Simplifying Integrated nested Laplace approximation with adaptive Gaussian Hermite quadrature

Adam Howes[a]

[a]*Department of Mathematics Imperial College London*

Abstract

Background:

Methods:

Results:

Conclusions:

1. Introduction

The `TMB` package [6] is gaining popularity in spatial statistics as a flexible alternative to `R-INLA` for fitting latent Gaussian models [7].

2. Background

2.1. Integrated nested Laplace approximation

Integrated nested Laplace approximation (INLA) [9] is an approximate Bayesian inference method which uses the Laplace approximation and numerical integration. INLA is designed for use with latent Gaussian models (LGMs) of the form

$$\text{(Observations)} \qquad y_i \sim p(y_i \,|\, x_i, \boldsymbol{\theta}), \quad i = 1, \ldots, n, \tag{1}$$

$$\text{(Latent field)} \qquad \mathbf{x} \sim \mathcal{N}(\mathbf{x} \,|\, \mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1}), \tag{2}$$

$$\text{(Parameters)} \qquad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \tag{3}$$

where $\dim(\mathbf{y}) = \dim(\mathbf{x}) = n$ and $\dim(\boldsymbol{\theta}) = m$, and $m < n$. The joint posterior of $(\mathbf{x}, \boldsymbol{\theta})$ is given by

$$p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{n/2} \exp\left(-\frac{1}{2} \mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i=1}^{n} \log p(y_i \,|\, x_i, \boldsymbol{\theta}) \right). \tag{4}$$

Rather than approximating the above full posterior, the INLA method instead approximates the posterior marginals of each latent random variable x_i and parameter θ_j given by

$$p(x_i \,|\, \mathbf{y}) = \int p(x_i, \boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta} = \int p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}, \quad i = 1, \ldots, n, \tag{5}$$

$$p(\theta_j \,|\, \mathbf{y}) = \int p(\boldsymbol{\theta} \,|\, \mathbf{y}) \mathrm{d}\boldsymbol{\theta}_{-j} \quad j = 1, \ldots, m. \tag{6}$$

An approximation is made to each of the two quantities, $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ and $p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y})$, nested inside the above integrals: (i) $p(\boldsymbol{\theta} \,|\, \mathbf{y}) \approx \tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y})$ and (ii) $p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) \approx \tilde{p}(x_i \,|\, \boldsymbol{\theta}, \mathbf{y})$, which we discuss in turn.

*Corresponding author
Email address: `ath19@ic.ac.uk` (Adam Howes)

2.1.1. Approximation (i)

The posterior marginal of the parameters $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ appears in both Equations (5) and (6). This distribution is approximated by $\tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y})$ and represented by a set of K integration points $\{\boldsymbol{\theta}^{(k)}\}$ and area-weights $\{\Delta^{(k)}\}$. The first step is to rewrite $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ as

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y})}{p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})} \propto \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})}. \tag{7}$$

Approximation (i) then uses a Gaussian approximation

$$p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \approx p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \triangleq \mathcal{N}(\mathbf{x} \,|\, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \hat{\mathbf{Q}}(\boldsymbol{\theta})^{-1}) \tag{8}$$

to the denominator of Equation 7. This approximation is accurate as the Gaussian prior on the latent field $\mathbf{x}$ makes the posterior distribution

$$p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_{i=1}^{n} \log p(y_i \,|\, x_i, \boldsymbol{\theta})\right) \tag{9}$$

close to being Gaussian since $\mathbf{y}$ is generally not that informative and the observation distribution $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta})$ is usually well-behaved [2].

As $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ does not depend on $\mathbf{x}$, any value may be chosen to evaluate the right hand side of Equation 7. As such, taking $\mathbf{x} = \hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$, the value where the Gaussian approximation is most accurate, gives the final approximation as

$$\tilde{p}(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto \frac{p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})} = \frac{p(\mathbf{y}, \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta})}{\det(\hat{\boldsymbol{Q}}(\boldsymbol{\theta}))^{1/2}}, \tag{10}$$

where the final equality is because $p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})$ is evaluated at its mode $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$.

2.1.2. Approximation (ii)

Having utilised used the approximation $p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \approx p_G(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y})$ in Section~2.1.1, a natural approach, and that taken by Rue and Martino [8], is to marginalise this distribution directly

$$\tilde{p}(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(x_i \,|\, \hat{\mu}_i(\boldsymbol{\theta}), 1/\hat{q}_i(\boldsymbol{\theta})), \tag{11}$$

where the marginal mean $\hat{\mu}_i(\boldsymbol{\theta})$ and precision $\hat{q}_i(\boldsymbol{\theta})$ are recovered directly from $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ and $\mathbf{Q}^\star(\boldsymbol{\theta})$ respectively. Although this approximation is fast, it tends not to be accurate, as it involves evaluating the Gaussian approximation away from its mode. As a result, although this method is available in `R-INLA` it is generally not advised. Instead, Rue et al. [9] propose two methods, a Laplace approximation and a simplified version which is less computationally demanding. The full Laplace approximation is

$$\begin{aligned}
p(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}) &= p(\mathbf{x} \,|\, \boldsymbol{\theta}, \mathbf{y}) \times \frac{1}{p(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})} && (12)\\
&= \frac{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y})}{p(\boldsymbol{\theta} \,|\, \mathbf{y})} \times \frac{1}{p(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})} && (13)\\
&\propto \frac{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y})}{p(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})} && (14)\\
&\approx \frac{p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y})}{p_G(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}_{-i}=\hat{\boldsymbol{\mu}}_{-i}(x_i,\boldsymbol{\theta})} = \tilde{p}_{LA}(x_i \,|\, \boldsymbol{\theta}, \mathbf{y}), && (15)
\end{aligned}$$

where $p_G(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation to $\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y}$ and $\hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta})$ is its modal configuration.[1] The set of distributions $\{p(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})\}_{i=1}^n$ are usually reasonably Gaussian so this approximation tends to work well. However, the Gaussian approximation $p_G(\mathbf{x}_{-i} \,|\, x_i, \boldsymbol{\theta}, \mathbf{y})$ must be recomputed for each value of x_i and $\boldsymbol{\theta}$, which is often computationally prohibitive. Two modifications to Equation (15) are proposed by Rue et al. [9] to reduce the computational cost:

[1] This notation is somewhat dangerous as $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ is the mode of the Gaussian approximation to the full latent field given $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta})$ is not the same as $\hat{\boldsymbol{\mu}}_{-i}(\boldsymbol{\theta})$.

1. Avoiding having to find the mode via optimisation by using the approximation $\hat{\boldsymbol{\mu}}_{-i}(x_i, \boldsymbol{\theta}) \approx \mathbb{E}_{p_G(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}(\mathbf{x}_{-i} \mid x_i)$
2. As only those x_j close to x_i should have an impact on the marginal of x_i, then by selecting some subset $R_i(\boldsymbol{\theta})$ of nodes j to impact j the matrix which needs to be factorised can be reduced in dimension to be $|R_i(\boldsymbol{\theta})| \times |R_i(\boldsymbol{\theta})|$ rather than $n \times n$

2.1.3. Combining the approximations

2.2. Simplified INLA

3. Template Model Builder

Template Model Builder (TMB) [6] is an R package for fitting random effect models, also known as latent variable models, hierarchical models or a host of other names. In TMB inference is based upon optimisation of a target function. This makes it very flexible, and able to handle non-linear, non-Gaussian random effect models.

The approach of TMB is inspired by the AD Model Builder (ADMB) package [5]. The "AD" in ADMB is automatic differentiation, a technique for calculating derivates of functions by repeated application of the chain rule. AD is popular in machine learning [1], for example as the basis for backpropagation algorithm and is beginning to gain popularity in statistics, including as a part of Stan [3]. TMB uses the derivatives from AD for multiple purposes including calculation of the Hessian used in Gaussian approximations and for numerical optimisation routines.

3.1. Statistical framework

Consider unobserved latent random effects $\mathbf{x} \in \mathbb{R}^n$ and parameters $\boldsymbol{\theta} \in \mathbb{R}^m$.[2] Let $\ell(\mathbf{x}, \boldsymbol{\theta}) \triangleq -\log p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$ be the negative joint log-likelihood. In TMB, the user writes C++ code to evaluate this negative log-likelihood function ℓ. A standard maximum likelihood approach is to optimise

$$L_\ell(\boldsymbol{\theta}) \triangleq \int_{\mathbb{R}^n} p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^n} \exp(-\ell(\mathbf{x}, \boldsymbol{\theta})) \mathrm{d}\mathbf{x} \tag{16}$$

with respect to $\boldsymbol{\theta}$ to find the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$. Taking a superficially more Bayesian approach than above, instead of ℓ, the user may instead write a function to evaluate the negative joint penalised log-likelihood given by

$$f(\mathbf{x}, \boldsymbol{\theta}) \triangleq -\log p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta}) = \ell(\mathbf{x}, \boldsymbol{\theta}) - \log p(\mathbf{x}, \boldsymbol{\theta}), \tag{17}$$

equivalent up to an additive constant to the negative log-posterior. Using f in place of ℓ, then the penalised likelihood is proportional to the posterior marginal of $\boldsymbol{\theta}$

$$L_f(\boldsymbol{\theta}) \triangleq \int_{\mathbb{R}^n} \exp(-f(\mathbf{x}, \boldsymbol{\theta})) \mathrm{d}\mathbf{x} \propto \int_{\mathbb{R}^n} p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \mathrm{d}\mathbf{x} = p(\boldsymbol{\theta} \mid \mathbf{y}). \tag{18}$$

Integrating out the random effects directly, as in Equation~18 above, is usually intractable because $\mathbf{x}$ is high-dimensional, so Kristensen et al. [6, Equation 3] use a Laplace approximation $L_f^\star(\boldsymbol{\theta})$ based instead upon integrating out a Gaussian approximation to the random effects. This Laplace approximation is analogous to the INLA approximation $\tilde{p}(\boldsymbol{\theta} \mid \mathbf{y})$ given in Section~2.1.1.

$$f''_{\mathbf{xx}}(\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \mathbf{x}^2} \log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \Big|_{\mathbf{x}=\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})} = -\frac{\partial^2}{\partial \mathbf{x}^2} \log p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}=\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})} = \hat{\boldsymbol{Q}}(\boldsymbol{\theta}).$$

Inference proceeds by optimising $L_f^\star(\boldsymbol{\theta})$ via minimisation of

$$-\log L_f^\star(\boldsymbol{\theta}) \propto \frac{1}{2} \log \det(\hat{\boldsymbol{Q}}(\boldsymbol{\theta})) + f(\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta}), \tag{19}$$

where $\propto$ is used to mean proportional up to an additive constant. The parameters of the Gaussian approximation (Equation~8), are found in terms of f via $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \arg\min_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\theta})$ and $\hat{\boldsymbol{Q}}(\boldsymbol{\theta}) = f''_{\mathbf{xx}}(\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ and must be recomputed for each value of $\boldsymbol{\theta}$. Obtaining $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ is known as the inner optimisation step.

[2] Kristensen et al. [6] use the notation u for random effects and θ for parameters. We aim for consistency with Section 2.1.

4. Examples

4.1. The Naomi small-area estimation model

Eaton et al. [4] specify a joint model linking small-area estimation models of HIV prevalence from household surveys, HIV prevalence from antenatal care clinics, and antiretroviral therapy (ART) coverage from routine health data collection. Modelling data from multiple sources concurrently increases statistical power, and may mitigate the biases of any single source giving a more complete picture of the situation as well as prompting investigation into any conflicts.

Prevalence component Consider a country partitioned into areas $i = 1, \ldots, n$, with a simple random household survey of m_i people is conducted in each area with y_i HIV positive cases observed. Cases may be modelled using a binomial logistic regression model

$$y_i \sim \text{Bin}(m_i, \rho_i), \tag{20}$$

$$\text{logit}(\rho_i) \sim \mathcal{N}(\beta_\rho, \sigma_\rho^2) \tag{21}$$

where HIV prevalence ρ_i is modelled by a Gaussian with mean β_ρ and standard deviation σ_ρ.

ANC component Routinely collected data from pregnant women attending antenatal care clinics (ANCs) is another important source of information about the HIV epidemic. If of m_i^{ANC} women, y_i^{ANC} are HIV positive, then an analogous binomial logistic regression model

$$y_i^{\text{ANC}} \sim \text{Bin}(m_i^{\text{ANC}}, \rho_i^{\text{ANC}}), \tag{22}$$

$$\text{logit}(\rho_i^{\text{ANC}}) = \text{logit}(\rho_i) + b_i, \tag{23}$$

$$b_i \sim \mathcal{N}(\beta_b, \sigma_b^2), \tag{24}$$

may be used to describe HIV prevalence amongst the sub-population of women attending ANCs. Reflecting the fact that prevalence in ANCs is related but importantly different to prevalence in the general population, bias terms b_i are used to offset ANC prevalence from HIV prevalence.

ART component The number of people receiving treatment at district health facilities A_i also provides additional information about HIV prevalence, whereby districts with high prevalence are likely to have a greater number of people receiving treatment. ART coverage, defined to be the proportion of PLHIV currently on ART on district i, is given by $\alpha_i = A_i/\rho_i N_i$, where N_i is the total population of district i and assumed to be fixed. As such, ART coverage may also be modelled using a binomial logistic regression model

$$A_i \sim \text{Bin}(N_i, \rho_i \alpha_i), \tag{25}$$

$$\text{logit}(\alpha_i) \sim \mathcal{N}(\beta_\alpha, \sigma_\alpha^2). \tag{26}$$

4.2. Supporting information

Appendix A: Statistical modelling
Appendix B: Supplementary tables and figures

4.3. Funding

AH was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1).

4.4. Disclaimer

References

[1] Atılım Günes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18 (1):5595–5637, 2017.

[2] Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

[3] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[4] Jeffrey W Eaton, Sumali Bajaj, Andreas Jahn, Thokozani Kalua, Andrew Mganga, Andrew F Auld, Evelyn Kim, Danielle Payne, Ray W Shiraishi, Steve Gutreuter, Timothy B Hallett, and Leigh F Johnson. Joint small-area estimation of HIV prevalence, ART coverage and HIV incidence. *Working paper*, 2019.

[5] David A Fournier, Hans J Skaug, Johnoel Ancheta, James Ianelli, Arni Magnusson, Mark N Maunder, Anders Nielsen, and John Sibert. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2): 233–249, 2012.

[6] Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, Bradley M Bell, et al. Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(i05), 2016.

[7] Aaron Osgood-Zimmerman and Jon Wakefield. A statistical introduction to template model builder: A flexible tool for spatial modeling, 2021.

[8] Håvard Rue and Sara Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192, 2007.

[9] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.