VB for INLA

King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية

May 2022

Outline

1. Introduction to INLA
2. Improved mean Gaussian approximation based on VB
3. Real examples
4. Current work

Model definition

Suppose we have response data $\boldsymbol{y}_{n\times 1}$ with density function $\pi(y|\mathcal{X},\boldsymbol{\theta})$ and link function $h(.)$, that is linked to some covariates $\boldsymbol{Z}=\{\boldsymbol{X},\boldsymbol{U}\}$ through linear predictors

$$\boldsymbol{\eta}_{n\times 1}=\beta_0\mathbf{1}+\boldsymbol{\beta}\boldsymbol{X}+\sum_{k=1}^{K}f^k(\boldsymbol{u}_k)$$

The inferential aim is to estimate the latent field $\mathcal{X}_{m_*\times 1}=\{\beta_0,\boldsymbol{\beta},\boldsymbol{f}\}$.
Define the augmented latent field

$$\mathcal{X}_{m\times 1}=\{\boldsymbol{\eta},\beta_0,\boldsymbol{\beta},\boldsymbol{f}\}.$$

Posterior approximations

$$
\begin{aligned}
\pi(\mathcal{X},\boldsymbol{\theta},\boldsymbol{y}) &= \pi(\boldsymbol{\theta})\pi(\mathcal{X}|\boldsymbol{\theta})\prod_{i=1}^{n}\pi(y_i|\mathcal{X}_i,\boldsymbol{\theta}) \\
\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \left.\frac{\pi(\mathcal{X},\boldsymbol{\theta},\boldsymbol{y})}{\pi_G(\mathcal{X}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\mathcal{X}=\boldsymbol{\mu}(\boldsymbol{\theta})} \\
\tilde{\pi}(\theta_j|\boldsymbol{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \\
\tilde{\pi}(\mathcal{X}_j|\boldsymbol{y}) &= \int \tilde{\pi}(\mathcal{X}_j|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},
\end{aligned}
$$

$\tilde{\pi}(\mathcal{X}_j|\boldsymbol{\theta},\boldsymbol{y})$ depends on the approximation used, for Gaussian it is straightforward for the Laplace approximation we do another Gaussian approximation to $\tilde{\pi}(\boldsymbol{\mathcal{X}}_{-j}|\boldsymbol{\theta},\boldsymbol{y})$.

Thoughts and ideas

- Can we improve $\pi_G(\mathcal{X}|\boldsymbol{\theta}, \boldsymbol{y})$ so we have cheap but good marginals $\tilde{\pi}(\mathcal{X}_j|\boldsymbol{y})$.
- For large data, $\boldsymbol{\mathcal{X}}$ is large - can we remove $\boldsymbol{\eta}$, but still produce cheap and accurate inference for $\boldsymbol{\eta}$.

Gaussian approximation

Laplace method - fit the best Gaussian at the mode of a curve where the variance is derived from the inverse Hessian at the mode.

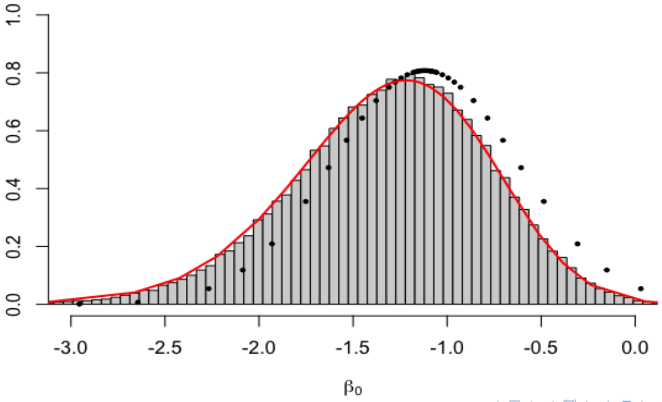
Mode and Hessian at the mode

$$\mathcal{X}|\boldsymbol{\theta},\boldsymbol{y} \sim N(\boldsymbol{\mu},(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))^{-1}) \tag{1}$$

with $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$ from a second-order expansion of $\pi(y_i|\mathcal{X}_i,\boldsymbol{\theta}) \approx a_i + b_i\mathcal{X}_i + c_i^2\mathcal{X}_i$.
Can we do better than this?

Poisson example

$$Y_i|\beta_0 = -1, \beta_1 = -0.5 \quad \sim \quad \text{Poisson}(\exp(\beta_0 + \beta_1 X_i))$$



Posterior approximations

$$
\begin{aligned}
\pi(\mathcal{X}, \boldsymbol{\theta}, \boldsymbol{y}) &= \pi(\boldsymbol{\theta})\pi(\mathcal{X}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \\
\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \left. \frac{\pi(\mathcal{X}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi_G(\mathcal{X}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\mathcal{X}=\boldsymbol{\mu}(\boldsymbol{\theta})} \\
\tilde{\pi}_G(\mathcal{X}_j|\boldsymbol{y}) &= \int \tilde{\pi}_G(\mathcal{X}_j|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \\
\tilde{\pi}_L(\mathcal{X}_j|\boldsymbol{y}) &= \int \tilde{\pi}_L(\mathcal{X}_j|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},
\end{aligned}
$$

with $\tilde{\pi}_L(\mathcal{X}_j|\boldsymbol{\theta}, \boldsymbol{y}) = \left. \frac{\pi(\mathcal{X}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi_G(\mathcal{X}_{-j}|\mathcal{X}_j, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\mathcal{X}_{-j}=\boldsymbol{\mu}_{-j}(\boldsymbol{\theta})}$

Better Gaussian approximation

Can we find a better Gaussian approximation (closer to LA) that is cheap and will scale well?

VB from Zellner (1988)

Based on prior information $\mathcal{I}$, data $\boldsymbol{y}$ and parameters $\boldsymbol{\theta}$, define the following:

1. $\pi(\boldsymbol{\theta}|\mathcal{I})$ is the prior model
2. $q(\boldsymbol{\theta}|\mathcal{D})$ is the learned model from the prior information and the data where $\mathcal{D} = \{\mathcal{I}, \boldsymbol{y}\}$
3. $l(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood
4. $p(\boldsymbol{y}|\mathcal{I})$ is the marginal model for the data where $p(\boldsymbol{y}|\mathcal{I}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathcal{I})d\boldsymbol{\theta}$

The input information in the learning of $\boldsymbol{\theta}$ is given by $\pi(\boldsymbol{\theta}|\mathcal{I})$ and $l(\boldsymbol{\theta}|\boldsymbol{y})$. An information processing rule (IPR) then delivers $q(\boldsymbol{\theta}|\mathcal{D})$ and $p(\boldsymbol{y}|\mathcal{I})$ as output information.

Bayes Rule as an efficient IPR

A stable and efficient IPR would provide the same amount of output information than received through the input information, thus being information conservative. Thus, we learn $q(\boldsymbol{\theta}|\mathcal{D})$ such that it minimizes

$$\begin{aligned} & - \int \left[\log \pi(\boldsymbol{\theta}|\mathcal{I}) + \log l(\boldsymbol{\theta}|\boldsymbol{y})\right] q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} + \int \left[\log q(\boldsymbol{\theta}|\mathcal{D}) + \log p(\boldsymbol{y}|\mathcal{I})\right] q(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\ = \;& E_{q(\boldsymbol{\theta}|\mathcal{D})}\left[- \log l(\boldsymbol{\theta}|\boldsymbol{y})\right] + \text{KLD}\left[q(\boldsymbol{\theta}|\mathcal{D}) || \pi(\boldsymbol{\theta}|\mathcal{I})\right]. \end{aligned} \tag{2}$$

Variational form of Bayes' theorem

Finding the best fit from a certain family $P(\Theta)$,

$$\arg \min_{p \in P(\Theta)} \left(\mathrm{E}_{p(\boldsymbol{\theta})} \left[- \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right) \tag{3}$$

How can we use this?

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{\delta}$.

$$\arg_{\boldsymbol{\delta}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu} + \boldsymbol{\delta}, \boldsymbol{Q}^{-1})$.
But $\mathcal{X}$ can be very large...

How can we use this?

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{\delta}$.

$$\arg_{\boldsymbol{\delta}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y}, \theta \sim N(\boldsymbol{\mu} + \boldsymbol{\delta}, \boldsymbol{Q}^{-1})$.
But $\mathcal{X}$ can be very large...

How can we use this?

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{\delta}$.

$$\arg_{\boldsymbol{\delta}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu} + \boldsymbol{\delta}, \boldsymbol{Q}^{-1})$.
But $\mathcal{X}$ can be very large...

How can we use this?

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{\delta}$.

$$\arg_{\boldsymbol{\delta}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu} + \boldsymbol{\delta}, \boldsymbol{Q}^{-1})$.
But $\mathcal{X}$ can be very large...

Implicit mean correction

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu}^* = \boldsymbol{b} + \boldsymbol{\lambda}$, so that

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{M}\boldsymbol{\lambda}$$

So now we solve for,

$$\arg_{\boldsymbol{\lambda}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}^*, \boldsymbol{Q}^{-1})$.
Low-rank correction $\rightarrow$ Only correct some $b's$, change to all $\mu's$.

Implicit mean correction

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu}^* = \boldsymbol{b} + \boldsymbol{\lambda}$, so that

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{M}\boldsymbol{\lambda}$$

So now we solve for,

$$\arg_{\boldsymbol{\lambda}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta} \sim N(\boldsymbol{\mu}^*, \boldsymbol{Q}^{-1})$.
Low-rank correction $\rightarrow$ Only correct some $b's$, change to all $\mu's$.

Implicit mean correction

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu}^* = \boldsymbol{b} + \boldsymbol{\lambda}$, so that

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{M}\boldsymbol{\lambda}$$

So now we solve for,

$$\arg_{\boldsymbol{\lambda}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta} \sim N(\boldsymbol{\mu}^*, \boldsymbol{Q}^{-1})$.
Low-rank correction $\rightarrow$ Only correct some b's, change to all μ's.

Implicit mean correction

Recall that $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$.
Now let's formulate $(\boldsymbol{Q}_\pi + diag(\boldsymbol{c}))\boldsymbol{\mu}^* = \boldsymbol{b} + \boldsymbol{\lambda}$, so that

$$\boldsymbol{\mu}^* = \boldsymbol{\mu} + \boldsymbol{M}\boldsymbol{\lambda}$$

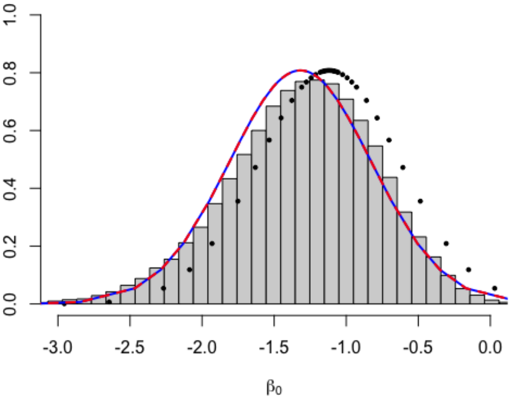
So now we solve for,

$$\arg_{\boldsymbol{\lambda}} \min_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left(\mathrm{E}_{p(\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta})} \left[-\sum_{i=1}^{n} \log f(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \right] + \mathrm{KLD}(p||\pi) \right)$$

where $\mathcal{X}|\boldsymbol{y}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}^*, \boldsymbol{Q}^{-1})$.
Low-rank correction $\rightarrow$ Only correct some $b's$, change to all $\mu's$.

Example (small data)

$$Y_i|\beta_0 = -1, \beta_1 = -0.5 \quad \sim \quad \text{Poisson}(\exp(\beta_0 + \beta_1 X_i))$$



Results of mean correction

	GA	INLA	VBC	MCMC
β_0	-1.119	-1.317	-1.316	-1.302
β_1	-0.361	-0.401	-0.401	-0.391

Table: Posterior means from the Gaussian method, INLA, VBC and MCMC

Example (large data)

$$Y_i|\beta_0 = -1, \beta_1 = -0.5 \quad \sim \quad \text{Poisson}(\exp(\beta_0 + \beta_1 X_i))$$

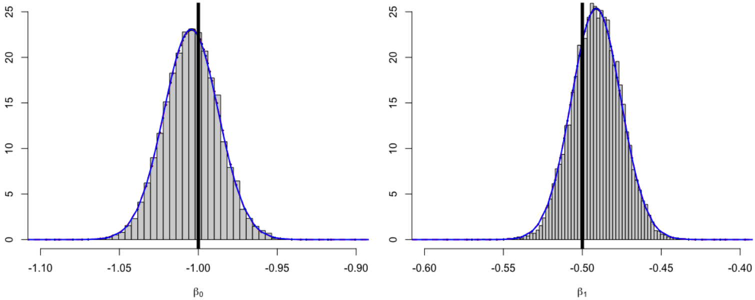


Figure: Marginal posterior of β_0 (center) and β_1 (right) from the Laplace method (points), VBC (solid line) and INLA (broken line) approximations

Results

	GA	INLA	VBC	MCMC
β_0	-1.004	-1.005	-1.005	-1.005
β_1	-0.491	-0.492	-0.492	-0.492
Time (s)	6.346	37.168	7.344	5779.42

Table: Posterior means from the Laplace method, INLA, VBC and MCMC

Tokyo example

The Tokyo dataset in the R INLA library contains information on the number of times the daily rainfall measurements in Tokyo was more than $1mm$ on a specific day t for two consecutive years. In order to model the annual rainfall pattern, a stochastic spline model with fixed precision is used to smooth the data.

$$
\begin{aligned}
y_i|\mathcal{X} &\sim Bin\left(n_i, p_i = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right) \\
(\alpha_{i+1} - 2\alpha_i + \alpha_{i-1})|\tau &\overset{\text{iid}}{\sim} N(0, \tau^{-1}),
\end{aligned}
$$

where $i = 1, 2, ..., 366$ on a torus, and $n_{60} = 1$ else $n_i = 2$.

Results

The mean of the absolute errors produced between the Laplace method and INLA is 0.0358 while for VBC it is 0.0009.

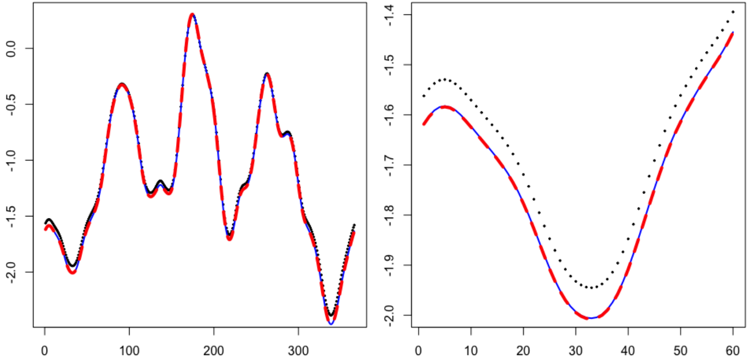
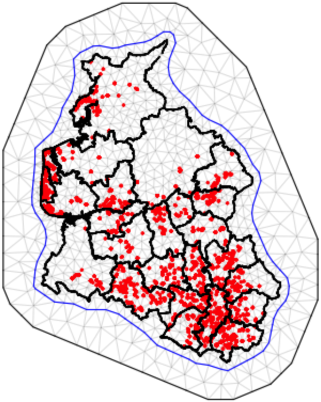


Figure: Posterior mean of α (left) (zoomed for the first two months (right)) from the Laplace method (points), VBC (solid line) and INLA (broken line)

Spatial survival example

Consider the R dataset `Leuk` that features the survival times of 1043 patients with acute myeloid leukemia (AML) in Northwest England between 1982 to 1998.



Cox spatial model

$$h(t, \boldsymbol{s}) = h_0(t) \exp(\beta \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{s})),$$

with

$$\eta_i(s) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{TPI}_i + u(s).$$

which implies a latent field of size $m = 39158$.

Results

	GA	INLA	VBC
β_0	-5.935	-6.312	-6.312
β_1	1.050	1.079	1.079
β_2	0.313	0.319	0.319
β_3	0.198	0.200	0.200
Time(s)	25.9	1276	26.3

Table: Posterior means from the Laplace method, INLA and VBC - all fixed effects are significant

Results

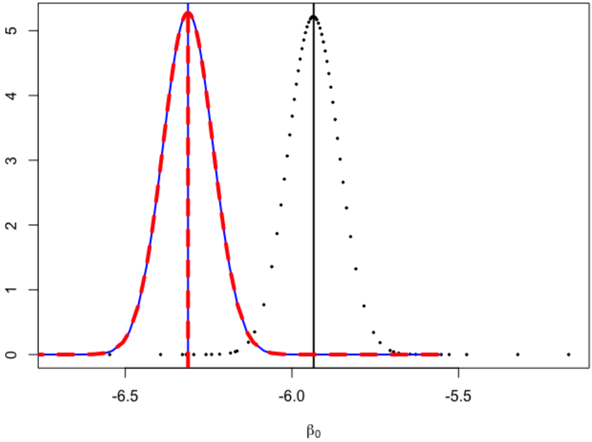


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Results

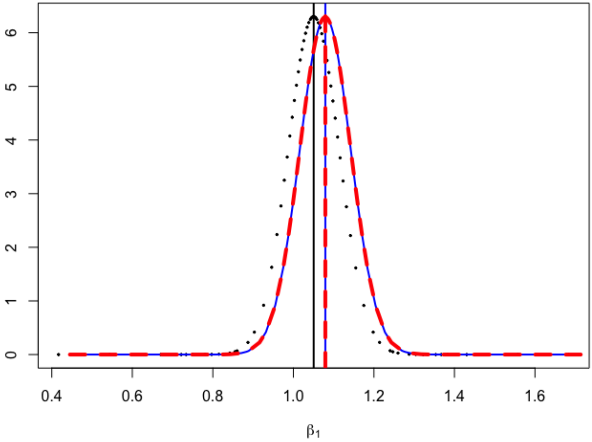


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Results

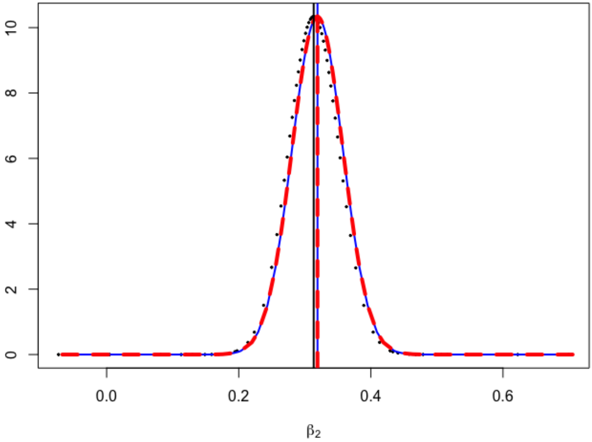


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Results

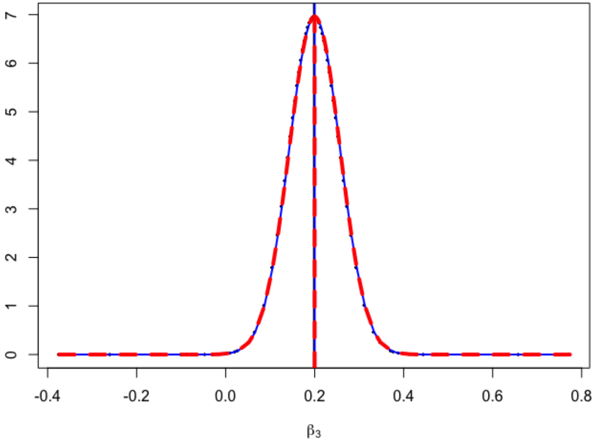


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Current work

- VB correction for the variance of the latent field

VB mean correction paper: `https://arxiv.org/abs/2111.12945`

Thank you • شكرا

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology