Beyond MCMC in fitting complex Bayesian models: The INLA method

Valeska Andreozzi

Centre of Statistics and Applications of Lisbon University
(valeska.andreozzi at fc.ul.pt)

European Congress of Epidemiology
Porto, Portugal
September 2012
This talk will be focused on an alternative method to perform Bayesian inference of complex models, namely the integrated nested Laplace approximation (INLA).

This alternative technique will be illustrated by an application of a generalized linear/additive mixed model in the analysis of a cohort study.

An R interface for INLA program (r-INLA package) will be demonstrated.
Outline

• Describe statistical modelling
• Give an illustrative example (Cohort study)
• Describe the Bayesian inference
• Summarize the MCMC methods
• Present a briefly introduction to INLA
• Illustrate the software available to implement INLA (r-INLA package)
• Provide some references to more details about INLA
• Concluding remarks
Complex models in Health

- By nature, Health is a complex system.
- Applied statisticians are more and more facing model complexity to represent Health problems satisfactory/appropriately.
Statistical Modelling

Set a hypothesis

⇓

Collect data \((y)\)

⇓

Trend + error

⇓

Statistical model for the data

⇓

Joint probability distribution for the data \(→ f(y|θ)\)
Statistical Modelling

- Actually, the joint probability distribution $f(y|\theta)$ for the data depend on the values of the unknown parameters $\theta$.

- In practice, when $f(y|\theta)$ is viewed as a function of $\theta$ rather than $y$, $f(y|\theta)$ is known as the **likelihood** for the data.

- The likelihood under the proposed model represent *how likely it is to observe the data $y$ we have collected, given specific values for the parameters $\theta$*
Illustrative example

Let’s introduce some example to illustrate the methods that will be discussed:

- Investigate the rate of infant weight gain.
  - Observational cohort study: infants attending the Public Health System of the city of Rio de Janeiro.
  - Project: *Quality assessment of care for infants under six months of age at the Public Health System of the city of Rio de Janeiro*
  - Coordinator: Doctor Maria do Carmo Leal (ENSP/Fiocruz/Brazil)
Infant weight gain in Rio de Janeiro

Data

- Data consist of repeated measures of infant weight collected from the child health handbook.

Covariates available:

- Infant characteristics: age, birth weight, gender, type of breastfeeding, whether attending nursery school, whether has been hospitalized.

- Mother characteristics: age, level of education, marital status, number of pregnancies, number of people in the household, number of child less than 5 years in the household, Kotelchuck Index
Model formulation

- The simplest model (likelihood - $f(y|\theta)$) for the data $y_i$ is the normal distribution
- To take into account the longitudinal structured of the data (within correlation among repeated measures of weight), random effects have been included.
- The result model for the infant weight gain can be expressed as

$$y_i = X_i\beta + Z_i\omega_i + \epsilon_i$$
Model formulation (cont.)

- Considering a **hierarchical** model formulation

\[
y_i \sim N(\mu_i, \Sigma^{-1})
\]

\[
\mu_i = X_i \beta + Z_i \omega_i
\]

\[
\omega_i \sim N(0, D^{-1})
\]

- Using the previous notation, the data \( y \) is assumed to have a **normal likelihood** \( f(y|\theta) \) and the vector of unknown parameters consist of \( \theta = (\beta, \Sigma, D) \)
Model inference

Maximum likelihood estimates (MLE)

- To find good estimates for $\theta$, one can choose values $\hat{\theta}$ that maximize the likelihood $f(y|\theta) \Rightarrow$ MLE
- Maximum likelihood methods is fine, but if the form of the likelihood $f(y|\theta)$ is complex and/or the number of individual parameters involved is large then the approach may prove either very difficult or infeasible to implement.
Bayes inference

- In this scenario, Bayesian inference is a appealing approach.
- Start by reviewing the Bayes theorem
  The posterior probability distribution for the parameters given the observed data is:

  \[ P(\theta | y) = \frac{P(\theta) f(y | \theta)}{f(y)} = \frac{P(\theta) f(y | \theta)}{\int_\theta P(\theta) f(y | \theta) d\theta} \]

- This results show that:
Bayes inference

\[ \text{Posterior} \propto \text{Prior} \times \text{Likelihood} \]

- The prior probability distribution for the parameters \( P(\theta) \) express the uncertainty about \( \theta \) before taking the data into account.
- Usually it is chosen to be non-informative.
- The posterior probability distribution for the parameters \( P(\theta|y) \) express the uncertainty about \( \theta \) after observed the data.
- So in Bayesian inference all parameter information comes from the posterior distribution.
- For example:

\[
\hat{\theta} = E(\theta|y) = \int_{\theta} \theta P(\theta|y)d\theta = \frac{\int_{\theta} \theta f(y|\theta)P(\theta)d\theta}{\int_{\theta} f(y|\theta)P(\theta)d\theta}
\]
Computing the posterior distribution

• Until recently, model solutions were very hard to obtain because integrals of the posterior distribution cannot be evaluated analytically.

• Solution: draw samples values from the posterior distribution approximating any characteristics of it by the characteristics of these samples.

• But how?
MCMC

Using Monte Carlo integration methods with Markov Chain (MCMC)

- This algorithm constructs a Markov chain with stationary distribution identical to the posterior and uses values from the Markov chain after a sufficiently long burn-in as simulated samples from the posterior.

- An attractive method to implement an MCMC algorithm is the Gibbs sampling.
Gibbs sampling

- Suppose the set of conditional distributions:

\[
\begin{align*}
[\theta_1 | \theta_2, \ldots, \theta_p, \text{data}] \\
[\theta_2 | \theta_1, \ldots, \theta_p, \text{data}] \\
\vdots \\
[\theta_p | \theta_1, \ldots, \theta_{p-1}, \text{data}]
\end{align*}
\]

- The idea behind Gibbs sampling is that we can set up a Markov chain simulation algorithm from the joint posterior distribution by successfully simulating individual parameters from the set of \( p \) conditional distributions.

- Under general conditions, draws from the simulating algorithm will converge to the target distribution of interest (the joint posterior distribution of \( \theta \) ).
MCMC estimation

- Although very flexible to program complex models.
- One has to monitor the performance of a MCMC algorithm to decide, at a long run (?), if the simulated sample provides a reasonable approximation to the posterior density.
- Issues to ensure good estimates:
  - Convergence (burn-in required).
  - Mixing (required number of samples after convergence).
MCMC estimation

- Difficult to construct a MCMC scheme that converges in a reasonable amount of time (can take hours or even days to deliver correct results).
- Difficulty in specifying prior distributions.
- In practice, the handicap of data analysis using MCMC is the large computational burden.

How to overcome this efficiency problem?
Alternative method

Use alternative approximation methods to MCMC, namely, the Integrated Nested Laplace Approximation.

• Designed to a class of hierarchical models called latent Gaussian models

• Examples where INLA can be applied:
  ○ Generalized linear mixed models
  ○ Generalized additive mixed models
  ○ Spline smoothing models
  ○ Disease mapping (including ecological studies)
  ○ Spatial and spatio-temporal models
  ○ Dynamic generalized linear models
  ○ Survival models
Latent Gaussian models

- It is a hierarchical model defined in 3 stages:

  **Observation model:** \( y_i | \theta \sim f(y_i | \theta) \)

  **Latent Gaussian field or Parameter model:** \( \theta | \gamma \sim N(\mu(\gamma), Q(\gamma)^{-1}) \)

  **Hyperparameter:** \( \gamma \sim f(\gamma) \)
Latent Gaussian models

- The observation $y_i$ belongs to the exponential family where the mean $\mu_i$ is linked to a structured additive predictor $\eta_i$ through a link function $g(\mu_i) = \eta_i$

$$\eta_i = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k x_{ki} + \sum_{j=1}^{n_h} h^{(j)}(z_{ji}) + \epsilon_i$$

- $\beta_k$ represent the linear effect of covariates $x$
- $h^{(j)}$ are unknown functions of covariates $z$. Can represent non-linear effects, time-trend, seasonal effects, random effects, spatial random effects.
- $\epsilon_i$ are the unstructured random effects.
Latent Gaussian models

Observation model: \( y_i | \theta \sim f(y_i | \theta) \)

Latent Gaussian field or Parameter model: \( \theta | \gamma \sim N(\mu(\gamma), Q(\gamma)^{-1}) \)

Hyperparameter: \( \gamma \sim f(\gamma) \)

The models are assumed to satisfy two properties:

1. The latent Gaussian field \( \theta \), usually of large dimension, is conditionally independent, so its precision matrix \( Q(\gamma) \) is sparse (full of zeros because many parameters are not correlated)
2. the dimension of the hyperparameter vector \( \gamma \) is small (\( \leq 6 \))
• With gaussian parameters and sparse precision matrix, it can be assumed a multivariated gaussian distribution for the ordinary scenario (normal response variable)

• In the Gaussian layout, inference is an easy task. It means that the posterior distribution of parameters is easy to calculate.

• For more complex models (non Gaussian response variable, for example), approximation method of integration of the posterior density has to be used ⇒ Laplace.
The marginal posterior density for $\theta$ is given by

$$P(\theta_i|y) = \int P(\theta_i|\gamma, y)P(\gamma|y)d\gamma$$

and INLA approximate this by

$$\tilde{P}(\theta_i|y) = \sum_k \tilde{P}(\theta_i|\gamma_k, y)\tilde{P}(\gamma_k|y)\Delta_k$$

where Laplace is applied to carry out the integrations required for evaluation of $\tilde{P}$.

So no simulations are needed to find estimate for $\theta$.

The INLA output is the marginal posterior distribution which can be summarized by means, variances and quantiles.

DIC and predictive measures (CPO, PIT) are also available.
## Results

\[ \text{weight}_i = \eta_i = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k x_{ki} + \sum_{j=1}^{n_h} h(j)(z_{ji}) + \epsilon_i \]

<table>
<thead>
<tr>
<th>Structured additive predictor</th>
<th>Processing total time (s)</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + h^{(1)}(\mathbf{u}_i) + \epsilon_i )</td>
<td>9.70</td>
<td>4647.21</td>
</tr>
<tr>
<td>( \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + h^{(1)}(\mathbf{u}_i) + \epsilon_i + \text{cov} )</td>
<td>15.62</td>
<td>4631.66</td>
</tr>
<tr>
<td>( \beta_0 + h^{(1)}(\mathbf{u}_i) + h^{(2)}(\text{age}_i) + \epsilon_i )</td>
<td>15.46</td>
<td>4641.46</td>
</tr>
<tr>
<td>( \beta_0 + h^{(1)}(\mathbf{u}_i) + h^{(2)}(\text{age}_i) + \epsilon_i + \text{cov} )</td>
<td>19.52</td>
<td>4627.81</td>
</tr>
</tbody>
</table>

\( h^{(1)}(\mathbf{u}_i) \sim N(0, \tau_u^{-1}) \) and \( \tau_1 \sim \text{Gamma}(a, b) \)

\( h^{(2)}(\text{age}_i) \) is a random walk smoothness prior with precision \( \tau_2 \)
## Results

### Model hyperparameters:

<table>
<thead>
<tr>
<th>Precision</th>
<th>mean</th>
<th>sd</th>
<th>0.025quant</th>
<th>0.975quant</th>
</tr>
</thead>
<tbody>
<tr>
<td>within child ($\epsilon$)</td>
<td>5.54</td>
<td>0.16</td>
<td>5.23</td>
<td>5.87</td>
</tr>
<tr>
<td>intercept ($u$)</td>
<td>3.14</td>
<td>0.16</td>
<td>2.83</td>
<td>3.47</td>
</tr>
<tr>
<td>within child ($\epsilon$)</td>
<td>5.55</td>
<td>0.16</td>
<td>5.24</td>
<td>5.87</td>
</tr>
<tr>
<td>intercept ($u$)</td>
<td>3.55</td>
<td>0.19</td>
<td>3.19</td>
<td>3.95</td>
</tr>
<tr>
<td>age</td>
<td>3710.58</td>
<td>2207.72</td>
<td>949.28</td>
<td>9313.81</td>
</tr>
<tr>
<td>within child ($\epsilon$)</td>
<td>5.56</td>
<td>0.16</td>
<td>5.25</td>
<td>5.89</td>
</tr>
<tr>
<td>intercept ($u$)</td>
<td>3.14</td>
<td>0.16</td>
<td>2.82</td>
<td>3.47</td>
</tr>
<tr>
<td>age</td>
<td>3756.31</td>
<td>2187.72</td>
<td>909.55</td>
<td>9182.95</td>
</tr>
</tbody>
</table>
Results

- Posterior density estimation of hyperparameters.
Results

- Posterior estimated of the non-linear effect of age by the model adjusted by mother and infant covariates
Results

- Posterior density estimation of covariate effects
INLA in R

```r
> library(INLA)
> model<-inla(y ~ x1+x2+
    f(age,model="rw2") +
    f(id,model="iid"),
    data=dataframe,family="gaussian",
    control.compute=list(dic=TRUE,cpo=T))

> summary(model)

> plot(model)
```
The R-INLA project

Bayesian computing with INLA!

This site provides documentation to the R-INLA package which solves a large class of statistical models using the INLA approach.

Here is a short introduction describing the class of models which can be solved using R-INLA.

All models implemented in R-INLA are described in details, moreover a large series of worked out examples are provided and we hope that this will help the user to gain familiarity with the library. Recent changes in the code can be viewed here.

Recent posts to the discussion group

Google Group

Welcome to this discussion group about r-inla. Please ask your questions here in case you think they will be useful for others, otherwise, send them to help@inla.org. You are of course free to comment on questions from others as well.

Best,
H

Show the appropriate size of the mesh (2)

- By Darek Kinyoko - 2 posts - 16 views - updated 12:14 PM (11 hours ago)

- by Max - 2 posts - 8 views - updated 4:56 AM (18 hours ago)

- can INLA implement leverage effect model now? (2)

- By zts21512 - 2 posts - 26 views - updated Aug 27 (7 days ago)

- can INLA useful new argument when doing prediction (4)

- By Harald Rue - 4 posts - 82 views - updated Aug 26 (8 days ago)

Recent announcements

Bayesian Biostatistics short course: WinBUGS/SAS/INLA, MUSC May 2013
Date: May 13, 2013
Time: 10:00 AM - 2:00 PM
Place: Room 201, Medical Education Building, MUSC

Bayesian disease mapping course, University of Edinburgh, April 2012

Bayesian computing with INLA: short course in Athens

Daniel Simpson will give a one-day short course in Athens on Bayesian computing with INLA, January 13, 2012. The announcement is found here.
Concluding remarks

Search: Integrated Nested Laplace Approximation(s)

Explore it!
Concluding remarks

• The main aim of this presentation was twofold:
  ○ to highlight the fact that MCMC sampling is not a simple tool to be used in routine analysis due to convergence and computational time issues.
  ○ to show that applied research can benefit, specially computationally, if an alternative method, such as INLA is adopted for Bayesian inference.
Concluding remarks

• The lessons are:
  - At the end of the day, there´s no free lunch. INLA does not substitute MCMC. It is designed for a specific class of models, whilst very flexible one.
  - It comes to complement and make Health problems statistical analysis (Bayesian speaking) more straightforward.
  - And beyond any doubt:
    
    *If you want r-INLA to have a particular feature, observation model or prior model, you need to ask us!*

    *Simpson D, Lindgren F, Rue H*
Selected references


Fong Y., Rue H. and Wakefield J. 2010, Bayesian inference for Generalized Linear Mixed Models. Biostatistics, 11, 397-412. (R-code and supplementary material)


A complete list can be found at www.r-inla.org/papers
This research has been partially supported by National Funds through FCT — Fundação para Ciência e Tecnologia, projects PTDC/MAT/118335/2010 and PEst-OE/MAT/UI0006/2011

Thank you very much for your attention!