INLA – Bayesian inference without MCMC

Finn Lindgren
finn.lindgren@ed.ac.uk

The University of Edinburgh

21 November 2019

Latent Gaussian models

Hierarchical model with latent jointly Gaussian variables

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \quad \text{(covariance parameters)}$$
$$(\boldsymbol{u} \mid \boldsymbol{\theta}) \sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}) \quad \text{(latent Gaussian variables)}$$
$$(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta}) \sim p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta}) \quad \text{(observation model)}$$

We are interested in the posterior densities $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, $p(\boldsymbol{u} \mid \boldsymbol{y})$ and $p(u_i \mid \boldsymbol{y})$.

Approximate conditional posterior distribution

Let $\widetilde{\boldsymbol{u}}(\boldsymbol{\theta})$ be the mode of the posterior density $p(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{u} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})$. Construct an approximate conditional posterior distribution:

$$p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) \sim \mathsf{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1})$$
$$\boldsymbol{0} = \nabla_{\boldsymbol{u}} \left\{ \ln p(\boldsymbol{u} \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta}) \right\}\big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}(\boldsymbol{\theta})}$$
$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_u - \nabla_{\boldsymbol{u}}^2 \ln p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})\big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}}$$

Use Newton optimisation to find the mode.

Integrated Nested laplace Approximation (INLA)

Exact posterior densities

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{\theta})p(\boldsymbol{u} \mid \boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{u}=\widetilde{\boldsymbol{u}}}$$

$$p(\boldsymbol{u}_i \mid \boldsymbol{y}) = \int p(\boldsymbol{u}_i \mid \boldsymbol{y}, \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$$

for arbitrary $\widetilde{\boldsymbol{u}}$.

The INLA approximation (`r-inla.org` & `inlabru.org`)

$$\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{\theta})p(\boldsymbol{u} \mid \boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}(\boldsymbol{\theta})}$$

$$\widetilde{p}(u_i \mid \boldsymbol{y}) = \int p_{GG}(u_i \mid \boldsymbol{y}, \boldsymbol{\theta})\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$$

The conditional posterior mode $\widetilde{\boldsymbol{\mu}}(\boldsymbol{\theta})$ tends to provide good approximations.

Integrated Nested Laplace Approximation (INLA) basics

1. Estimate the posterior mode for $p(\boldsymbol{\theta}|\boldsymbol{y})$ by optimisation of the approximation

$$\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{\theta})p(\boldsymbol{u} \mid \boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}(\boldsymbol{\theta})}$$

   where $p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})$ is a Gaussian approximation matching the low order derivatives at the mode of the exact conditional log-posterior for $\boldsymbol{u}$. (In a fully Gaussian model this is exact.)
   This is a Laplace approximation of $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
2. Find the mode of $\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$.
3. Construct a numerical integration grid/scheme $(\boldsymbol{\theta}_k, w_k)$ for $\boldsymbol{\theta}$, where w_k are integration weights; this step also estimates the normalisation constant.
4. Construct $p_{GG}(u_i \mid \boldsymbol{y}, \boldsymbol{\theta}_k)$ as Laplace approximations of the marginal conditional posterior densities, integrating out $\boldsymbol{u}_{-i} = \{u_j, j \neq i\}$.
5. Combine to form marginal posterior densities:

$$\widetilde{p}(u_i \mid \boldsymbol{y}) = \sum_k p_{GG}(u_i \mid \boldsymbol{y}, \boldsymbol{\theta}_k)\, \widetilde{p}(\boldsymbol{\theta}_k \mid \boldsymbol{y})\, w_k$$

For better accuracy, non-Gaussian conditional approximations can be used instead of $p_{GG}(u_i \mid \boldsymbol{y}, \boldsymbol{\theta}_k)$.

Example: Point process data

Log-Gaussian Cox processes

Point intensity:

$$\lambda(\boldsymbol{s}) = \exp\left(\sum_i b_i(\boldsymbol{s})\beta_i + u(\boldsymbol{s})\right)$$

Inhomogeneous Poisson process log-likelihood:

$$\ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) = |D| - \int_D \lambda(\boldsymbol{s})\, d\boldsymbol{s} + \sum_{k=1}^N \ln \lambda(\boldsymbol{y}_k)$$

The likelihood can be approximated numerically, e.g.

$$\int_D \lambda(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{j=1}^n \lambda(\boldsymbol{s}_j) w_j,$$

where s_j are finite element mesh nodes, and $w_j = \langle \psi_j, 1 \rangle_D$

Example: Point process data (cont)

Discretised field and likelihood:

$$\lambda(\boldsymbol{s}) = \exp\left(\sum_i b_i(\boldsymbol{s})\beta_i + \sum_j \psi_j(\mathbf{s})u_j\right)$$

$$\ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx |D| - \sum_{j=1}^n \lambda(\boldsymbol{s}_j)w_j + \sum_{k=1}^N \ln \lambda(\boldsymbol{y}_k)$$

Then, with $\boldsymbol{\lambda_D} = \left[\lambda(s_i)\right]$, $\boldsymbol{A}_D = \left[\psi_j(s_i)\right]$, and $\boldsymbol{A}_y = \left[\psi_j(y_i)\right]$,

$$\nabla_{\boldsymbol{u}} \ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx -\boldsymbol{A}_D^\top \operatorname{diag}(\boldsymbol{w})\boldsymbol{\lambda}_D + \boldsymbol{A}_y^\top \mathbf{1}$$

$$\nabla_{\boldsymbol{u}}^2 \ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx -\boldsymbol{A}_D^\top \operatorname{diag}(\boldsymbol{w}) \operatorname{diag}(\boldsymbol{\lambda}_D)\boldsymbol{A}_D$$

and similarly for ∇_β, ∇_β^2, and $\nabla_u \nabla_\beta$.

Standard INLA: Generalised linear models

In the R-INLA implementation, the data likelihood is only allowed to depend on a single, linear combination of latent Gaussian variables:

$$(\boldsymbol{u} \mid \boldsymbol{\theta}) \sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}) \qquad \text{(Prior)}$$

$$(y_i \mid \boldsymbol{u}, \boldsymbol{\theta}) \sim \mathsf{N}(g^{-1}(\eta_i), \boldsymbol{Q}_{y|u}^{-1}), \quad \boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{u} \qquad \text{(Gaussian observations, link function } g(\cdot)\text{)}$$

$$p(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{u} \mid \boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta}) \qquad \text{(Conditional posterior)}$$

Linear Gaussian observations

For a linear $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{u}$ and identity link $g^{-1}(\eta) = \eta$:

$$(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) \overset{\text{approx}}{\sim} \mathsf{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1}) \qquad \text{(Conditional posterior)}$$

$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_u + \boldsymbol{A}^\top \boldsymbol{Q}_{y|u} \boldsymbol{A}$$

$$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_u + \widetilde{\boldsymbol{Q}}^{-1} \left\{ \boldsymbol{A}^\top \boldsymbol{Q}_{y|u} [\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}_u)] \right\}$$

(For non-Gaussian observations or non-identity link, this is replaced by a Newton iteration.)

Extended INLA: inlabru

In the inlabru implementation, a linearisation step and iterated INLA allows non-linear predictors $\boldsymbol{\eta}$.:

$$(\boldsymbol{u} \mid \boldsymbol{\theta}) \sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}) \qquad \text{(Prior)}$$

$$(y_i \mid \boldsymbol{u}, \boldsymbol{\theta}) \sim \mathsf{N}(g^{-1}(\eta_i), \boldsymbol{Q}_{y|u}^{-1}), \quad \boldsymbol{\eta} = h(\boldsymbol{u}) \qquad \text{(Gaussian observations, link function } g(\cdot)\text{)}$$

$$p(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{u} \mid \boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta}) \qquad \text{(Conditional posterior)}$$

Non-linear Gaussian observations

For identity link, and a non-linear $h(\boldsymbol{u})$ with Jacobian $\boldsymbol{J}$ at $\boldsymbol{u} = \widetilde{\boldsymbol{\mu}}$, iterate:

$$(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) \overset{\text{approx}}{\sim} \mathsf{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1}) \qquad \text{(Approximate conditional posterior)}$$

$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_u + \boldsymbol{J}^\top \boldsymbol{Q}_{y|u} \boldsymbol{J}$$

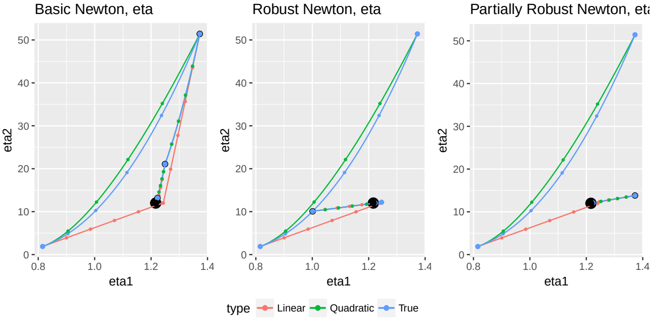
$$\widetilde{\boldsymbol{\mu}}' = \widetilde{\boldsymbol{\mu}} + a \widetilde{\boldsymbol{Q}}^{-1} \left\{ \boldsymbol{J}^\top \boldsymbol{Q}_{y|u} [\boldsymbol{y} - h(\widetilde{\boldsymbol{\mu}})] - \boldsymbol{Q}_u(\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}_u) \right\}$$

where a is chosen by line search.

(For non-Gaussian observations and non-identity link functions, this is modified appropriately.)

Extensions from the EUSTACE project, with $\sim 10^{11}$ latent variables

- Nonlinear Newton iteration with robust line-search



  The transformation $\eta = h(u)$ is much cheaper than the optimisation target.
- Preconditioned conjugate gradient (PCG) iteration for $\boldsymbol{Q}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Q}\widehat{\boldsymbol{\mu}}$
- Local and multiscale approximations for preconditioning: $\boldsymbol{M}^{-1}\boldsymbol{Q} \approx \boldsymbol{I}$
- Sampling with PCG: $\boldsymbol{Q}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{L}\boldsymbol{w}$
  Requires only a rectangular pseudo-Cholesky factorisation $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{Q}$.
  Sometimes possible, due to a kronecker product sum precision structure.

References (1/4)

- F. Lindgren, H. Rue and J. Lindström (2011), *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion)*, JRSSB, 73(4):423–498. Code available in `R-INLA`, see http://r-inla.org/
- Preliminary draft book chapter on GMRF computation basics (2015): http://www.maths.ed.ac.uk/~flindgre/tmp/gmrf.pdf
- R. Ingebrigtsen, F. Lindgren, I. Steinsland (2014), *Spatial models with explanatory variables in the dependence structure*, Spatial Statistics, 8:20–38. http://dx.doi.org/10.1016/j.spasta.2013.06.002
- G-A. Fuglstad, F. Lindgren, D. Simpson, H. Rue (2015), *Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy*, Statistica Sinica, 25:115–133. http://arxiv.org/abs/1304.6949
- Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian (2019) *inlabru: an R package for Bayesian spatial modelling from ecological survey data*, Methods in Ecology and Evolution, 10(6):760–766. https://doi.org/10.1111/2041-210X.13168
  CRAN package: `inlabru`, webpage https://inlabru.org/

References (2/4)

- Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, Andrew Gelman (Submitted on 18 Apr 2018)
  Validating Bayesian Inference Algorithms with Simulation-Based Calibration,
  http://arxiv.org/abs/1804.06788
- D. Bolin, F. Lindgren (2015),
  Excursion and contour uncertainty regions for latent Gaussian models,
  JRSSB, 77(1):85–106. http://arxiv.org/abs/1211.3946
  CRAN package: `excursions`
- H. Cramér and R. M. Leadbetter (2004, reprint of 1967 edition),
  Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications,
  Dover Books on Mathematics.
- G. Lindgren (2012),
  Stationary Stochastic Processes: Theory and applications,
  Chapman & Hall/CRC, Texts in Statistical Science.

References (3/4)

- J. Besag (1974),
  Spatial interaction and the statistical analysis of latice systems (with discussion), JRSSB, 36(2):192–225.
- J. Besag and C. Kooperberg (1995),
  On conditional and intrinsic autoregressions, Biometrika, 82(4):733–746.
- J. Besag (1981),
  On a system of two-dimensional recurrence equations, JRSSB, 43(3):302–309.
- J. Besag and D. Mondal (2005),
  First-order intrinsic autoregressions and the de Wij process, Biometrika, 92(4):909–920.
- Y. A. Rozanov (1977),
  Markov random fields and stochastic partial differential equations, Mathematics of the USSR – Sbornik, 32(4):515–534
- Y. A. Rozanov (1977),
  On the paper "Markov random fields and stochastic partial differential equations", Mathematics of the USSR – Sbornik, 35(1):157–164
- Y. A. Rozanov (1982, Russian original 1980),
  Markov Random Fields, Springer Verlag, New York.

References (4/4)

- Y. A. Rozanov (2010, reprint of 1st ed. 1998),
  Random fields and stochastic partial differential equations, Springer Verlag, Mathematics and its Applications.
- R. J. Adler and J. Taylor (2007),
  Random Fields and Geometry, Springer Monographs in Mathematics
- P. Whittle (1954),
  On stationary processes in the plane, Biometrika, 41(3/4):434–449.
- P. Whittle (1963),
  Stochastic processes in several dimensions, Bull. Inst. Internat. Statist., 40:974–994.
- B. Matérn (1960),
  Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. (Swedish: Stokastiska modeller och deras tillämpning på några problem i skogstaxering och andra samplingundersökningar.), Stockholm University, PhD thesis in Mathematical Statistics. You may be lucky and find someone who has a copy!