INLA: Integrated Nested Laplace Approximations

John Paige

Statistics Department
University of Washington

October 10, 2017

The problem

- Markov Chain Monte Carlo (MCMC) takes too long in many settings. How can we get a good approximation fast?

Integrated Nested Laplace Approximation (INLA)

- One possible solution is INLA
- The idea: for certain types of models, we can break the posterior integration up into a nested product of low-dimensional integrals. We can then approximately numerically integrate these with high accuracy
- Laplace approximation

Latent Gaussian models

- Assume exponential family observations: y_i
- Mean: μ_i
- Link function: $g(\cdot)$
- Predictor: $\eta_i = g(\mu_i)$ with,

$$\eta_i = \alpha + \sum_{j=1}^{n_f} \underbrace{f^{(j)}(u_{ji})}_{\text{(smooth fxns)}} + \sum_{k=1}^{n_\beta} \underbrace{\beta_k z_{ki}}_{\text{(linear fxns)}} + \varepsilon_i$$

- Give Gaussian priors to these parameters
- Applications: spatial models, GMRFs, regression models, dynamic models
- Set $f^{(j)}(u_\mathbf{s}) = f_\mathbf{s}$ for spatial application

INLA Overview

- Let $\mathbf{x}$ be all n Gaussian variables (parameters) $\{\eta_i\}$, α, $\left\{f^{(j)}\right\}$, $\{\beta_k\}$, and $\{\varepsilon_i\}$
- Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ be the hyperparameters
- $\pi(\mathbf{x}|\boldsymbol{\theta}_1)$: Gaussian with zero mean and precision $\mathbf{Q}(\boldsymbol{\theta}_1)$
- $\pi(\mathbf{x}|\boldsymbol{\theta})$: Gaussian with zero mean and precision $\mathbf{Q}(\boldsymbol{\theta})$

INLA Overview

Then:

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta})$$
$$\propto \pi(\boldsymbol{\theta}) \left|\mathbf{Q}(\boldsymbol{\theta})\right|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_{i \in \mathcal{I}} \log\left\{\pi(y_i|x_i, \boldsymbol{\theta})\right\}\right\}$$

- Goal:
  approximate posterior marginals $\pi(x_i|\mathbf{y})$, $\pi(\boldsymbol{\theta}|\mathbf{y})$, $\pi(\theta_j|\mathbf{y})$
- If $\mathbf{Q}(\boldsymbol{\theta})$ is sparse (i.e. under conditional independence)

INLA Overview

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

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

approximated with:

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

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

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

- $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$: Gaussian (Laplace) approximation
- $\mathbf{x}^*(\boldsymbol{\theta})$: mode of $\pi(\mathbf{x}|\boldsymbol{\theta})$

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

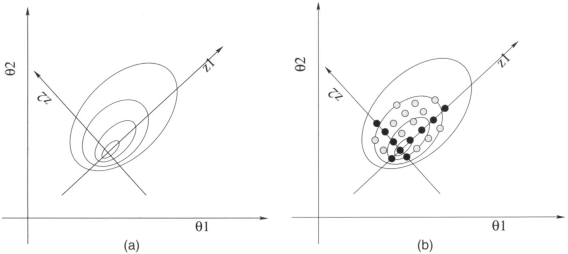

Fig. 1. Illustration of the exploration of the posterior marginal for $\boldsymbol{\theta}$: in (a) the mode is located and the Hessian and the co-ordinate system for **z** are computed; in (b) each co-ordinate direction is explored (•) until the log-density drops below a certain limit; finally the new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

Step 1: Find mode of $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ using quasi-Newton method, $\boldsymbol{\theta}^*$

Step 2: Compute the negative Hessian, $\mathbf{H}$, at $\boldsymbol{\theta}^*$, and set $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$

Step 3: Explore $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ near $\boldsymbol{\theta}^*$ along principal axes of $\mathbf{H}$ on grid

Now we can construct interpolant and numerically evaluate $\log \pi(\boldsymbol{\theta}|\mathbf{y})$ (since $\boldsymbol{\theta}$ low-dim.)

Approximating $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$: Gaussian approximation

- Cheapest approximation
- Use Gaussian Fisher-scoring. Already need to do this for evaluating:

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

Approximating $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$: Laplace approximations

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

$$\mathbf{x}^*_{-i}(x_i, \boldsymbol{\theta}) \approx E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)$$

$$\Rightarrow \frac{E_{\tilde{\pi}_G}(x_j|x_i) - \mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta}) \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})}$$

$$\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \mathcal{N}\left(x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\right) \exp\left\{\text{cubic spline}(x_i)\right\}$$

- $\tilde{\pi}_{GG}$ is the Gaussian approximation
- $\mu_i(\boldsymbol{\theta})$, $\sigma_i^2(\boldsymbol{\theta})$: marginal mean, var of $\tilde{\pi}_G(x_i|\boldsymbol{\theta}, \mathbf{y})$
- Gauss-Hermite quadrature $\rightarrow \tilde{\pi}_{LA}$ normalizing constant

Approximating densities: Accuracy

- Accuracy depends on 'effective number of parameters' of $\mathbf{x}$:

  $$p_D(\boldsymbol{\theta}) \approx n - \operatorname{tr}\left\{\mathbf{Q}(\boldsymbol{\theta})\mathbf{Q}^*(\boldsymbol{\theta})^{-1}\right\}$$

  Low $p_D(\boldsymbol{\theta}) \Rightarrow$ high accuracy
- For GMRF, asymptotic error rate is $\mathcal{O}(q/n_d)$ for n_d number observations, q rank of $\mathbf{x}$ Gaussian distribution
- In most cases, approximation errors cancel out reducing error rates from $\mathcal{O}(n_d^{-1})$ to $\mathcal{O}(n_d^{-3/2})$
- These results apply to both approximations of $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{x})$ and of $\pi(\boldsymbol{\theta}|\mathbf{y})$

Approximating densities: Assessing errors

Idea 1:

$$\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})} \propto E_{\tilde{\pi}_G}\left[\exp\left\{\sum_i h_i(x_i)\right\}\right]$$

Where h_i is $\log\{\pi(y_i|x_i, \boldsymbol{\theta})\}$ minus the second order term of its Taylor expansion around $x_i^*(\boldsymbol{\theta})$

Idea 2: Compute symmetric Kullback Leibler divergence (SKLD) for Gaussian, Laplace, and simplified Laplace approximations

Examples: AR(1) model with unknown mean

0.08s vs. 25s of computation time:

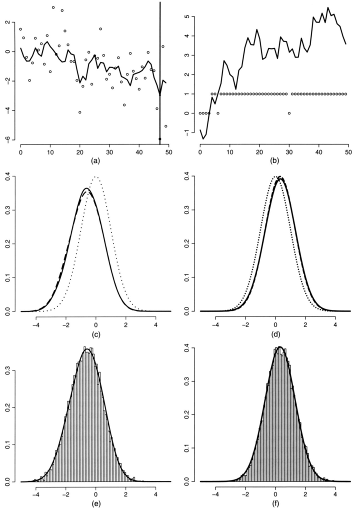

Fig. 2. (a), (b) True latent Gaussian field (———), observed Student t_3-data and Bernoulli data (∘), (c), (d) approximate marginal for a selected node by using various approximations (·······, Gaussian; – – –, simplified Laplace; ———, Laplace) and (e), (f) comparison of samples from a long MCMC chain with the marginal computed with the simplified Laplace approximation

Examples: A generalized linear mixed model for longitudinal data

1.5s vs. "hours" of computation time:

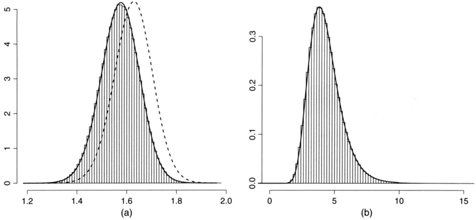

Fig. 3. Posterior marginal for (a) β_0 (———, simplified Laplace approximation; - - - - - - -, Gaussian approximation) and (b) τ_ε (———, after integrating out τ_ν) for the example in Section 5.2: the histograms result from a long MCMC run using OpenBUGS

Examples: Stochastic volatility model

11s vs. "long":

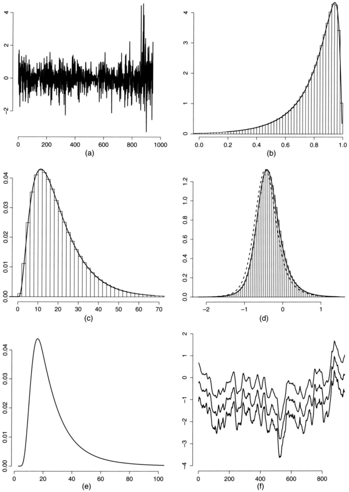

Fig. 4. (a) Log-daily-difference of the pound–dollar exchange rate from October 1st, 1981, to June 28th, 1985, (b), (c) approximated posterior marginals for ϕ and τ by using only the first $n=50$ observations in (a) (overlaid are the histograms that were obtained from a long MCMC run using OpenBUGS), (d) approximated posterior marginal by using simplified Laplace approximations (———) and Gaussian approximations (-------) for μ, which is the node in the latent field with maximum SKLD, (e) posterior marginal for the degrees of freedom assuming Student t_ν-distributed observations and (f) 0.025, 0.5 and 0.975 posterior quantiles for η_t

Examples: Mapping cancer incidence rates

34s vs. "long":

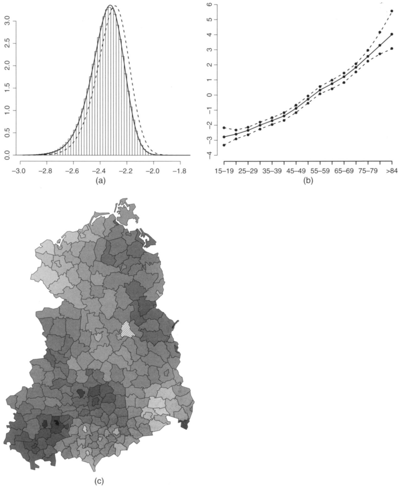

Fig. 5. Results for the cancer incidence example: (a) posterior marginal for $f_3^{(a)}$ by using simplified Laplace approximations (———), Gaussian approximations (-------) and samples from a long MCMC run (□); (b) posterior median (———) and 0.025- and 0.975-quantiles (-------) of the age–class effect and results obtained from a long MCMC run (•); (c) posterior median of the (smooth) spatial effect

Examples: Log-Gaussian Cox process

10hrs (simplified Laplace) vs. 10 min (Gaussian) vs. "long" (MCMC):

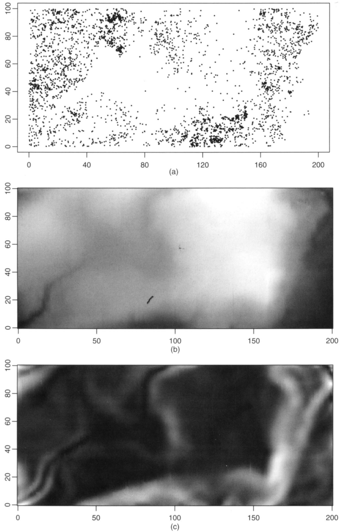

Fig. 6. Data and covariates from the log-Gaussian Cox process example: (a) locations of the 3605 trees, (b) altitude and (c) norm of the gradient

Questions?