Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations

H. Rue, S. Martino and N. Chopin
Journal of the Royal Statistical Society, Series B

Presented by Esther Salazar
Duke University

April 16, 2012

Aim of the paper

- They consider approximate Bayesian inference for *additive regression models*, where the latent field/component is Gaussian
- They show that, by using an integrated nested Laplace approximation (INLA), we can directly compute very accurate approximations to the posterior marginals
- The methodology is particularly attractive if the latent Gaussian model is a GMRF
- **Main benefit**: computational time. Where MCMC algorithms need hours or days to run, the INLA approximations provide more precise estimates in seconds or minutes

Class of models

They consider a subclass of *structured additive regression models*, named latent Gaussian models:

Structured additive regression models

- Linear predictor: $\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(\mu_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i$
- Observations: $\boldsymbol{y} \sim \pi(\boldsymbol{y}|\boldsymbol{\eta}) = \prod_i \pi(y_i|\eta_i)$

Class of models

They consider a subclass of *structured additive regression models*, named latent Gaussian models:

Structured additive regression models

- Linear predictor: $\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(\mu_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i$
- Observations: $\boldsymbol{y} \sim \pi(\boldsymbol{y}|\boldsymbol{\eta}) = \prod_i \pi(y_i|\eta_i)$

Latent Gaussian models

If we assign Gaussian priors on α, $\{f^{(j)}(\cdot)\}$, $\{\beta_k\}$ and $\{\epsilon_i\}$, let $\boldsymbol{x}$ denote the vector of all the latent Gaussian variables and $\boldsymbol{\theta}$ the vector of hyperparameters we will have the three-stage Bayesian hierarchical model

$$\begin{aligned}
\text{Hyperprior:} \quad \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \\
\text{Parameter model:} \quad \boldsymbol{x}|\boldsymbol{\theta} &\sim \pi(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \\
\text{Observation model:} \quad \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} &\sim \prod_i \pi(y_i|\eta_i, \boldsymbol{\theta})
\end{aligned}$$

Latent Gaussian models: notation and basic properties

- Observed data: $y_i|x_i \sim \pi(y_i|x_i, \boldsymbol{\theta})$
- Latent Gaussian field: $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$
- Hyperparameters: $\boldsymbol{\theta}$
- Posterior distribution:

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

Features:

- y_i is often non-Gaussian (Poisson, binomial, etc)
- Dimension of the latent Gaussian field: n large between 10^2 - 10^5
- Dimension of $\boldsymbol{\theta}$: $\dim(\boldsymbol{\theta})$ is small, between $1 - 5$

Main goal: compute marginal posterior distribution

From

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

compute the posterior marginals

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

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_{-j},$$

The key feature of the approach is to use this form to construct nested approximations

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

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_{-j}.$$

What is the main idea?

The approach os based on the the identity

$$\pi(z) = \frac{\pi(x,z)}{\pi(x|z)} \quad \text{leading to} \quad \tilde{\pi}(z) = \frac{\pi(x,z)}{\tilde{\pi}(x|z)}$$

where $\tilde{\pi}(x|z)$ is the Gaussian approximation (Tierney and Kadane's 1986 Laplace approximation)

INLA approximates

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

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}_{-j},$$

by

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

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\theta_k, \boldsymbol{y})\tilde{\pi}(\theta_k|\boldsymbol{y})\Delta_k$$

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

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

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}, \quad \forall \boldsymbol{x}$$

- INLA approximation:

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

  where $\tilde{\pi}_G$ is the Gaussian approximation to $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode

Steps

(1) locate the mode of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ by optimizing $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ with respect to $\boldsymbol{\theta}$ (using e.g. quasi-Newton method)

(2) at the modal configuration $\boldsymbol{\theta}^*$ compute the negative Hessian matrix $\boldsymbol{H} > 0$. Let $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ and use the standardized variable $\boldsymbol{z}$ instead of $\boldsymbol{\theta}$ and compute $\boldsymbol{\theta}(\boldsymbol{z}) = \boldsymbol{\theta}^* + \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}$

(3) explore $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ by using the $\boldsymbol{z}$-parameterization

(4) posterior marginals $\pi(\theta_j|\boldsymbol{y})$ can be obtained directly from $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

we can start from the mode $\boldsymbol{z} = 0$ and go in the positive direction of z_1 with step length δ_z say $\delta_z = 1$ as long as

$$\log[\tilde{\pi}\{\boldsymbol{\theta}(0)|\boldsymbol{y}\}] - \log[\tilde{\pi}\{\boldsymbol{\theta}(\boldsymbol{z})|\boldsymbol{y}\}] < \delta_z$$

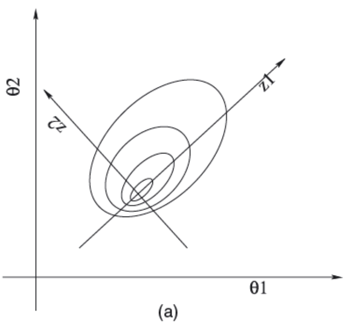

Fig. 1. Illustration of the exploration of the posterior marginal for θ: in (a) the mode is located and the Hessian and the co-ordinate system for **z** are computed; in (b) each co-ordinate direction is explored (•) until the log-density drops below a certain limit; finally the new points (◦) are explored

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

Recall that

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\theta_k, \boldsymbol{y})\tilde{\pi}(\theta_k|\boldsymbol{y})\Delta_k$$

with a set of weighted points $\{\theta_k\}$ to be used in the previous integration.

Three alternatives for approximation $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

- Gaussian approximation (Rue and Martino, 2007), easily extractable from $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ where
$$\tilde{\pi}_G(x_i|\boldsymbol{\theta}, \boldsymbol{y}) = N(x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta}))$$
- Laplace approximations
$$\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) = N(x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})) \ \exp\{\text{cubic spline}(x_i)\}$$
- Simplified Laplace approximation based on the skew-normal distribution (Azzalini and Capitano, 1999)

  The simplified Laplace approximation appears to be highly accurate for many observational models.

Approximation methods in machine learning

- Variational Bayes (VB): The principle of VB is to use as an approximation the joint density $q(\boldsymbol{x}, \boldsymbol{\theta})$ that minimizes the Kullback-Leibler contrast of $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ wrt $q(\boldsymbol{x}, \boldsymbol{\theta})$

  However, even though VB seem often to approximate well the posterior mode, the posterior variance can be (sometimes) underestimated.

- Expectation propagation (EP): (Minka, 2001). For latent Gaussian models can be demonstrated that EP usually overestimates the posterior variance (Bishop, 2006, chapter 10)

Disease mapping with covariate

Example: Larynx cancer mortality counts are observed in the 544 district of Germany from 1986 to 1990. The data are conditionally independently Poisson counts

$$y_i|\eta_i \sim \text{Poisson}(E_i \exp(\eta_i)), \quad , i = 1, \ldots, 544$$

where E_i is fixed and accounts for demographic variation, and η_i is the log relative risk. Together with the counts, for each district, the level of smoking consumption c_i is registered.

The model for η_i is

$$\eta_i = \mu + f_s(s_i) + \beta c_i + u_i$$

where $f_s(s_i)$ is the spatial effect and u_i is the unstructured random effect.
The model has three hyperparameters $\boldsymbol{\theta} = (\log \lambda_s, \log \lambda_f, \log \lambda_\eta)$ (unknown precisions)

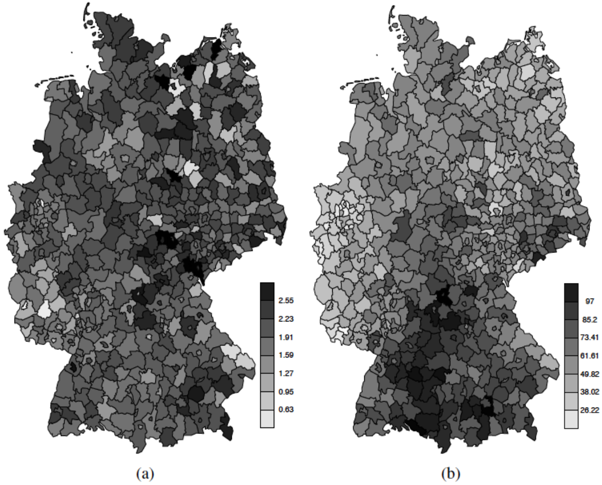

Figure 8: Standardised mortality ratio for larynx cancer, panel (a) and observed covariate values, panel(b)

Implementing using the INLA package for R

```
require(rgl)
require(INLA)
require(lattice)

# Disease mapping with covariate
data(Germany)
Germany<-cbind(Germany,region.struct=Germany$region)


# Model (INLA approximation)
formula<-Y~f(region.struct,model="besag",graph.file="germany.graph",
param=c(1,0.00005),initial=2.8)+x+f(region,model="iid")

mod<-inla(formula, family="poisson", data=Germany, E=E,
control.inla=list(h=0.01), verbose=TRUE)

# Plots
source("draw-map.r")
res = matrix(mod$mode$x[1:1632],544,3)
germany.map(res[,2])

plot(mod)
```