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: $\mu_i$
- Link function: $g(\cdot)$
- Predictor: $\eta_i = g(\mu_i)$ with,

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f(j)(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i$$

(smooth fxns)  (linear fxns)

- Give Gaussian priors to these parameters
- Applications: spatial models, GMRFs, regression models, dynamic models
- Set $f(j)(u_s) = f_s$ for spatial application
Let $x$ be all $n$ Gaussian variables (parameters) $\{\eta_i\}, \alpha, \{f(j)\}, \{\beta_k\}$, and $\{\varepsilon_i\}$.

Let $\theta = (\theta_1, \theta_2)^T$ be the hyperparameters.

$\pi(x|\theta_1)$: Gaussian with zero mean and precision $Q(\theta_1)$.

$\pi(x|\theta)$: Gaussian with zero mean and precision $Q(\theta)$. 


INLA Overview
Then:

\[
\pi(x, \theta|y) \propto \pi(\theta) \pi(x|\theta) \prod_{i \in I} \pi(y_i|x_i, \theta)
\]

\[
\propto \pi(\theta) |Q(\theta)|^{1/2} \exp \left\{ -\frac{1}{2} x^T Q(\theta) x + \sum_{i \in I} \log \{ \pi(y_i|x_i, \theta) \} \right\}
\]

▶ Goal:
   approximate posterior marginals \( \pi(x_i|y), \pi(\theta|y), \pi(\theta_j|y) \)

▶ If \( Q(\theta) \) is sparse (i.e. under conditional independence)
\[ \pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y) \, d\theta \]

\[ \pi(\theta_j|y) = \int \pi(\theta|y) \, d\theta_{-j} \]

approximated with:

\[ \tilde{\pi}(x_i|y) = \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y) \, d\theta \]

\[ \tilde{\pi}(\theta_j|y) = \int \tilde{\pi}(\theta|y) \, d\theta_{-j} \]

\[ \tilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x|\theta, y)} \bigg|_{x=x^*(\theta)} \]

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

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

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

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

Now we can construct interpolant and numerically evaluate $\log \pi(\theta|y)$ (since $\theta$ low-dim.)
Approximating $\pi(x_i|\theta, y)$: Gaussian approximation

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

$$\tilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x|\theta, y)}\bigg|_{x=x^*(\theta)}$$
Approximating $\pi(x_i|\theta, y)$: Laplace approximations

\[
\tilde{\pi}_{LA}(x_i|\theta, y) \propto \frac{\pi(x, \theta, y)}{\tilde{\pi}_{GG}(x_{-i}|x_i, \theta, y)} \bigg|_{x_{-i}=x_{-i}^*(x_i, \theta)}
\]

\[
x_{-i}^*(x_i, \theta) \approx E_{\tilde{\pi}_G}(x_{-i}|x_i)
\]

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

\[
\tilde{\pi}_{LA}(x_i|\theta, y) \propto \mathcal{N}(x_i; \mu_i(\theta), \sigma_i^2(\theta)) \exp\{\text{cubic spline}(x_i)\}
\]

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

- Accuracy depends on ‘effective number of parameters’ of $x$:

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

Low $p_D(\theta) \Rightarrow$ high accuracy

- For GMRF, asymptotic error rate is $O(q/n_d)$ for $n_d$ number observations, $q$ rank of $x$ Gaussian distribution

- In most cases, approximation errors cancel out reducing error rates from $O(n_d^{-1})$ to $O(n_d^{-3/2})$

- These results apply to both approximations of $\pi(x|\theta, x)$ and of $\pi(\theta|y)$
Approximating densities: Assessing errors

Idea 1:

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

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

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:
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 $\omega$ and $\nu$ 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 $\omega$, which is the node in the latent field with maximum SKLD, (e) posterior marginal for the degrees of freedom assuming Student $t_{\nu}$-distributed observations and (f) 0.025, 0.5 and 0.975 posterior quantiles for $\nu$. 
Examples: Mapping cancer incidence rates

34s vs. “long”:

Fig. 5. Results for the cancer incidence example: (a) posterior marginal for \( r^{(a)} \) 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?