INLA - Introduction

Elias T. Krainski
eliaskr@ufpr.br

62ª RBras & 17º SEAGRO,
Jul–2017, Lavras/MG

Outline

Outline

A model for Tokyo data

Observation model

$$y_i \sim \text{Binomial}(n_i, p_i)$$

$$p_i = \frac{1}{1 + \exp(-x_i)}$$

the likelihood has no $\boldsymbol{\theta}$

$$\pi(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{366} \pi(y_i|x_i)$$

Latent model

$$
\begin{aligned}
\pi(\boldsymbol{x}|\boldsymbol{\theta}) &\propto \exp\left\{-\frac{\theta}{2}\left[(x_1 - x_{366})^2 + \sum_{i=2}^{366}(x_i - x_{i-1})^2\right]\right\} && (1)\\
&= \exp\left\{-\frac{\theta}{2}\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{x}\right\} && (2)
\end{aligned}
$$

Latent model

$$
\begin{aligned}
\pi(\boldsymbol{x}|\boldsymbol{\theta}) &\propto \exp\left\{-\frac{\theta}{2}\left[(x_1 - x_{366})^2 + \sum_{i=2}^{366}(x_i - x_{i-1})^2\right]\right\} && (1) \\
&= \exp\left\{-\frac{\theta}{2}\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{x}\right\} && (2)
\end{aligned}
$$

where $\boldsymbol{R} = \begin{pmatrix} 2 & \text{-1} & & & & & & \text{-1} \\ \text{-1} & 2 & \text{-1} & & & & & \\ & \text{-1} & 2 & \text{-1} & & & & \\ & & & & \ddots & & & \\ & & & & & \text{-1} & 2 & \text{-1} \\ & & & & & & \text{-1} & 2 & \text{-1} \\ \text{-1} & & & & & & \text{-1} & 2 \end{pmatrix}$

$\boldsymbol{Q}(\boldsymbol{\theta}) = \theta\boldsymbol{R}$

Latent model warning

$$\exp\left\{-\frac{\theta}{2}\left[(x_1 - x_{366})^2 + \sum_{i=2}^{366}(x_i - x_{i-1})^2\right]\right\} \tag{3}$$

$$\tag{4}$$

intrinsic/improper

$$\begin{array}{rrcl} x_i = 20, & x_{i-1} = 10 & \rightarrow & x_i - x_{i-1} = 10 \\ x_i = 10020, & x_{i-1} = 10010 & \rightarrow & x_i - x_{i-1} = 10 \end{array}$$

constraint or take the intercept out

$\pi(\boldsymbol{\theta})$ problem

- Tokyo example: $Q(\boldsymbol{\theta}) = \theta \boldsymbol{R}$
  - bigger θ less variation of $\boldsymbol{x}$
    - related to the variation of p_i
- $\theta > 0$: people usually use $\theta \sim \text{Gamma}(a, b)$
- improper distribution: θ values depends on $\boldsymbol{R}$
  - hard to interpret θ (a=?????, b=?????)

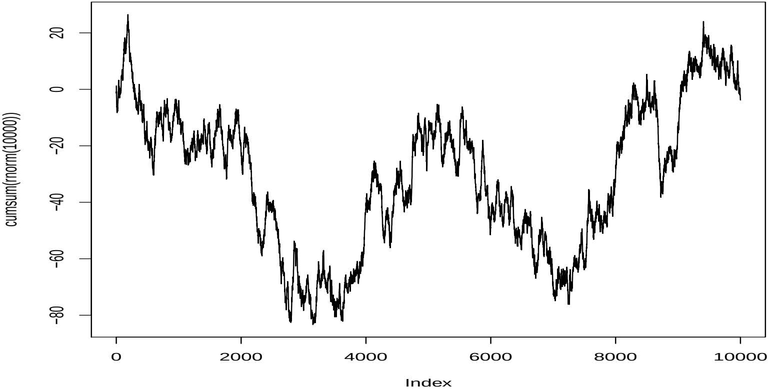
$\pi(\boldsymbol{x}|\theta = 1)$ and n

The marginal variance and n relation

```
rw.var <- function(n, order) {
    R <- as.matrix(INLA::inla.rw(n, order=order))
    mean(diag(INLA::inla.ginv(R, rankdef=order)))
}
n <- c(10, 100, 366, 1000); names(n) <- n
rbind(rw1=sapply(n, rw.var, order=1),
      rw2=sapply(n, rw.var, order=2))

##       10      100          366         1000
## rw1 1.65   16.665     60.99954     166.6665
## rw2 2.40 2381.190 116733.95702 2380955.1304
```

$\pi(\boldsymbol{x}|\theta = 1)$: one realization



We need to control the marginal variance!

$\pi(\boldsymbol{\theta})$ solution

1. scale the model $\rightarrow$ easy to interpret θ
   - Tutorial on `scale.option` at `www.r-inla.org/`

$\pi(\boldsymbol{\theta})$ solution

1. scale the model $\rightarrow$ easy to interpret θ
   - Tutorial on `scale.option` at `www.r-inla.org/`
2. AND (new idea) Penalized complexity prior
   - P0: basic model: $p_i = p_0$
   - P1: complex model: p_i varies
   - Kullback-Leibler divergence (KLD)
     - a distance from P1 model to P0, KLD(P0/P0) = 0
   - allow variation on p_i
   - AND supports the basic model
     - Gamma(a, b) always overfits

Outline

On our Bayesian hierarchical model

- Inference on (what we know about) $\boldsymbol{\theta}$ and $\boldsymbol{x}$ given $\boldsymbol{y}$
  - in maths: $\pi(\boldsymbol{x}|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
- considering $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$

On our Bayesian hierarchical model

- Inference on (what we know about) $\boldsymbol{\theta}$ and $\boldsymbol{x}$ given $\boldsymbol{y}$
  - in maths: $\pi(\boldsymbol{x}|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
- considering $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$
- using the Bayes theorem,

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

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

On our Bayesian hierarchical model

- Inference on (what we know about) $\boldsymbol{\theta}$ and $\boldsymbol{x}$ given $\boldsymbol{y}$
  - in maths: $\pi(\boldsymbol{x}|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
- considering $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$
- using the Bayes theorem,

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

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

- even more...
  - $\pi(\theta_j|\boldsymbol{y})$, $j = 1, ..., \dim(\boldsymbol{\theta})$
  - $\pi(x_i|\boldsymbol{y})$, $i = 1, ..., \dim(\boldsymbol{x})$

The inference problem

- we have to compute

$$\pi(x_i|\boldsymbol{y}) \propto \int_{x_{\{-i\}}} \int_{\boldsymbol{\theta}} \pi(y|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{x}_{\{-i\}}$$

The inference problem

- we have to compute

$$\pi(x_i|\boldsymbol{y}) \propto \int_{x_{\{-i\}}} \int_{\boldsymbol{\theta}} \pi(y|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{x}_{\{-i\}}$$

and

$$\pi(\theta_j|\boldsymbol{y}) \propto \int_{x} \int_{\boldsymbol{\theta}_{\{-j\}}} \pi(y|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{\{-j\}}\, d\boldsymbol{x}$$

The inference problem

- we have to compute

$$\pi(x_i|\boldsymbol{y}) \propto \int_{x_{\{-i\}}} \int_{\boldsymbol{\theta}} \pi(y|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{x}_{\{-i\}}$$

  and

$$\pi(\theta_j|\boldsymbol{y}) \propto \int_{x} \int_{\boldsymbol{\theta}_{\{-j\}}} \pi(y|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{\{-j\}}\, d\boldsymbol{x}$$

- remember
  - dim($\boldsymbol{\theta}$) is small
  - dim($\boldsymbol{x}$) is not small

  - we have to compute very high dimensional integrals

The inference problem

- we have to compute

$$\pi(x_i|\boldsymbol{y}) \propto \int_{x_{\{-i\}}} \int_{\boldsymbol{\theta}} \pi(y|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{x}_{\{-i\}}$$

  and

$$\pi(\theta_j|\boldsymbol{y}) \propto \int_{x} \int_{\boldsymbol{\theta}_{\{-j\}}} \pi(y|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{\{-j\}}\, d\boldsymbol{x}$$

- remember
  - $\dim(\boldsymbol{\theta})$ is small
  - $\dim(\boldsymbol{x})$ is not small

  - we have to compute very high dimensional integrals
- typically they are not analytically tractable
  - $\rightarrow$ we have to approach

- single-site: compute (the expressions) for
  - $p(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{x}, \boldsymbol{y})$
  - $p(x_i|\boldsymbol{x}_{-i}, \boldsymbol{\theta}, \boldsymbol{y})$

using MCMC

- single-site: compute (the expressions) for
  - $p(\theta_j | \boldsymbol{\theta}_{-j}, \boldsymbol{x}, \boldsymbol{y})$
  - $p(x_i | \boldsymbol{x}_{-i}, \boldsymbol{\theta}, \boldsymbol{y})$
- draw samples from such conditionals
  - WinBUGS, OpenBUGS, JAGS, and others
- use these samples to summarize $p(\boldsymbol{x})$ and $p(\boldsymbol{\theta})$

using MCMC

- single-site: compute (the expressions) for
  - $p(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{x}, \boldsymbol{y})$
  - $p(x_i|\boldsymbol{x}_{-i}, \boldsymbol{\theta}, \boldsymbol{y})$
- draw samples from such conditionals
  - WinBUGS, OpenBUGS, JAGS, and others
- use these samples to summarize $p(\boldsymbol{x})$ and $p(\boldsymbol{\theta})$
- warning
  - sampling from $x_i|\boldsymbol{x}_{-i}, \boldsymbol{\theta}, y$
    - slow convergence when strong dependence
    - does not works for our example...
  - better: draw joint sample from $\boldsymbol{x}|\boldsymbol{\theta}, y$
  - best: use INLA

Outline

What INLA does

- INLA does:
  - compute marginals of $\pi(x_i|\boldsymbol{y})$ and $\pi(\theta_j|\boldsymbol{y})$
- how?
  - approach $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ to approach $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
  - explore $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
    - approach $\pi(\theta_j|\boldsymbol{y})$
  - approach $\pi(x_i|\boldsymbol{x}_{-i})$

Important ingredient

The GMRF-approximation

$$\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) \quad \propto \quad \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_i \log \pi(y_i|x_i)\right)$$

Important ingredient

The GMRF-approximation

$$
\begin{aligned}
\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}\boldsymbol{x} + \sum_i \log \pi(y_i|x_i)\right) \\
&\approx \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{Q}+\mathrm{diag}(\boldsymbol{c}))(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \pi_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})
\end{aligned}
$$

$c_i = -\frac{dl_i^2}{dx_i^2}$ where $l_i = \log(\pi(y_i|x_i))$, $i = 1, ..., \#$ data

Important ingredient

The GMRF-approximation

$$
\begin{aligned}
\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \sum_i \log \pi(y_i|x_i)\right) \\
&\approx \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{Q}+\mathrm{diag}(\boldsymbol{c}))(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \pi_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})
\end{aligned}
$$

$c_i = -\frac{dl_i^2}{dx_i^2}$ where $l_i = \log(\pi(y_i|x_i))$, $i = 1, ..., \#$ data

- Markov and computational properties (on $\boldsymbol{Q}$) are preserved

Important ingredient

The GMRF-approximation

$$\begin{aligned}
\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_i \log\pi(y_i|x_i)\right) \\
&\approx \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{Q}+\mathrm{diag}(\boldsymbol{c}))(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= \pi_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})
\end{aligned}$$

$c_i = -\frac{dl_i^2}{dx_i^2}$ where $l_i = \log(\pi(y_i|x_i))$, $i = 1, ..., \#$ data

- Markov and computational properties (on $\boldsymbol{Q}$) are preserved
- $\widetilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ costs
  - temporal: $O(n)$
  - spatial: $O(n\log(n))$

If $\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}$ is *Gaussian*, the "approximation" is exact.

INLA, $\pi(\theta_j|\boldsymbol{y})$

- Considering

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}$$

INLA, $\pi(\theta_j|\boldsymbol{y})$

- Considering

$$
\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\
&\propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}
\end{aligned}
$$

INLA, $\pi(\theta_j|\boldsymbol{y})$

- Considering

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\ &\propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\end{aligned}$$

- Gaussian approximation to denominator

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\pi_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})}$$

INLA, $\pi(\theta_j|\boldsymbol{y})$

- Considering

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \\ &\propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \end{aligned}$$

- Gaussian approximation to denominator

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\pi_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})}$$

- mode of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ (optimization)
  - explore $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$
    - approach $\pi(\theta_j|\boldsymbol{y})$ (numerical integration)

INLA, $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$

Approaching $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$

- Problem
  - dim($\boldsymbol{x}$)=n is not small
  - n marginals to compute
- Laplace approximation

$$\widetilde{\pi}(x_i \mid \boldsymbol{y},\boldsymbol{\theta}) \approx \left.\frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\widetilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i,\boldsymbol{y},\boldsymbol{\theta})}\right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})}$$

INLA, $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$

Approaching $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$

- Problem
  - dim($\boldsymbol{x}$)=n is not small
  - n marginals to compute
- Laplace approximation

$$\widetilde{\pi}(x_i \mid \boldsymbol{y},\boldsymbol{\theta}) \approx \left.\frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\widetilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i,\boldsymbol{y},\boldsymbol{\theta})}\right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*_{-i}(x_i,\boldsymbol{\theta})}$$

- simpler/cruder (fast) approximation (from $\pi_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$)

$$\hat{\pi}(x_i|\boldsymbol{y},\boldsymbol{\theta}) = N(x_i;\mu_i(\boldsymbol{\theta}),\sigma_i^2(\boldsymbol{\theta}))$$

INLA, $\pi(x_i|\boldsymbol{y})$

Approaching $\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})$

- integrate $\boldsymbol{\theta}$ out from $\widetilde{\pi}(x_i \mid \boldsymbol{y}, \boldsymbol{\theta})$
- select values for $\boldsymbol{\theta}$
- use weighted sum

$$\widetilde{\pi}(x_i \mid \boldsymbol{y}) \propto \sum_j \widetilde{\pi}(x_i|\boldsymbol{y}, \boldsymbol{\theta}_j) \times \widetilde{\pi}(\boldsymbol{\theta}_j|\boldsymbol{y})$$

Remarks

1. Expect $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ to be accurate, since
    - $\boldsymbol{x}|\boldsymbol{\theta}$ is *a priori* Gaussian
    - Likelihood models are 'well-behaved' so

      $$\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$$

      is *almost* Gaussian.
2. There are no distributional assumptions on $\boldsymbol{\theta}|\boldsymbol{y}$
3. Similar remarks are valid to

$$\widetilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$$

How can we assess the error in the approximations?

Tool 1: Compare a sequence of improved approximations

1. Gaussian approximation
2. Simplified Laplace
3. Laplace

No big differences $\rightarrow$ good approximation

How can we assess the error in the approximations?

Tool 2: Estimate the "effective" number of parameters as defined in the Deviance Information Criteria:

$$p_{\mathrm{D}}(\boldsymbol{\theta}) = \overline{D}(\boldsymbol{x}; \boldsymbol{\theta}) - D(\overline{\boldsymbol{x}}; \boldsymbol{\theta})$$

and compare this with the number of observations
Low ratio is good.
This criteria has theoretical justification.