Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 2: Introduction to INLA & R-INLA

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce Laplace approximation, and Integrated Laplace Approximation computational methods. To introduce

Reading list & references:

(1) Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.
- Ch. 4.6-4.9; pp.104-126

(2) Turkman, M. A. A., Paulino, C. D., & Müller, P. (2019). Computational Bayesian statistics: an introduction (Vol. 11). Cambridge University Press.
- Ch. 8

(3) Rue, H., Martino, S., & Chopin, N. (2009). 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.

1. Laplace Approximation (LA)

Proposition 1. *Consider integral*

$$I = \int \exp\left(nL\left(\theta\right)\right) d\theta$$

where $\theta \in \mathbb{R}^d$. *Laplace approximation (LA) method produces approximation* $I \approx \hat{I}$

$$\hat{I} = \left(2\pi\right)^{\frac{d}{2}} \left(n\right)^{-\frac{d}{2}} \left(\det\left(\Sigma\right)\right)^{\frac{1}{2}} \exp\left(nL\left(\hat{\theta}\right)\right)$$

where $\hat{\theta}$ *is the maximum of* $L\left(\cdot\right)$ *and* $\Sigma = -\left(H\left(\hat{\theta}\right)\right)^{-1}$ *with Hessian* $H\left(\hat{\theta}\right) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(L\left(\theta\right)\right)\Big|_{\theta=\hat{\theta}}$.

Proof. Sketch of the proof. Take 2nd order Taylor expansion of $L\left(\theta\right)$ around $\hat{\theta}$ i.e.

$$L\left(\theta\right) \approx L\left(\hat{\theta}\right) + \left(\theta - \hat{\theta}\right) \underbrace{\nabla L\left(\hat{\theta}\right)}_{=0} + \frac{1}{2}\left(\theta - \hat{\theta}\right)^{\top} H\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right) \tag{1.1}$$

then

$$\begin{aligned} I &\approx \int \exp\left(nL\left(\hat{\theta}\right) + n\left(\theta - \hat{\theta}\right)^{\top} H\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right)\right) \mathrm{d}\theta \\ &= \exp\left(nL\left(\hat{\theta}\right)\right) \int \exp\left(-\frac{1}{2}\left(\theta - \hat{\theta}\right)^{\top} \left(\left(-nH\left(\hat{\theta}\right)\right)^{-1}\right)^{-1} \left(\theta - \hat{\theta}\right)\right) \mathrm{d}\theta \\ &= \exp\left(nL\left(\hat{\theta}\right)\right) \left(2\pi\right)^{\frac{d}{2}} \left(\det\left(\left(-nH\left(\hat{\theta}\right)\right)^{-1}\right)\right)^{\frac{1}{2}} \end{aligned}$$

Given regularity conditions related to the Taylor expansions (1.1), it can be shown that $I = \hat{I}\left(1 + O\left(n^{-1}\right)\right)$ (not discussed here). □

Example 2. Consider posterior expectation

$$\mathrm{E}\left(g\left(\theta\right)|z\right) = \int g\left(\theta\right) \mathrm{pr}\left(\theta|z\right) \mathrm{d}\theta \tag{1.2}$$

of a function $g\left(\cdot\right)$ of the parameter $\theta \in \mathbb{R}^d$ given observables z. Laplace method can produce approximation$\mathrm{E}\left(g\left(\theta\right)|z\right) \approx \widehat{\mathrm{E}\left(g\left(\theta\right)|z\right)}$

$$\widehat{\mathrm{E}\left(g\left(\theta\right)|z\right)} = \left(\frac{\det\left(\Sigma^*\right)}{\det\left(\Sigma\right)}\right)^{\frac{1}{2}} \exp\left(n\left(L^*\left(\hat{\theta}^*\right) - L\left(\hat{\theta}\right)\right)\right) \tag{1.3}$$

where $\hat{\theta}$ and Σ are the mode and minus the inverse Hessian of $L\left(\theta\right) = \log\left(\mathrm{pr}\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(z|\theta\right)\right)/n$ while $\hat{\theta}^*$ and Σ^* are the mode and minus the inverse Hessian of $L^*\left(\theta\right) = \log\left(g\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(z|\theta\right)\right)/n$.

Solution. (Sketch of the solution) It is

$$\mathrm{E}\left(g\left(\theta\right)|z\right) = \frac{\int g\left(\theta\right)\mathrm{pr}\left(z|\theta\right)\mathrm{pr}\left(\theta\right)\mathrm{d}\theta}{\int \mathrm{pr}\left(z|\theta\right)\mathrm{pr}\left(\theta\right)\mathrm{d}\theta} = \frac{\int \exp\left(nL^*\left(\theta\right)\right)\mathrm{d}\theta}{\int \exp\left(nL\ \ \left(\theta\right)\right)\mathrm{d}\theta} \overset{(\star)}{\approx} \frac{\left(2\pi n\right)^{d/2}\sqrt{\det\left(\Sigma^*\right)}\exp\left(nL^*\left(\hat{\theta}^*\right)\right)}{\left(2\pi n\right)^{d/2}\sqrt{\det\left(\Sigma\ \right)}\exp\left(nL\left(\hat{\theta}\ \right)\right)}$$

where $(\star)$ is by applying Proposition 1 once at the top and once at the bottom of the fraction. Then the result follows.

Under regularity conditions related to Taylor expansion (not discussed here), it is $\mathrm{pr}\left(\theta_1|z\right) = \widehat{\mathrm{pr}\left(\theta_1|z\right)}\left(1 + O_{\theta_1}\left(n^{-1}\right)\right)$ where the lower index indicates the dependence of the constant on θ_1.

Example 3. Consider the marginal posterior density of $\theta_1 \in \mathbb{R}$

$$\mathrm{pr}\left(\theta_1|z\right) = \int \mathrm{pr}\left(\theta_1, \theta_2|z\right)\mathrm{d}\theta_2 \tag{1.4}$$

under a Bayesian model with observable $z \sim \mathrm{pr}\left(z|\theta\right)$ and unknown parameter $\theta = \left(\theta_1, \theta_2\right) \in \mathbb{R}^d$ with $\theta \sim \mathrm{pr}\left(\theta\right)$. Laplace method can produce approximation

$$\widehat{\mathrm{pr}\left(\theta_1|z\right)} = \left(\frac{\det\left(\Sigma^*\left(\theta_1\right)\right)}{2\pi n \det\left(\Sigma\right)}\right)^{\frac{1}{2}} \frac{\mathrm{pr}\left(\theta_1, \hat{\theta}_2\left(\theta_1\right)\right)\exp\left(\log\left(\mathrm{pr}\left(z|\theta_1, \hat{\theta}_2\left(\theta_1\right)\right)\right)\right)}{\mathrm{pr}\left(\hat{\theta}\right)\exp\left(\log\left(\mathrm{pr}\left(z|\hat{\theta}\right)\right)\right)} \tag{1.5}$$

where $\hat{\theta}$ is the maximizer of $\log\left(\mathrm{pr}\left(\theta_1, \theta_2\right)\right) + \log\left(\mathrm{pr}\left(z|\theta_1, \theta_2\right)\right)$,

Σ is the minus Hessian of $n^{-1}\left(\log\left(\mathrm{pr}\left(\theta_1, \theta_2\right)\right) + \log\left(\mathrm{pr}\left(z|\theta_1, \theta_2\right)\right)\right)$,

$\hat{\theta}_2\left(\theta_1\right)$ is the maximizer of $\log\left(\mathrm{pr}\left(z|\theta_1, \cdot\right)\right) + \log\left(\mathrm{pr}\left(\theta_1, \cdot\right)\right)$,

$\Sigma^*\left(\theta_1\right)$ is the minus Hessian of $n^{-1}\left(\log\left(\mathrm{pr}\left(z|\theta_1, \cdot\right)\right) + \log\left(\mathrm{pr}\left(\theta_1, \cdot\right)\right)\right)$

Solution. (Sketch of the solution) It is

$$\begin{aligned}\mathrm{pr}\left(\theta_1|z\right) &= \frac{\int \mathrm{pr}\left(\theta_1,\theta_2\right)\mathrm{pr}\left(z|\theta_1,\theta_2\right)\mathrm{d}\theta_2}{\int \mathrm{pr}\left(\theta_1,\theta_2\right)\mathrm{pr}\left(z|\theta_1,\theta_2\right)\mathrm{d}\theta} = \frac{\int \exp\left(nL^*_{\theta_1}\left(\theta_2\right)\right)\mathrm{d}\theta_2}{\int \exp\left(nL\ \ \left(\theta\right)\right)\mathrm{d}\theta} \\ &\overset{(\star)}{\approx} \left(\frac{\det\left(\Sigma^*\left(\theta_1\right)\right)}{2\pi n \det\left(\Sigma\right)}\right)^{\frac{1}{2}} \frac{\mathrm{pr}\left(\theta_1,\hat{\theta}_2\left(\theta_1\right)\right)\exp\left(\log\left(\mathrm{pr}\left(z|\theta_1,\hat{\theta}_2\left(\theta_1\right)\right)\right)\right)}{\mathrm{pr}\left(\hat{\theta}\right)\exp\left(\log\left(\mathrm{pr}\left(z|\hat{\theta}\right)\right)\right)}\end{aligned}$$

where $L^*_{\theta_1}\left(\theta_2\right) = n^{-1}\left(\log\left(\mathrm{pr}\left(\theta_1,\theta_2\right)\right) + \log\left(\mathrm{pr}\left(z|\theta_1,\theta_2\right)\right)\right)$ and $L\left(\theta\right) = n^{-1}\left(\log\left(\mathrm{pr}\left(\theta\right)\right) + \log\left(\mathrm{pr}\left(z|\theta\right)\right)\right)$. Here $(\star)$ results by applying Proposition 1 once at the top and once at the bottom of the fraction. Then the result is implied.

Under regularity conditions related to Taylor expansion (not discussed here), it is $\mathrm{pr}\left(\theta_1|z\right) = \widehat{\mathrm{pr}\left(\theta_1|z\right)}\left(1 + O_{\theta_1}\left(n^{-1}\right)\right)$ where the lower index indicates the dependence of the constant on θ_1.

2. Integrated Nested Laplace Approximation (INLA)

2.1. **Motivations.**

Note 4. Integrated Nested Laplace Approximation (INLA) can directly compute very accurate approximations to posterior marginals and summary statistics of statistical models with a specific type (such as those discussed in the module) even if they are high-dimensional or involve large datasets. In such models, MCMC methods may need hours or days to run, which INLA can provide more precise estimates in seconds or minutes for a certain type of models we will discuss.

2.2. **Where it can be applied; implementations.**

Note 5. INLA is suitable to facilitate Bayesian inference in spatial statistical problems related to Latent Gaussian Models (LGM).

Note 6. The class of Latent Gaussian Models (LGM) can be represented in a three level hierarchical model structure. The first level is the sampling model where the observations $z = \left(z_1,...,z_n\right)^\top$ can be assumed to be conditionally independent, given a latent random field $y = \left(y_1,...,y_n\right)^\top$ and hyper-parameter θ_1, i.e.

$$z|y,\theta_1 \sim \mathrm{pr}\left(z|y,\theta_1\right) = \prod_{i=1}^{n}\mathrm{pr}\left(z_i|y_i,\theta_1\right). \tag{2.1}$$

The second level assumes that y follows a multivariate Gaussian distribution (Essentially a Gaussian random field) given hyper-parameter θ_2 , i.e.

$$y|\theta \sim \mathrm{N}\left(\mu\left(\theta_2\right),\left(Q\left(\theta_2\right)\right)^{-1}\right) \tag{2.2}$$

The third level (relevant only to fully Bayesian statistical models) specifies a prior on the unknown parameter $\theta = (\theta_1, \theta_2)^\top$, i.e.

$$\theta \sim \text{pr}(\theta)$$

Assumption 7. *For the computational purposes of INLA, we make assumption that (2.2) is defined wrt an undirected graph* $\mathcal{G} = (\mathcal{V} = \{1, ...n\}, \mathcal{E})$ *such that*

$$y_l \perp y_m | y_{-\{l,m\}}, \quad \forall \{l, m\} \notin \mathcal{E} \tag{2.3}$$

This leads to sparse precision matrix $Q(\theta_2)$ *because*

$$y_l \perp y_m | y_{-\{l,m\}} \Leftrightarrow [Q(\theta_2)]_{l,m} = 0$$

This makes (2.2) be a Gaussian Markov Random Field (GMRF).

Note 8. The LGM (under consideration) is summarized to

$$\begin{aligned} z|y, \theta &\sim \text{pr}(z|y, \theta) = \prod_{i=1}^{n} \text{pr}(z_i|y_i, \theta) && \text{(sampling model for } z\text{)} \\ y|\theta &\sim \text{pr}_{\mathcal{G}}(y|\theta) && \text{(GMRF prior for } y\text{)} \\ \theta &\sim \text{pr}(\theta) && \text{(hyperprior for } \theta\text{)} \end{aligned} \tag{2.4}$$

Note 9. The joint posterior probability model is

$$\begin{aligned} \text{pr}(y, \theta|z) &\propto \prod_{i=1}^{n} \text{pr}(z_i|y_i, \theta) \, \text{pr}(y|\theta) \, \text{pr}(\theta) \\ &\propto \exp\left(-\frac{1}{2} (y - \mu(\theta))^\top Q(\theta) (y - \mu(\theta)) + \sum_{i=1}^{n} \log(\text{pr}(z_i|y_i, \theta)) \right) \text{pr}(\theta) \end{aligned} \tag{2.5}$$

and hence there is interest in computing the marginal densities and expectations of $y_i|z$, and $\theta_i|z$ as well as predictions of unseen y's.

Assumption 10. *For INLA to perform most efficiently (fast) and accurately (due to approximations), we make the following critical assumptions:*

1. *The number of hyperparameters* θ *is small, typically* 2 *to* 5*, but not exceeding* 20*.*
2. *pr*$(y|\theta)$ *is required to be a GMRF (or close to one) when the dimension* n *is high* (103–105)*.*
3. *The data* $\{z_i\}$ *are mutually conditionally independent of* y *and* θ*, implying that each observation* z_i *only depends on one component of the latent field, for example,* y_i*. Most components of* y_i *will not be observed.*

Note 11. LGM in (2.4) can be specified as a special case of a regression model whose response z_i are assumed to follow an exponential family distribution with mean $\mu_i = \text{E}(z_i|y_i, \theta)$ linked

to a Gaussian linear predictor η_i via a known link function $g(\cdot)$, as $g(\mu_i) = \eta_i$ and

$$\eta_i = \alpha + \sum_j \beta_j x_{j,i} + \sum_k f_k(u_{ki}) + \epsilon_i \tag{2.6}$$

where α is the intercept, $\{\beta_j\}$ are coefficients (fixed effects) of covariates $\{x_{j,i}\}$, and $f_k(\cdot)$ are unknown functions of covariates u, and ϵ_i is a random error. Casting it as an LGM, we can set

$$y = (\alpha, \{\beta_j\}, \{f_k(u_{ki})\}, \{\eta_i\})$$

is the latent field in (2.4) (for conveniency, we consider η_i instead of ϵ), and the rest hyperparameters (to be learned) constitute θ.

Note 12. Consequently the class LGM involves many computationally challenging models, such as the spatial models (geostatistical, latent, point process), the associated spatiotemporal models, and the mixed effect GLM.

2.3. **The general idea.**

Note 13. We are interested in computing the following marginals of (2.5)

$$\mathrm{pr}(\theta_j|z) = \int\int \mathrm{pr}(y,\theta|z)\,\mathrm{d}y\mathrm{d}\theta_{-j} = \int \mathrm{pr}(\theta|z)\,\mathrm{d}\theta_{-j} \tag{2.7}$$

$$\mathrm{pr}(y_i|z) = \int\int \mathrm{pr}(y,\theta|z)\,\mathrm{d}y_{-i}\mathrm{d}\theta = \int \mathrm{pr}(y_i|z,\theta)\,\mathrm{pr}(\theta|z)\,\mathrm{d}\theta \tag{2.8}$$

where integrals (2.7) and (2.8) can be of high dimentionality wrt y.

Note 14. For the approximation of (2.7) and (2.8), INLA involves three steps: evaluation of $\mathrm{pr}(y_i|z,\theta)$ via Laplace approx, evaluation of $\mathrm{pr}(\theta|z)$ via Laplace approx, and finally numerical integration.

Note 15. To compute an approximate for $\mathrm{pr}(\theta|z)$, notice that at any point y it is

$$\mathrm{pr}(\theta|z) = \frac{\mathrm{pr}(y,\theta|z)}{\mathrm{pr}(y|z,\theta)} \propto \frac{\mathrm{pr}(z|y,\theta)\,\mathrm{pr}(y|\theta)\,\mathrm{pr}(\theta)}{\mathrm{pr}(y|z,\theta)} \tag{2.9}$$

Unlike the numerator, the denumerator is not available in closed form and is hard to compute. INLA employs the approximation of $\mathrm{pr}(y|z,\theta)$ by a multivariate Gaussian distribution $\widetilde{\mathrm{pr}}_{\mathrm{G}}(y|z,\theta)$ whose mean is the mode $y^*(\theta)$ of $\mathrm{pr}(y|z,\theta)$ and covariance matrix is the minus inverse Hessian at that mode. Essentially, the approximation of (2.9) at a specific value of θ is

$$\widetilde{\mathrm{pr}}(\theta|z) \propto \left.\frac{\mathrm{pr}(z|y,\theta)\,\mathrm{pr}(y|\theta)\,\mathrm{pr}(\theta)}{\widetilde{\mathrm{pr}}_{\mathrm{G}}(y|z,\theta)}\right|_{y=y^*(\theta)} \tag{2.10}$$

which is equivalent to the Laplace approximation method for marginal densities.

Note 16. To compute an approximate for $\text{pr}(y_i|z,\theta)$ at each y_i there are three main approaches:

Gaussian approximation approach.: Compute the marginal from the Gaussian approximation $\widetilde{\text{pr}}_{\text{G}}(y|z,\theta)$ of $\text{pr}(y|z,\theta)$ in Note 15. This is fast but not generally accurate.

Laplace approximation: Similar to Note 15, compute

$$\widetilde{\text{pr}}(y_i|z,\theta) \propto \left.\frac{\text{pr}(z|y,\theta)\,\text{pr}(y|\theta)\,\text{pr}(\theta)}{\widetilde{\text{pr}}_{\text{G}}(y_{-i}|y_i,z,\theta)}\right|_{y=y^*(\theta)} \tag{2.11}$$

where $\widetilde{\text{pr}}_{\text{G}}(y_{-i}|y_i,z,\theta)$ is a multivariate Gaussian distribution whose mean is the mode $y^*_{-i}(y_i,\theta)$ and covariance matrix is the minus inverse Hessian at that mode. It is more accurate than the previous one but computational demanding because it requires the the re-calculation of the precision matrix for each y_i.

Simplified Laplace approximation: It builds on third order Taylor series expansions both in numerator and denumerator of (2.11), which improves the approximation wrt asymmetry. We skip the mathematical details here. It has improved accuracy.

2.4. **The schematic of the procedure.**

Algorithm 17. *Summing up, the INLA method proceeds as follows:*

(1) Explore the space of θ.
   (a) Locate a collection of points $\left\{\theta^{(k)}; k=1,...,K\right\}$ in the area of high density of $\widetilde{\text{pr}}(\theta|z)$.
   (b) Find the mode of $\widetilde{\text{pr}}(\theta|z)$.

(2) Compute approximation $\widetilde{\text{pr}}(\theta|z)$ at points $\left\{\theta^{(k)}; k=1,...,K\right\}$ by using (2.10).

(3) Compute approximation $\widetilde{\text{pr}}(y_i|z,\theta)$ at points $\left\{\theta^{(k)}; k=1,...,K\right\}$ of θ by using the Laplace approximation in (2.11) or the simplified Laplace approximation, or the Gaussian approximation, as said in Note 16.

(4) Compute the approximation $\widetilde{\text{pr}}(y_i|z)$ of (2.8) via standard numerical approximation as

$$\widetilde{\text{pr}}(y_i|z) = \sum_{k=1}^{K} \widetilde{\text{pr}}\left(y_i|z,\theta^{(k)}\right) \widetilde{\text{pr}}\left(\theta^{(k)}|z\right) \Delta^{(k)} \tag{2.12}$$

where $\Delta^{(k)}$ as weights depending on the locations $\left\{\theta^{(k)}\right\}$ and the numerical integration scheme. If $\left\{\theta^{(k)}\right\}$ are equal-distant then $\Delta^{(k)} = 1$.

(5) Compute the approximation $\widetilde{\text{pr}}(y_i|z)$ of (2.8) via standard numerical approximation as

$$\widetilde{\text{pr}}(\theta_j|z) = \sum_{k=1}^{K} \widetilde{\text{pr}}\left(\theta_{-j}, \theta_{-j}^{(k)}|z\right) \Delta^{(k)} \tag{2.13}$$

where $\Delta^{(k)}$ as weights depending on the locations $\left\{\theta_{-j}^{(k)}\right\}$ and the numerical integration scheme. If $\left\{\theta^{(k)}\right\}$ are equal-distant then $\Delta^{(k)} = 1$.

Note 18. The error in (2.12) comes from the Laplace approximations in $\widetilde{\text{pr}}\left(\theta^{(k)}|z\right)$ and $\widetilde{\text{pr}}\left(y_i|z, \theta^{(k)}\right)$, as well as the numerical integration and the choice of locations $\left\{\theta^{(k)}\right\}$. When the likelihood $\text{pr}\left(y|z, \theta^{(k)}\right)$ is Gaussian then its marginals are Gaussian and hence this error is eliminated.

2.5. **Byproducts.**

Note 19. Marginal likelihood $\text{pr}(z)$ is often used in Bayesian model comparison, and model averaging. A natural approximation for the marginal likelihood $\text{pr}(z)$ is

$$\widetilde{\text{pr}}(z) = \int \left. \frac{\text{pr}(z|y,\theta)\,\text{pr}(y|\theta)\,\text{pr}(\theta)}{\widetilde{\text{pr}}_{\text{G}}(y|z,\theta)} \right|_{y=y^*(\theta)} \text{d}\theta$$

The approx can fail when $\text{pr}(\theta|z)$ is multimodal, however LGM generate unimodal posteriors in most cases.

Note 20. Deviance Information Criterion (DIC) can be used in Bayesian model comparison. Analogously to AIC, the deviance of the model is

$$D(\theta) = -2\log(\text{pr}(z|\theta)),$$

the model complexity here is measured via effective number of parameters

$$p_D = \text{E}(D(\theta)|z) - D(\text{E}(\theta|z))$$

and hence DIC is defined as

$$\text{DIC} = \text{E}(D(\theta)|z) + p_D.$$

Models with smaller DIC are better supported by the data. INLA approximates integrals/expectations numerically after (2.10) has been approximated.

Note 21. Predictive distribution of an unseen value z^{new} (includes missing data) given the observables z and model (2.4) is

$$\text{pr}\left(z^{\text{new}}|z\right) = \int \text{pr}\left(z^{\text{new}}|y^{\text{new}}\right) \text{pr}\left(y^{\text{new}}|z\right) \text{d}y^{\text{new}} \tag{2.14}$$

$$\text{pr}\left(y^{\text{new}}|z\right) = \int \text{pr}\left(y^{\text{new}}|\theta\right) \text{pr}\left(\theta|z\right) \text{d}\theta \tag{2.15}$$

due to the conditional independence in (2.1). Given that (2.10) has been approximated, INLA employs numerical integration for the integral (2.15) firstly and 2.14 secondly.

3. The R-INLA software (An empirical introduction)

Note 22. All the info is int he website of the software `https://www.r-inla.org`

3.1. **How to install R-INLA.**

Note 23. To install R-INLA do the following from `https://www.r-inla.org/download-install`.

```
# install the stable version, do
install.packages("INLA",repos=c(getOption("repos"),
    INLA="https://inla.r-inla-download.org/R/stable"),
    dep=TRUE)
install.packages("INLA",repos=c(getOption("repos"),
    INLA="https://inla.r-inla-download.org/R/testing"),
    dep=TRUE)
# update the stable version the package
inla.upgrade()
# install dependency fmesh R package
options(repos=c( inlabruorg = "https://inlabru-org.r-universe.dev",
     INLA = "https://inla.r-inla-download.org/R/testing",
     CRAN = "https://cran.rstudio.com")
     )
install.packages("fmesher")
```

3.2. **How to use R-INLA.**

Note 24. There are two essential steps:

(1) Define the linear predictor (2.6) through a formula object
(2) Complete the model definition and fit the model using the R function `inla{INLA}`.

The fitted model is returned as an inla object.

Example 25. We analyze the R dataset `Salm{INLA}`.

- Bayesian model

$$\begin{cases} z_{i,j}|\lambda_{i,j} \sim \text{Poi}\left(\lambda_{i,j}\right) & i = 1, ..., 6 \text{ and } j = 1, 2, 3 \\ \log\left(\lambda_{i,j}\right) = \beta_0 + \beta_1 \log\left(x_i + 10\right) + \beta_2 x_i + u_{i,j} & i = 1, ..., 6 \text{ and } j = 1, 2, 3 \end{cases}$$

  where $\{z_{i,j}\}$ (the observables) are number of colonies found on plate j for dose i and x_i indicate the ith dose. Let $u_{ij}|\tau \sim \text{N}\left(0, \sigma^2\right)$ be the so-called random effects, while $\{\beta_i\}$ are unknown parameters called fixed effects.
- In terms of model (2.4), the GMRF is $y = \left(\{\lambda_{i,j}\}, \{\beta_i\}, \{u_{i,j}\}\right)$.
- We consider prior on σ^2 such that

$$\tau = -\log\left(\sigma^2\right) \sim \text{type-2 Gimbel}(1/2, -\log\left(a\right)/u)$$

  This is because R-INLA specifies prior on $\tau = -\log\left(\sigma^2\right)$.

Data loading.

- Load R-INLA

```
# load the data set
library(‘‘INLA’’)
```

- We import the R data set `Salm{INLA}` as follows

```
# load the data set
data(Salm)
# get info about the R dataset
?Salm
# rename the columns to fit the notation
names(Salm) = c("z", "x", "u")
```

Training via R-INLA.

- Code the model in R-INLA language, and produce the inla object

```
# specify the prior for the log precision parameter
my.hyper <- list(theta = list(prior="pc.prec", param=c(1,0.01)))
# specify the linear predictor
formula <- z ~ log(x + 10) + x + f(u, model = "iid", hyper = my.hyper)
# run R-INLA and get the result object
result <- inla(formula=formula, data=Salm, family="Poisson",
    control.inla = list(strategy=’’laplace’’))
```

- The ‘`formula`’ is as in `lm{stats}` command.
- Function ‘`inla.list.models()`’ provides a list of available distributions for the different parts of the model, such as the "prior" (available priors for the hyper-parameters), "likelihood" (all implemented likelihoods) and "latent" (available models for the latent field).

- Function `f()` is used to specify the latent Gaussian model for the non-linear terms and random effect u_{ij}; here an independent noise model (hence the use of model = "iid"), and the hyperprior for its corresponding hyperparameters (here σ^2).
- R function `inla{INLA}` (given the input above) generates an inla object similar to that of `lm{stats}`. The data object should be data.frame or list. The likelihood is specified in form of a string. `strategy=''laplace''` refers to the approximation strategy in Note 16 and has options "gaussian", "simplified.laplace", "laplace".

Parametric inference.

- Post-processing the results from `inla` object.

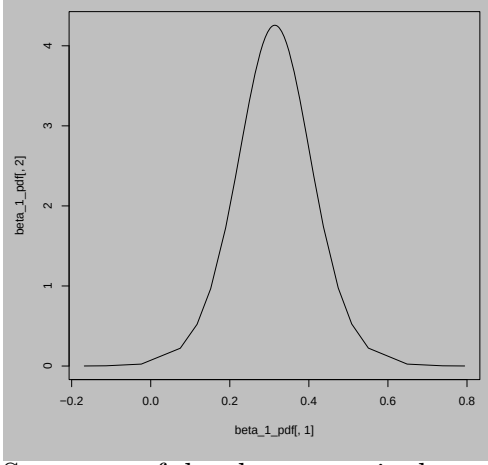
```
summary(result)
```

```
Time used:
    Pre = 0.343, Running = 0.156, Post = 0.0147, Total = 0.514
Fixed effects:
              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept)  2.165 0.362      1.445    2.166      2.880  2.167   0
log(x + 10)  0.313 0.099      0.117    0.313      0.508  0.314   0
x           -0.001 0.000     -0.002   -0.001      0.000 -0.001   0

Random effects:
  Name    Model
    u IID model

Model hyperparameters:
                  mean    sd 0.025quant 0.5quant 0.975quant  mode
Precision for u 20.64 16.52       5.72    16.44      59.79 11.91

Marginal log-Likelihood:  -83.69
 is computed
Posterior summaries for the linear predictor and the fitted values are computed
(Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

It provides summary statistics of the posterior of the fixed effect, random effect, and precision parameters, as well as the marginal log-likelihood $\log(\text{pr}(z))$.

- Marginal posteriors for the fixed effect, random effect, and hyperparameters are stored in `result$marginals.fixed`, `result$marginals.random`, `result$marginals.hyperpar`. E.g., one can plot the posterior of β_1 as

```
beta_1_pdf <- result$marginals.fixed$`log(x + 10)`
plot(beta_1_pdf[,1], beta_1_pdf[,2], type="l")
```



- Summary of the above marginal posteriors can be obtained by using `result$summary.fixed`, `result$summary.random`, `result$summary.hyperpar`

result$marginals.fixed

```
> result$summary.fixed
                     mean           sd   0.025quant      0.5quant    0.975quant          mode
(Intercept)  2.1647643605 0.3620126799  1.444666455  2.1655831923  2.879995e+00  2.1669703669
log(x + 10)  0.3132991434 0.0985605383  0.117201855  0.3134878885  5.084337e-01  0.3139144159
x           -0.0009656845 0.0004357064 -0.001827388 -0.0009671395 -9.635679e-05 -0.0009702587
                     kld
(Intercept) 1.419280e-08
log(x + 10) 2.901292e-08
x           4.525820e-08
```

result$summary.hyperpar

```
> result$summary.hyperpar
                      mean       sd 0.025quant 0.5quant 0.975quant     mode
Precision for u 20.64402 16.51935    5.72236 16.44435   59.78984 11.90988
```

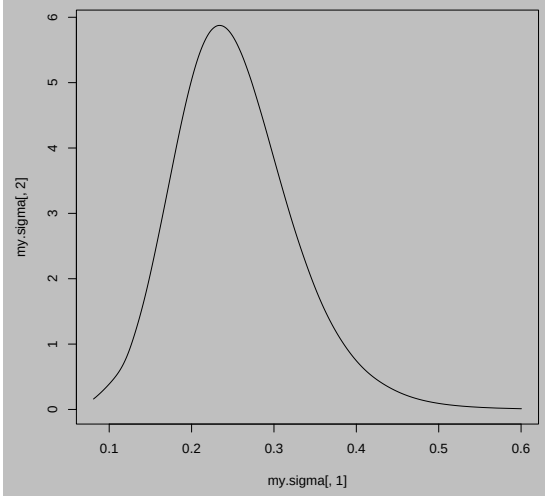
- To get the posterior summary of a function of the parameters, e.g. the posterior mean and standard deviation of $\sigma^2 = \exp(\tau)$

```
# Select the right hyperparameter marginal
tau <- result$marginals.hyperpar[[1]]
# Compute the expected value for 1/sqrt(tau) and 1/sqrt(tau)^2
E = inla.emarginal(function(x) c(1/sqrt(x),(1/sqrt(x))^2), tau)
# From this we computed the posterior standard deviation as
mysd = sqrt(E[2] - E[1]^2)
# so that we obtain the posterior mean and standard deviation
print(c(mean=E[1], sd=mysd))
```

```
      mean         sd
0.25353753 0.07325247
```

- To compute the marginal posterior distribution of $\sigma^2 = \exp(\tau)$ use the `inla.tmarginal()`

```
# Select the right hyperparameter marginal
tau <- result$marginals.hyperpar[[1]]
# Do the transformation
my.sigma <- inla.tmarginal(function(x){1/sqrt(x)}, tau)
# plot
plot(my.sigma[,1], my.sigma[,2], type="l")
```



- Other R-INLA functions providing operations on posterior marginals can be found in R help documentation,

```
?inla.marginal
```

Predictive inference.

- In R-INLA there is no function `predict{stats}` as for `glm{stats}` or `lm{stats}`. Predictions must be done as a part of the model fitting itself. Prediction can be regarded as fitting a model with missing data, hence we can simply set `y[i]=NA` for those "locations" we want to predict. Predictive distributions, which are often of interest, are however not returned directly, and the user needs to some extra "hacks". There are two reasonable "hacks".
- For illustration, pretend 7th observation is unknown, by removing it from the training data, and try to predict it.

```
## set observation 7 to NA
Salm.predict = Salm
Salm.predict[7, "y"] <- NA
# re-run the model
res.predict = inla(formula=formula, data=Salm,     family="Poisson",
    control.predictor = list(compute = TRUE),
    control.family = list(control.link=list(model="log")) )
```

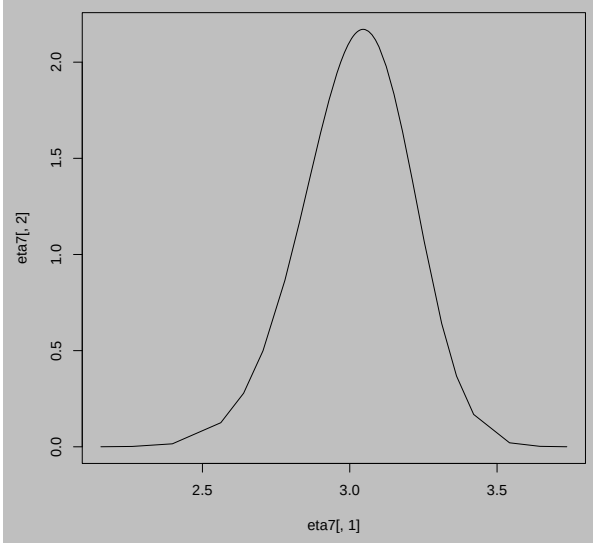
- Using the same settings as before, train the model by function `inla(INLA)`.

```
# re-run the model
res.predict = inla(formula=formula, data=Salm,     family="Poisson",
    control.predictor = list(compute = TRUE),
    control.compute=list(return.marginals.predictor=TRUE),
    control.family = list(control.link=list(model="log")) )
```

By specifying `control.predictor=list(compute=TRUE)` the posterior marginals will be included in the results object. We also need to explicitly specify the link function g connecting $g(\lambda_i) = \eta_i$, where $\lambda_i = \mathrm{E}(z_i)$, using the `control.family` object in order for `inla()` to compute the linear predictor η_i. Note that here $\lambda_i = \exp(\eta_i)$. By specifying `control.compute=list(return.marginals.predictor=TRUE)`, we ask function `inla(INLA)` to compute and return the marginal pdf of the linear predictor, which be default are not due to computational cost.

- We can compute $\mathrm{pr}(\eta_7|z_{-7})$ by

```
# marginal posterior for the linear predictor
eta7 = res.predict$marginals.linear.predictor[[7]]
```

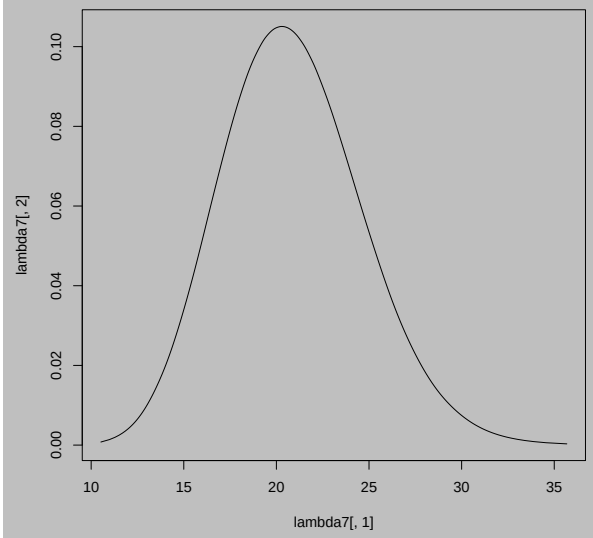


- Summary about is $\mathrm{pr}(\eta_7|z_{-7})$ taken by

```
# some summary statistics round(res.predict$summary.linear.predictor[7,], 3)
> res.predict$summary.linear.predictor[7,]
                  mean        sd 0.025quant 0.5quant 0.975quant     mode          kld
Predictor.07 3.021652 0.1847223   2.639797 3.029161   3.362469 3.045581 1.224947e-07
```

- We can compute $\text{pr}\,(\lambda_7|z_{-7})$ by

```
# marginal posterior for lambda
eta7 = res.predict$marginals.linear.predictor[[7]]
lambda7 = inla.tmarginal(function(x){exp(x)}, eta7)
## or alternatively
# lambda7_bis = inla.tmarginal(function(x){exp(x)},eta7)
## or alternatively
# lambda7_bis = inla.tmarginal(function(x){exp(x)},eta7)
# plot
plot(lambda7[,1], lambda7[,2], type="l")
```

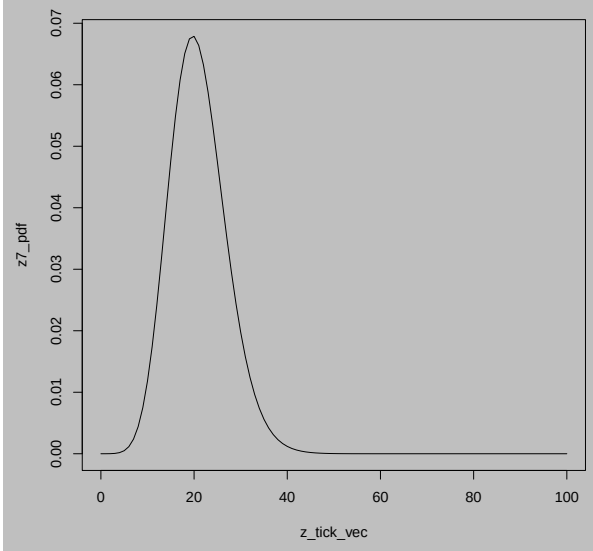


- To compute $\text{pr}\,(z_7|z_{-7})$ i.e. the predictive distribution (in this case) or the posterior distribution of the missing value (in principle), we can consider the following integration

$$\begin{aligned} \text{pr}\,(z_7|z_{-7}) &= \int \text{pr}\,(z_7|\lambda_7)\,\text{pr}\,(\lambda_7|z_{-7})\,\text{d}\lambda_7 \\ &\approx \int \widetilde{\text{pr}}\,(z_7|\lambda_7)\,\widetilde{\text{pr}}\,(\lambda_7|z_{-7})\,\text{d}\lambda_7 \end{aligned} \tag{3.1}$$

and either approximated by using numerical integration, e.g. trapezoid rule with R function `trapz{caTools}`

```
# library supporting trapezoid rule integration.
library(caTools)
# specify the support at which we want to compute the density
z_tick_vec = 0:100
z7_pdf = rep(0,101)
# go over the posterior marginal of the fitted value
for(j in 1:(length(lambda7[ ,1])-1)) {
    z7_pdf <- z7_pdf + dpois(z_tick_vec,
        lambda = ((lambda7[j,1]+ lambda7[j+1,1])/2))
            * trapz(lambda7[j:(j+1), 1], lambda7[j:(j+1), 2])
}
# plot
plot(z_tick_vec,z7_pdf, type="l")
```

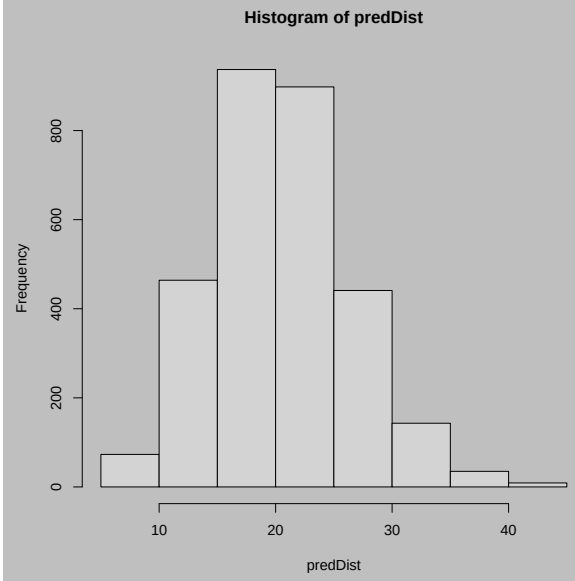


- alternatively one approximate (3.1) by Monte Carlo integration

$$\begin{aligned} \mathrm{pr}\left(z_7|z_{-7}\right) &\approx \mathrm{E}_{\widetilde{\mathrm{pr}}(\lambda_7|z_{-7})}\left(\widetilde{\mathrm{pr}}\left(z_7|\lambda_7\right)\right) \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\widetilde{\mathrm{pr}}\left(z_7|\lambda_7^{(t)}\right) \end{aligned} \tag{3.2}$$

where $\left\{\lambda_7^{(t)}\right\}_{t=1}^{T}$ is a sample drawn from $\widetilde{\mathrm{pr}}\left(\lambda_7|z_{-7}\right)$ by using function inla.rmarginal{INLA} as follows.

```
# set the number of samples (T)
n.samples = 3000
# sample from the marginal latent distribution
samples_lambda = inla.rmarginal(n.samples, lambda7)
# sample from the likelihood model
predDist = rpois(n.samples, lambda = samples_lambda)
```



Introduced in ASM3, MLNN3

Note 26. Assume we wish to address the minimization problem

$$\hat{\theta} = \arg\min_{\theta} \left(C\left(\theta\right) \right) \tag{A.1}$$

for some cost function $C\left(\cdot\right)$.

Note 27. For instance, Proposition 1, it is $C\left(\theta\right) = -2\log\left(L\left(\theta\right)\right)$.

Note 28. Newton algorithm and Gradient descent algorithms are two optimization algorithms aiming to address the minimization problem (A.1). Each of them generate a convergence sequence $\left\{\theta^{(t)}\right\}$ to $\hat{\theta}$ as $\theta^{(t)} \to \hat{\theta}$ under regularity conditions (omitted here).

Algorithm 29. *Newton algorithm consist of the recursion*

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla_{\theta}^{2} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}\right]^{-1} \nabla_{\theta} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}$$

where $\nabla_{\theta} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}$ *is the gradient of* $C\left(\theta\right)$ *at* $\theta = \theta^{(t)}$, $\nabla_{\theta}^{2} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}$ *is the Hessian matrix of* $C\left(\theta\right)$ *at* $\theta = \theta^{(t)}$. *It requires a user specified seed* $\theta^{(0)}$. *The recursion stops when a termination criterion such as* $t \geq T_{\max}$, *for some user specified* $T_{\max} > 0$, *is satisfied.*

Algorithm 30. *Gradient descent algorithm consist of the recursion*

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t \, \nabla_{\theta} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}$$

where $\nabla_{\theta} C\left(\theta\right)\big|_{\theta=\theta^{(t)}}$ *is the gradient of* $C\left(\theta\right)$ *at* $\theta = \theta^{(t)}$. *It requires a user specified positive non-increasing sequence* $\{\eta_t\}$ *such as* $\eta_t = \sqrt{1/t}$, *and a user specified seed* $\theta^{(0)}$. *The recursion stops when a termination criterion such as* $t \geq T_{\max}$ *for some user-specified* $T_{\max} > 0$, *is satisfied.*

Example 31. Consider the marginal likelihood

$$f(x|a,b) = \left(\frac{1}{\Gamma(a) b^{a}}\right)^{n} \prod_{i=1}^{n} x_i^{a} e^{-n\bar{x}\frac{1}{b}}$$

where $a > 0$, $b > 0$. Write the Newton alg., and Gradient descent alg. recursions for to find $\theta^{*} = \arg\min_{\theta} \left(-\ell_n\left(\theta\right)\right)$ where $\ell_n\left(\theta\right) = \log f(x|\theta)$ and $\theta = (a, b)$.

Hint-1: Digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$

Hint-2: Trigamma function $\psi_1(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\log\Gamma(x)$

Hint-3: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Proof. Gradient descent's recursion is

$$\begin{bmatrix} a^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} a^{(t)} \\ b^{(t)} \end{bmatrix} + \eta_t \left.\nabla_\theta \ell_n(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}$$

for $\eta_t = \sqrt{1/t}$, where

$$\ell_n(\theta) = -n\log\Gamma(a) - na\log(b) - \frac{1}{b}\sum_{i=1}^n x_i + (a-1)\sum_{i=1}^n \log(x_i)$$

$$\nabla_\theta \ell_n(\theta) = \begin{bmatrix} -n\psi(a) - n\log(b) + \sum_{i=1}^n \log(x_i) \\ -n\frac{a}{b} + n\frac{1}{b^2}\bar{x} \end{bmatrix}, \text{ and } \nabla^2_\theta \ell_n(\theta) = -n\begin{bmatrix} \psi_1(a) & \frac{1}{b} \\ \frac{1}{b} & \frac{2\bar{x}-ab}{b^3} \end{bmatrix}$$

$$\nabla^2_\theta \ell_n(\theta) = -n\begin{bmatrix} \psi_1(a) & \frac{1}{b} \\ \frac{1}{b} & \frac{2\bar{x}-ab}{b^3} \end{bmatrix} = -\frac{1}{n}\frac{1}{\psi_1(a)\frac{2\bar{x}-ab}{b} - 1}\begin{bmatrix} \frac{2\bar{x}-ab}{b} & -b \\ -b & b^2\psi_1(a) \end{bmatrix}$$

$$\begin{bmatrix} a^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} a^{(t)} \\ b^{(t)} \end{bmatrix} + \eta_t \left.\nabla_\theta \ell_n(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}$$

Newton algorithm's recursion is

$$\begin{bmatrix} a^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} a^{(t)} \\ b^{(t)} \end{bmatrix} + \left[\left.\nabla^2_\theta C(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}\right]^{-1} \left.\nabla_\theta \ell_n(\theta)\right|_{\theta=\left(a^{(t)},b^{(t)}\right)}$$

where additionally

$$\nabla^2_\theta \ell_n(\theta) = -n\begin{bmatrix} \psi_1(a) & \frac{1}{b} \\ \frac{1}{b} & \frac{2\bar{x}-ab}{b^3} \end{bmatrix}; \text{ hence } \left[\nabla^2_\theta \ell_n(\theta)\right]^{-1} = -\frac{1}{n}\frac{1}{\psi_1(a)\frac{2\bar{x}-ab}{b} - 1}\begin{bmatrix} \frac{2\bar{x}-ab}{b} & -b \\ -b & b^2\psi_1(a) \end{bmatrix}$$

□

Appendix B. Gaussian approximation of a (posterior) distribution

Introduced in SI2

Note 32. A well known approximation of the posterior distribution is the Gaussian posterior approximation.

Theorem 33. *The posterior density $pr(\theta|z_{1:n})$ of θ given n observables $z_{1:n}$ can be approximated by a multivariate Gaussian distribution density $pr_G(\theta|\mu_n, \Sigma_n)$ with mean μ_n being the mode i.e. $\frac{\partial}{\partial\theta_i}\log\left(pr(\theta|z_{1:n})\right)\Big|_{\theta=\mu_n} = 0$, and with covariance matrix $\Sigma_n > 0$* being the inverse Hessian at the mode i.e. $\Sigma_n = \left(H_{pr}(\mu_n)\right)^{-1}$ *where* $\left[H_{pr}(\mu_n)\right]_{i,j} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\left(pr(\theta|z_{1:n})\right)\Big|_{\theta=\mu_n}$.

Example 34. Consider a Bayesian model with sampling distribution $x_i|\theta \overset{\text{iid}}{\sim} \text{pr}(x_i|\theta) \propto \theta^{x_i}(1-\theta)^{x_i-1}$ and prior $\theta \sim \text{pr}(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. Find the Gaussian approximation of the posterior $\text{pr}(\theta|x)$ of θ given $x = (x_1, ..., x_n)$.

Solution. The log posterior density is

$$\log\left(\text{pr}\left(\theta|x\right)\right) = \left(a_n - 1\right)\log(\theta) + \left(b_n - 1\right)\log(1-\theta)$$

where $a_n = a + n\bar{x}$, and $b_n = b + n - n\bar{x}$. So

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\log\left(\text{pr}\left(\theta|x\right)\right)\Big|_{\theta=\mu_n} = \frac{a_n - 1}{\theta} - \frac{b_n - 1}{1-\theta}\Big|_{\theta=\mu_n} \implies \mu_n = \frac{a_n - 1}{a_n + b_n - 2}$$

$$\Sigma_n = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log\left(\text{pr}\left(\theta|x\right)\right)\Big|_{\theta=\mu_n} = \frac{a_n - 1}{\theta^2} - \frac{b_n - 1}{(1-\theta)^2}\Big|_{\theta=\mu_n} \implies \Sigma_n = \frac{(a_n - 1)(b_n - 1)}{(a_n + b_n - 2)^3}$$

Therefore, θ has asymptotic posterior density is that of $\text{N}\left(\mu_n, \Sigma_n\right)$; i.e. $\text{pr}\left(\theta|x\right) \approx \text{N}\left(\theta|\mu_n, \Sigma_n\right)$.