) + \boldsymbol{z}_i' \beta + \epsilon_i,$$

$$i = 0, \ldots, n_\eta - 1.$$

- $f_k(c_{ki})$: nonlinear effect of covariate k for observation i
  1. Nonlinear effects: time trends and seasonal effects, two dimensional surfaces, iid random intercepts, slopes and spatial random effects.
  2. The unknown functions $\boldsymbol{f}_k = (f_{0k}, \ldots, f_{m_k - 1, k})'$ are modelled as GMRF given some parameter $\boldsymbol{\theta}_{f_k}$: $\boldsymbol{f}_k \mid \boldsymbol{\theta}_{f_k} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_k^{-1})$
- $\boldsymbol{z}_i$: A vector of n_β covariates assumed to have a linear effect
- β: The vector of unknown parameters

Supported Models (continued)

- $\boldsymbol{x} = (\boldsymbol{\eta}', \boldsymbol{f}'_0, \ldots, \boldsymbol{f}'_{n_f-1}, \boldsymbol{\beta}')'$: full vector of latent variables; Dimension: $n = n_\eta + \sum_{j=0}^{n_f-1} m_j + n_\beta$; note we parameterized $\boldsymbol{x}$ by $\boldsymbol{\eta}$ instead of $\boldsymbol{\epsilon}$
- All the elements of vector $\boldsymbol{x}$ are defined as GMRFs:

$$pr(\boldsymbol{x} \mid \boldsymbol{\theta}_2) = \prod_{i=0}^{n_\eta - 1} pr(\eta_i \mid \boldsymbol{f}_0, \ldots, \boldsymbol{f}_{n_f-1}, \boldsymbol{\beta}, \lambda_\eta^{-1}) \prod_{k=0}^{n_f-1} pr(\boldsymbol{f}_k \mid \boldsymbol{\theta}_{f_k}) \times \prod_{m=0}^{n_\beta - 1} pr(\beta_m),$$

where

$$\eta_i \mid \boldsymbol{f}_0, \ldots, \boldsymbol{f}_{n_f-1}, \boldsymbol{\beta} \sim \mathcal{N}\left(\sum_{k=0}^{n_f-1} f_k(c_{ki}) + \boldsymbol{z}'_i \boldsymbol{\beta}, \lambda_\eta\right), \beta_m \overset{iid}{\sim} \mathcal{N}(0, \lambda_\beta)$$

and $\boldsymbol{\theta}_2 = \{\log \lambda_\eta, \boldsymbol{\theta}_{f_0}, \ldots, \boldsymbol{\theta}_{f_{n_f-1}}\}$ is a vector of unknown hyperparameters.

Prior

- Specify priors on the hyperparameters:
  $\boldsymbol{\theta}_2 = \{\log \lambda_\eta, \boldsymbol{\theta}_{f_0}, \ldots, \boldsymbol{\theta}_{f_{n_f}-1}, \log \lambda_\beta\}$
- Need not be Gaussian

Examples

- *Time series model*: $c_k = t$ for time, f_k for nonlinear trends or seasonal effects

$$\eta_t = f_{trend}(t) + f_{seasonal}(t) + \boldsymbol{z}'_t \beta$$

- *Generalized additive models (GAM)*:
  1. $pr(y_i \mid \eta_i, \boldsymbol{\theta}_1)$ belongs to an exponential family
  2. c_{ki}'s are univariate, continuous covariates
  3. f_k's are smooth functions

Examples

- *Generalized additive mixed models (GAMM)* for longitudinal data
  - Individuals: $i = 0, \cdots, n_i - 1$, observed at time points $t_0, t_1, \ldots$. A GAMM extends a GAM by introducing individual specific random effects:

    $$\eta_{it} = f_0(c_{it0}) + \ldots + f_{n_f-1}(c_{it,n_f-1}) + b_{oi}w_{it0} + \ldots + b_{n_b-1,i}w_{it,n_b-1},$$

    where η_{it} is the linear predictor for individual i at time t, $c_{itk}, k = 0, \ldots, n_f - 1, w_{itq}, q = 0, \ldots, n_b - 1$ are covariate values for individual i at time t, and $b_{0i}, \ldots, b_{n_b-1,i}$ is a vector of n_b individual-specific random intercepts (if $w_{itq} = 1$) or slopes.
  - Just define $r = (i, t)$ and $c_{kr} = c_{kit}$ for $k = 0, \ldots, n_f - 1$ and $c_{n_f-1+q,r} = w_{qit}$, $f_{n_f-1+q}(c_{(n_f-1+q),r}) = b_{qi}w_{qit}$ for $q = 0, \ldots, n_b$.

Examples

- *Geoadditive models* (Kammann and Wand, 2003, JRSS-C):

$$\eta_i = f_1(c_{0i}) + \ldots + f_{n_f-1}(c_{n_f-1,i}) + f_{spatial}(s_i) + \boldsymbol{z}_i'\beta,$$

where s_i indicates the location of observation i and $f_{spatial}$ is a spatially correlated effect.

Examples

- *ANOVA-type interaction model*: For the effect of two continuous covariates w and v:

  $$\eta_i = f_1(w_i) + f_2(v_i) + f_{1,2}(w_i, v_i) + \dots,$$

  where f_1, f_2 are the main effects and $f_{1,2}$ is a two dimensional interaction surface. As a special case, we just define $c_{1i} = w_i$, $c_{2i} = v_i$ and $c_{3i} = (w_i, v_i)$,
- *Univariate stochastic volatility model*
  - Time series models with Gaussian likelihood where the variance (not the mean) of the observed data is part of the latent GMRF model:

    $$y_i \mid \eta_i \sim \mathcal{N}(0, \exp(\eta_i)),$$

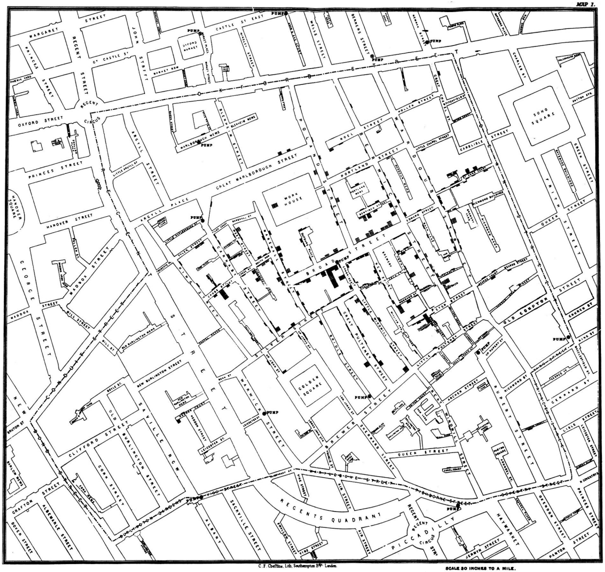
    and, for example, model the latent field $\boldsymbol{\eta}$ as an autoregressive model of order 1.

Bayesian for Spatial and Spatio-temporal Models (Blangiardo and Cameletti, 2015, Wiley)

INLA for Spatial Area Data: Suicides in London

- Disease mapping is commonly used in small area studies to assess the pattern of a disease
- To identify areas characterized by unusually high or low relative risk (Lawson 2009)

London Cholera Outbreak in 1854



- John Snow's Cholera map in dot style; dots represent deaths from cholera in London in 1854 to detect the source of the disease

Example: Suicide Mortality

- 32 Boroughs in London; 1989-1993
- For the i-th area, the number of suicides y_i:

$$y_i \sim Poisson(\lambda_i),$$

  where $\lambda_i = \rho_i E_i$, a product of rate ρ_i and the expected number of suicides E_i
- Linear predictor defined on logarithmic scale:

$$\eta_i = \log(\rho_i) = \alpha + v_i + \nu_i,$$

  where α is the intercept, $v_i = f_1(i)$ and $\nu_i = f_2(i)$ are two area specific effects.

Besag-York-Mollie (BYM) model (Besag et al. 1991)

- v_i: spatially structured residual, modeled using an intrinsic conditional autoregressive structure (iCAR):

$$v_i \mid v_{j \neq i} \sim Normal(m_i, s_i^2)$$
$$m_i = \frac{\sum_{j \in \mathcal{N}(i)} v_j}{|\mathcal{N}(i)|}$$
$$s_i^2 = \frac{\sigma_v^2}{|\mathcal{N}(i)|},$$

  where $|\mathcal{N}(i)|$ is the number of areas which share boundaries with the i-th one.
- ν_i: unstructured residual; modeled by exchangeable prior:

$$\nu_i \sim Normal(0, \sigma^2)$$

Priors

- $\log \tau_{\nu} \sim \log - Gamma(1, 0.0005)$
- $\log \tau_{v} \sim \log - Gamma(1, 0.0005)$

Goal

1. Posterior for borough-specific relative risks of suicides, compared to the whole of London: $pr(\exp(v_i + \nu_i) \mid \boldsymbol{y})$
2. Posterior exceedance probability: $pr(\exp(v_i + \nu_i) > 1 \mid \boldsymbol{y})$
3. Fraction of structured variance component

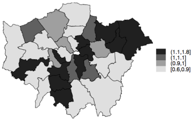
Incorporating Risk Factors

- Extension: when risk factors are available and the aim of the study is to evaluate their effect on the risk of death (or disease)
- Ecological regression model
- For example: Index of social deprivation (x_1), index of social fragmentation (describing lack of social connections and of sense of community) (x_2)
- Model:

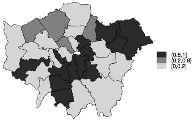
$$\eta_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + v_i + \nu_i$$

- Can be fitted using the R-INLA package

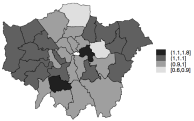
London Suicide Rates Mapping



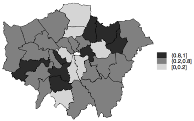
(a) Distribution of the borough specific relative risks of suicides $\zeta_i = \exp(\upsilon_i + \nu_i)$ in the disease mapping model



(b) Distribution of the borough specific posterior probability $p(\zeta_i > 1 \mid \boldsymbol{y})$ in the disease mapping model



(c) Distribution of the borough specific relative risks of suicides $\zeta_i = \exp(\upsilon_i + \nu_i)$ in the ecological regression model



(d) Distribution of the borough specific posterior probability $p(\zeta_i > 1 \mid \boldsymbol{y})$ in the ecological regression model

Figure 1: Borough specific relative risks and posterior probabilities.

Figure 1: suicide_rates

Other Spatial Examples (FiveThirtyEight)

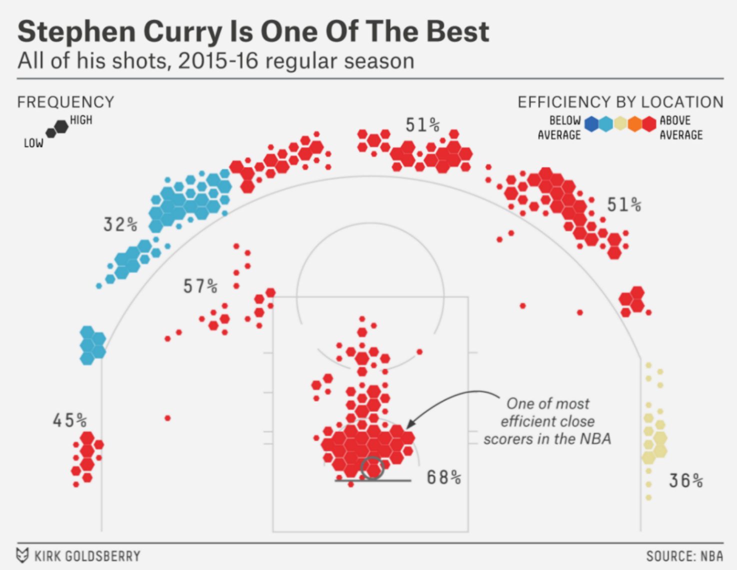


Figure 2: Stehpen Curry

Background: Gaussian Markov Random Fields (GMRF)

- GMRF: $\boldsymbol{x} = (x_1, \ldots, x_n)'$ with Markov property that for some $i \neq j$, $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$
- Can be encoded by precision matrix $\boldsymbol{Q}$: $Q_{ij} = 0$ if and only if $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$
- Density function with mean vector $\boldsymbol{\mu}$:

$$pr(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

- Most cases: $\boldsymbol{Q}$ is sparse: only $\mathcal{O}(n)$ of the n^2 entries are nonzero
- Can handle extra linear constraints: $\boldsymbol{Ax} = \boldsymbol{e}$ for a $k \times n$ matrix $\boldsymbol{A}$ of rank k
- *Computational note*: Simulation usually based on lower Cholesky decomposition $\boldsymbol{Q} = \boldsymbol{LL}'$, with $\boldsymbol{L}$ preserving the sparseness in $\boldsymbol{Q}$. *See Section 2.1 in Rue et al. (2009) for more details.*

Background: Gaussian approximation (under regularity conditions)

- Find a Gaussian density $q(\boldsymbol{z})$ to approximate a density $p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$, where $Z = \int f(\boldsymbol{z}) d\boldsymbol{z}$
  - One-dimensional case
  - Multi-dimensional case
- Need to find mode $\boldsymbol{z}_0$ (Newton or quasi-Newton methods)
- Need not know the normalizing constant Z
- Central Limit Theorem, approximate becomes better as sample size n increases if $f(\boldsymbol{z}; \boldsymbol{Data})$ is a posterior distribution of model parameters
- Typically better for marginal and conditional posteriors than joint posteriors (marginals are averages across other distributions!)
- Can use transformations (e.g., logit or log) to approximate a distribution over a constrained space
- Not so useful if there is skewness, or if interested in extreme values that are far from the mode

Background: Gaussian approximation - Density at Extreme Values (Bishop CM, 2006, Sec 4.4)

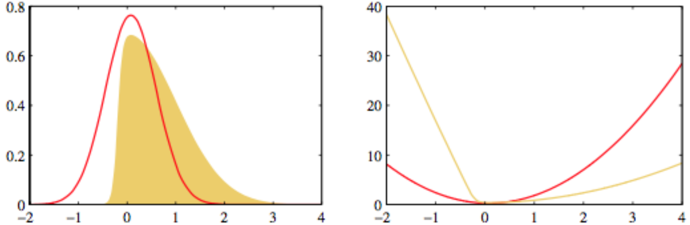


Figure 4.14 Illustration of the Laplace approximation applied to the distribution $p(z) \propto \exp(-z^2/2)\sigma(20z+4)$ where $\sigma(z)$ is the logistic sigmoid function defined by $\sigma(z) = (1+e^{-z})^{-1}$. The left plot shows the normalized distribution $p(z)$ in yellow, together with the Laplace approximation centred on the mode z_0 of $p(z)$ in red. The right plot shows the negative logarithms of the corresponding curves.

Background: Gaussian Approximations

- Approximate density of the form

$$pr(\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in\mathcal{I}} g_i(x_i)\right\},$$

  where $g_i(x_i) = \log(pr(y_i \mid x_i, \boldsymbol{\theta}))$ in our setting.
- Gaussian approximation $\tilde{pr}_G(\boldsymbol{x})$: obtained by matching the modal configuration and curvature at the mode (model could be computed by Newton-Raphson method)
- Let the mode be $\boldsymbol{x}^*$, the precision matrix be $\boldsymbol{Q}^* + diag(\boldsymbol{c}^*)$ (hint: use expansion to the second order - $g_i(x_i) \approx g_i(\boldsymbol{x}^*) + b_i x_i - \frac{1}{2}c_i x_i^2$)
- *Property*: because the second summation does not involve x_i and x_j in one $g()$, the resulting $\boldsymbol{Q}^*$ preserves the Markov property in the original latent Gaussian model on $\boldsymbol{x}$

Background: Laplace Approximation

- Approximate marginal posterior:

$$pr(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\int pr(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{x}}{\int pr(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{\theta}} \propto \left. \frac{pr(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\tilde{pr}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})},$$

where $\boldsymbol{x}^*(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{x}} pr(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$.

- Key difference with Tierney and Kadane (1986) JASA: here in latent Gaussian models, the dimension of latent field $\boldsymbol{x}$ is n, could change with the number of observations n_d; Not the case in TK1986

Goal of INLA: Approximate Marginal Posteriors

- Marginal posterior for each θ_k and x_j by numerical integration over $\boldsymbol{\theta}$:
-
$$pr(\theta_k \mid \boldsymbol{y}) \approx \int \tilde{pr}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-k}$$
-
$$pr(x_j \mid \boldsymbol{y}) \approx \int \tilde{pr}(x_j \mid \boldsymbol{\theta}, \boldsymbol{y}) \tilde{pr}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}$$

INLA in Three Steps

- **Goal**: Compute posterior marginal $pr(x_i \mid \boldsymbol{y})$, $i = 1, \ldots, n$.
- **Step I**: Laplace approximation to $pr(\boldsymbol{\theta} \mid \boldsymbol{y})$; Will be used to integrate out uncertainty about $\boldsymbol{\theta}$
- **Step II**: Simplified Laplace approximation to $pr(\boldsymbol{x}_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ over selected $\boldsymbol{\theta}$ values: $\{\boldsymbol{\theta}_k\}$
- **Step III**: Combines the previous two steps using numerical integration

INLA - Step I: Approximate $pr(\boldsymbol{\theta} \mid \boldsymbol{y})$

- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$

1. Locate the mode $\boldsymbol{\theta}^*$ for $\tilde{pr}(\boldsymbol{\theta} \mid \boldsymbol{y})$: optimize $\log(\tilde{pr}(\boldsymbol{\theta} \mid \boldsymbol{y}))$ by quasi-Newton method; Compute the Hession matrix $\boldsymbol{H}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$
2. Construct a representation for general $\boldsymbol{\theta}$ values for exploration: $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{z}) = \boldsymbol{\theta}^* + \boldsymbol{V\Lambda}^{1/2}\boldsymbol{z}$, where $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$ and $\boldsymbol{\Sigma}$ has been spectrally decomposed as $\boldsymbol{\Sigma} = \boldsymbol{V\Lambda V}'$
3. Explore $\log(\tilde{pr}(\boldsymbol{\theta} \mid \boldsymbol{y}))$ over a grid of $\{\boldsymbol{\theta}_k\}$ by using the $\boldsymbol{z}$-parametrization. Need stepsize δ_z in each $\boldsymbol{z}$-direction. For each grid points, assign weight Δ_k (see next slide for an example with $m = 2$)
4. Can approximate $pr(\theta_j \mid \boldsymbol{y})$ already!

INLA - Step I-3

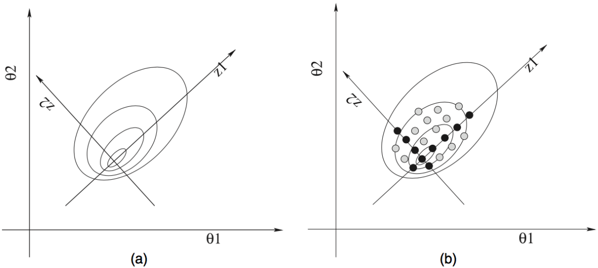


Fig. 1. Illustration of the exploration of the posterior marginal for $\boldsymbol{\theta}$: in (a) the mode is located and the Hessian and the co-ordinate system for **z** are computed; in (b) each co-ordinate direction is explored (●) until the log-density drops below a certain limit; finally the new points (◦) are explored

INLA - Step II: Approximate $pr(x_i \mid \boldsymbol{\theta}_k, \boldsymbol{y})$

- Now we have a set of weighted points $\{\boldsymbol{\theta}_k\}$, we obtain for each x_i the marginal posterior given each selected $\boldsymbol{\theta}_k$ Three options:

1. Gaussian approximation: simplest and cheapest: $\tilde{pr}_G(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$; There could be errors in the location or due to the lack of skewness
2. Laplace approximation

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

   Too expensive: recomputed $\tilde{pr}_{GG}()$ at every x_i. Has some fixes (see Section 3.2.3 of Rue et al. 2009)
3. **Simplified Laplace approximation**: Correct Gaussian approximation for location and skewness AND has computing time $\mathcal{O}(n^2 \log n)\exp(m)$.

Comparing MCMC and INLA

- **MCMC**: Stochastic simulation of the posterior; Accurate if computing time is not a concern (rarely true)
- Easy posterior inferene for functions of unknowns
- Components of latent field $\boldsymbol{x}$ strongly dependent; $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are also strongly dependent. Chains will mix painfully slow
- Usually requires blockwise proposal-and-rejection scheme (aka block MCMC)
- The Monte Carlo error decays at rate $\mathcal{O}(N^{-1/2})$.
- Time: hours to days for some spatial models (see Rue et al, 2009)

Comparing MCMC and INLA

- **INLA**: Deterministic; Using analytic approximations
- Suitable for latent GMRF; Sparse precision matrix can speed up computations; Approximation bias found to be smaller than typical MCMC in some cases
- *Variational Bayes*: Also deterministic approximation; Iterative algorithm; Usually require exponential-family likelihood and priors on $\boldsymbol{\theta}$
- Time: seconds or minutes

INLA - Summary

- Compute the posterior marginals for latent Gaussian Markov Random Field Models based on deterministic Laplace approximations
- Much faster than MCMC with small approximation biases
- Practically exact results by INLA over a randge of commonly used latent Gaussian models; Also has tools for assessing approximation errors to decide if they are non-neglegible (not discussed see Section 4 of Rue et al. 2009)
- Could be a basis for greater automation and parallel implementation; Core is the sparse matrix algorithms; Essentially no tunning.
- Disadvantage: computing time exponential of m, the dimension of hyperparameters $\boldsymbol{\theta}$
- Could be used as a baseline model to explore smooth effects

Extensions (Not Discussed)

- Approximate posterior marginals for a subset of $\boldsymbol{x}_S$
- Approximate marginal likelihood (e.g. for Bayes factor)
- Approximate predictive measures for model crticism and comparison
- Approximate Deviance Information Criteria (Spiegelhalter, 2002, Bayesian measure of model complexity and fit)

Comment

- Next and Final lecture: Network Analysis
- **Required reading**:
  - Rue, Martino and Chopin (2009) Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations. JRSS-B, 71(2): 319-392.
- Additional References:
  - INLA Tutorials
  - Simpson et al. (2015). Going off grid: computationally efficient inference for log-Gaussian Cox processes. Biometrika.
- Other resouces:
  - R-INLA project
  - All models implemented in R inla package