



RESEARCH ARTICLE · SIMULATION STUDY · FOUNDING ARTICLE

Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA)

Aliaksandr Hubin^{1*} and Geir Storvik²

¹ Norwegian University of Life Sciences, Ås, Norway and Integreat, University of Oslo, Oslo, Norway

² Department of Mathematics, University of Oslo, Oslo, Norway and Integreat, University of Oslo, Oslo, Norway

*Corresponding author: aliaksandr.hubin@nmbu.no

ARTICLE INFO

Keywords

Integrated nested Laplace approximation
Marginal likelihood
Model Evidence
Bayes Factor
Markov chain Monte Carlo
Numerical Integration

History

Received 14 Oct 2026
Final version 26 Oct 2026
Accepted 23 Oct 2026
Version 1

Handling editor

Matthias Templ

ABSTRACT

The marginal likelihood is a well established model selection criterion in Bayesian statistics. It also allows to efficiently calculate the marginal posterior model probabilities that can be used for Bayesian model averaging of quantities of interest. For many complex models, including latent modeling approaches, marginal likelihoods are however difficult to compute. One recent promising approach for approximating the marginal likelihood is Integrated Nested Laplace Approximation (INLA), designed for models with latent Gaussian structures. In this study we compare the approximations obtained with INLA to alternative approaches across a progression of models of varying complexity. In particular, we evaluate a simple linear latent model with exact analytical truth, a Bayesian linear regression model, probit and logit Bayesian regression models, a longitudinal Poisson generalized linear mixed model, and two extra simulation scenarios designed to probe the limits of the latent Gaussian approximation.

✓ **Reproduced** verified 2026-10-23

REPRODUCIBILITY RECORD

Result	Reported results reproduced. Reported values reproduce under INLA 25.06.07, Table 5 within 0.10 log units.	✓ measured
Entry point	cd code/INLA-marginal-likelihood-ddb7377399d94d012f1d4604db91ef8972dbeea7 && Rscript run_all.R	✓ executed
Environment	R 4.5.2, INLA 25.06.07, MCMCpack 1.7.1, rstan 2.32.7, bridgesampling 1.2.1, macOS arm64, one core, no network	✓ measured
Determinism	Seeded.	declared
Archive	estimating-the-marginal-likelihood-with-integrated-nested-laplace-approximation-inla-v1-repro.zip · reviewed bundle SHA-256 cd56e4bf0cb1b76b9600ecdde9a034fa6813a9266c15cd0cd0a379f1e4a118	✓ computed
Code	https://github.com/aliaksah/INLA-marginal-likelihood	declared
Data	Open.	declared, route checked

1 Introduction

Marginal likelihoods have been commonly accepted to be an extremely important quantity within Bayesian statistics. For data $\mathbf{y}$ and model $\mathcal{M}$, which includes some unknown parameters θ , the marginal likelihood is given by

$$p(\mathbf{y}|\mathcal{M}) = \int_{\Omega_{\theta}} p(\mathbf{y}|\mathcal{M}, \theta) p(\theta|\mathcal{M}) d\theta \quad (1)$$

CITE AS Hubin, A., & Storvik, G. (2026). Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA). Journal of Reproducible Statistics, 1, 9. <https://doi.org/10.68365/jrepstat.2026.9>

where $p(\theta|\mathcal{M})$ is the prior for θ under model $\mathcal{M}$ while $p(\mathbf{y}|\mathcal{M}, \theta)$ is the likelihood function conditional on θ . Consider first the problem of comparing models $\mathcal{M}_i$ and $\mathcal{M}_j$ through the ratio between their posterior probabilities

$$\frac{p(\mathcal{M}_i|\mathbf{y})}{p(\mathcal{M}_j|\mathbf{y})} = \frac{p(\mathbf{y}|\mathcal{M}_i)}{p(\mathbf{y}|\mathcal{M}_j)} \times \frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}. \quad (2)$$

The first term of the right hand side is the Bayes Factor (Kass & Raftery, 1995). In this way one usually performs Bayesian model selection with respect to the posterior marginal model probabilities without the need to calculate them explicitly. However if we are interested in Bayesian model averaging and marginalizing some quantity Δ over the given set of models $\Omega_{\mathcal{M}}$ we are calculating the posterior marginal distribution, which in our notation becomes

$$p(\Delta|\mathbf{y}) = \sum_{\mathcal{M} \in \Omega_{\mathcal{M}}} p(\Delta|\mathcal{M}, \mathbf{y}) p(\mathcal{M}|\mathbf{y}). \quad (3)$$

Here, $p(\mathcal{M}|\mathbf{y})$ is the posterior marginal model probability for model $\mathcal{M}$ that can be calculated with respect to Bayes theorem as

$$p(\mathcal{M}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M})p(\mathcal{M})}{\sum_{\mathcal{M}' \in \Omega_{\mathcal{M}}} p(\mathbf{y}|\mathcal{M}')p(\mathcal{M}')}. \quad (4)$$

Thus, one requires marginal likelihoods $p(\mathbf{y}|\mathcal{M})$ in (2), (3) and (4). Metropolis-Hastings algorithms searching through models within a Monte Carlo setting (e.g. Hubin & Storvik, 2018) requires acceptance ratios of the form

$$r_m(\mathcal{M}, \mathcal{M}^*) = \min \left\{ 1, \frac{p(\mathbf{y}|\mathcal{M}^*)p(\mathcal{M}^*)q(\mathcal{M}|\mathcal{M}^*)}{p(\mathbf{y}|\mathcal{M})p(\mathcal{M})q(\mathcal{M}^*|\mathcal{M})} \right\}, \quad (5)$$

also involving the marginal likelihoods. All these examples show the fundamental importance of being able to calculate marginal likelihoods $p(\mathbf{y}|\mathcal{M})$ in Bayesian statistics.

Unfortunately, for most of the models that include both unknown parameters θ and some latent variables $\boldsymbol{\eta}$ analytical calculation of $p(\mathbf{y}|\mathcal{M})$ is impossible. In such situations one must use approximative methods that hopefully are accurate enough to neglect the approximation errors involved. Different approximative approaches have been mentioned in various settings of Bayesian variable selection and Bayesian model averaging. Laplace's method (Tierney & Kadane, 1986) has been widely used, but it is based on rather strong assumptions. The Harmonic mean estimator (Newton & Raftery, 1994) is an easy to implement MCMC based method, but it can give high variability in the estimates. Chib's method (Chib, 1995), and its extension (Chib & Jeliazkov, 2001), are also MCMC based approaches that have gained increasing popularity. They can be very accurate provided enough MCMC iterations are performed, but need to be adopted to each application and the specific algorithm used. Approximate Bayesian Computation (ABC, Marin et al., 2012) has also been considered in this context, being much faster than MCMC alternatives, but also giving cruder approximations. Variational methods (Jordan et al., 1999) provide lower bounds for the marginal likelihoods and have been used for model selection in e.g. mixture models (McGrory & Titterton, 2007). Integrated nested Laplace approximation (INLA, Rue et al., 2009) provides estimates of marginal likelihoods within the class of latent Gaussian models and has become extremely popular. The reason for it is that Bayesian inference within INLA is extremely fast and remains at the same time reasonably precise.

A number of recent works build directly on INLA's reported marginal likelihoods. For example, Hubin & Storvik (2018) embed INLA marginal likelihoods within mode-jumping MCMC algorithms for Bayesian variable selection and model averaging in generalized linear mixed models; Hubin et al. (2020) additionally use them for latent process selection in binomial latent Gaussian models for DNA methylation; Gómez-Rubio et al. (2020) utilize INLA marginal likelihoods for Bayesian model averaging; Berild et al. (2022) combine INLA with importance sampling to estimate normalizing constants; and López-Gómez et al. (2026) rely on INLA marginal likelihoods for Bayesian model and variable selection. While these studies demonstrate the practical utility of INLA's marginal likelihood output, they largely take

its accuracy as an established premise or use it as an exploratory proposal. Earlier comparative surveys, such as Friel & Wyse (2012), benchmarked Laplace approximations, harmonic mean estimators, Chib's method, power posteriors, annealed importance sampling, and nested sampling, but did not include INLA in their comparisons. Consequently, systematic empirical evidence directly evaluating INLA's marginal likelihood approximations against exact analytical benchmarks, converged MCMC estimators, and thermodynamic alternatives has remained scarce. The main goal of this article is to fill this gap by evaluating the accuracy, numerical stability, and computational efficiency of INLA's marginal likelihood approximations on a decision-relevant scale.

INLA approximates the marginal likelihood by

$$p(\mathbf{y}|\mathcal{M}) \approx \int_{\Omega_{\theta}} \frac{p(\mathbf{y}, \theta, \boldsymbol{\eta}|\mathcal{M})}{\tilde{\pi}_G(\boldsymbol{\eta}|\mathbf{y}, \theta, \mathcal{M})} \Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*(\theta|\mathcal{M})} d\theta, \quad (6)$$

where $\boldsymbol{\eta}^*(\theta|\mathcal{M})$ is the mode of the full conditional of the latent field $\boldsymbol{\eta}$, while $\tilde{\pi}_G(\boldsymbol{\eta}|\mathbf{y}, \theta, \mathcal{M})$ is a Gaussian approximation to $\pi(\boldsymbol{\eta}|\mathbf{y}, \theta, \mathcal{M})$. Within the INLA framework, both random effects and regression parameters are treated as latent Gaussian variables, while the hyperparameter vector θ is typically of low dimension. The integration of θ over Ω_{θ} can be performed by an empirical Bayes (EB) approach, or by numerical integration using a central composite design (CCD) or a regular grid (see Rue et al., 2009, for details). INLA correspondingly reports two estimates: the integration-based marginal likelihood $\tilde{p}(\mathbf{y}|\mathcal{M})$ and the Gaussian approximation at the hyperparameter mode $\tilde{p}_G(\mathbf{y}|\mathcal{M})$.

In the following sections we evaluate the performance of INLA across a series of models of increasing complexity, beginning with an analytically tractable Gaussian model and continuing through Gaussian regression, probit and logit GLMs, a longitudinal Poisson GLMM, and two simulation scenarios specifically designed to stress the latent Gaussian approximation.

2 INLA versus truth and the harmonic mean

To begin with we address an extremely simple example suggested by Neal (2008), in which we consider the following model $\mathcal{M}$:

$$\begin{aligned} Y|\boldsymbol{\eta}, \mathcal{M} &\sim N(\boldsymbol{\eta}, \sigma_1^2); \\ \boldsymbol{\eta}|\mathcal{M} &\sim N(0, \sigma_0^2). \end{aligned} \quad (7)$$

Then, obviously, the marginal likelihood is available analytically as

$$Y|\mathcal{M} \sim N(0, \sigma_0^2 + \sigma_1^2),$$

and we have a benchmark to compare approximations to. The harmonic mean estimator (Raftery et al., 2007) is given by

$$p(\mathbf{y}|\mathcal{M}) \approx \frac{n}{\sum_{i=1}^n \frac{1}{p(\mathbf{y}|\boldsymbol{\eta}_i, \mathcal{M})}},$$

where $\boldsymbol{\eta}_i \sim p(\boldsymbol{\eta}|\mathbf{y}, \mathcal{M})$. This estimator is consistent, however often requires too many iterations to converge. Following Neal (2008), we evaluated a single observed datum $y = 2$ ($D = 2$) with observation standard deviation $\sigma_1 = 1$, varying the prior standard deviation across $\sigma_0 \in \{1000, 10, 0.1\}$. The harmonic mean is obtained based on $n = 10^7$ Monte Carlo simulations across 5 independent replications with fixed random seeds. For INLA we used R-INLA release 24.06.27 with default tuning parameters¹.

¹In this conjugate linear Gaussian example, INLA's approximation is exact by construction, so different hyperparameter integration strategies yield identical results.

Table 1 Comparison of INLA, harmonic mean (10^7 iterations, 5 replications), and exact analytical marginal likelihood. The column $D = 2$ indicates the observed value $y = 2$ in Neal (2008). The HM Error column reports the bias Mean – Exact on the log scale.

σ_0	σ_1	D	Exact	INLA	Harmonic Mean		HM Error
6-7 (lr)8-8					Mean	SD	(Mean – Exact)
1000	1	2	−7.8267	−7.8267	−2.5248	0.1359	+5.3019
10	1	2	−3.2463	−3.2463	−2.3509	0.0683	+0.8954
0.1	1	2	−2.9041	−2.9041	−2.9041	0.0001	0.0000

As shown in Table 1, INLA matches the analytical ground truth to four decimal places across all prior standard deviations. In contrast, the harmonic mean estimator suffers from severe infinite-variance pseudo-convergence: at $\sigma_0 = 1000$, its estimate of -2.5248 carries an error of $+5.3019$ log units against the analytical truth of -7.8267 (a Bayes factor error of over $\exp(5.3) \approx 200$), and at $\sigma_0 = 10$ it overestimates the evidence by $+0.8954$ log units. Only when the prior is strongly informative ($\sigma_0 = 0.1$) does the harmonic mean agree with the exact value. Due to this well-documented pathology (see also Neal, 2008), the harmonic mean will be dropped from most of the primary comparisons to follow.

3 INLA versus Chib's method in Gaussian Bayesian regression

In the second example we address INLA versus Chib's method (Chib, 1995) for the US crime data (Vandaele, 1978) based on the following model $\mathcal{M}$:

$$\begin{aligned}
 Y_t | \mu_t, \mathcal{M} &\stackrel{iid}{\sim} N(\mu_t, \sigma^2); \\
 \mu_t | \mathcal{M} &= \beta_0 + \sum_{i=1}^{p_{\mathcal{M}}} \beta_i x_{ti}^{\mathcal{M}}; \\
 \frac{1}{\sigma^2} | \mathcal{M} &\sim \text{Gamma}(\alpha_{\sigma}, \beta_{\sigma}); \\
 \beta_i | \mathcal{M} &\sim N(\mu_{\beta}, \sigma_{\beta}^2),
 \end{aligned} \tag{8}$$

where $t = 1, \dots, 47$ and $i = 0, \dots, p_{\mathcal{M}}$. We also addressed two different model variants, $\mathcal{M}_1$ and $\mathcal{M}_2$, induced by different sets of the explanatory variables with cardinalities $p_{\mathcal{M}_1} = 8$ and $p_{\mathcal{M}_2} = 11$ respectively.

In model (8), the hyperparameters were specified to $\mu_{\beta} = 0$, $\alpha_{\sigma} = 1$ and $\beta_{\sigma} = 1$. Because the regression coefficients β follow a conjugate multivariate Gaussian conditional distribution, they can be integrated out analytically. The residual precision $\tau = \sigma^{-2}$ can then be integrated via one-dimensional numerical quadrature over $\log \tau$, providing a benchmark against which both INLA and MCMC estimates can be rigorously validated.

For Chib's method, computations were performed using the `MCMCregress` function from the `MCMCpack` package (Martin et al., 2011) with `marginal.likelihood = "Chib95"` and 100 000 iterations per replication across 5 independent runs with fixed random seeds. INLA computations were performed using R-INLA (version 24.06.27) in default mode. For INLA, we used `strategy = "gaussian"` (which is exact for linear Gaussian models) and hyperparameter integration `int.strategy = "ccd"` over the residual precision τ . Different precisions σ_{β}^{-2} in the range $[0, 100]$ were evaluated to explore the behavior across prior specifications. Figure 1 shows the estimated marginal log-likelihoods for Chib's method (x -axis) and INLA (y -axis) for model $\mathcal{M}_1$ (left) and $\mathcal{M}_2$ (right), with the dashed line representing the identity line $y = x$. Across all 100 evaluated prior precisions, the median absolute discrepancy between INLA and the exact analytical benchmark is 0.0091 log units for $\mathcal{M}_1$ (95th percentile 0.0092, maximum 0.0092) and 0.0089 for $\mathcal{M}_2$ (95th percentile 0.0090, maximum 0.0091), while the discrepancy in the log Bayes factor $\log \text{BF}_{12}$ between INLA and the exact benchmark has a median of only 0.0002 (95th percentile 0.0004, maximum 0.0005) log units.

Table 2 reports numerical estimates across three standard prior standard deviations ($\sigma_{\beta} \in \{1000, 10, 0.1\}$), including the exact benchmark, the INLA estimate, the mean and standard deviation

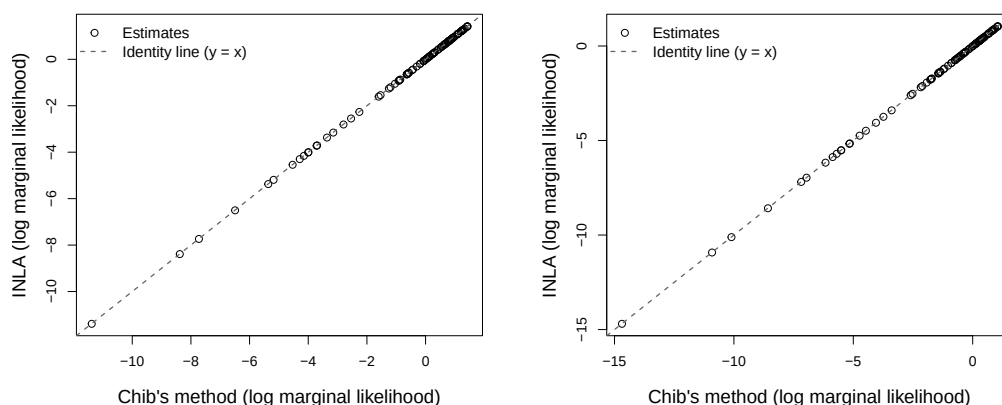


Figure 1 Chib's-INLA plots of estimated marginal log-likelihoods obtained by Chib's method (x -axis) and INLA (y -axis) for 100 different values of $\sigma_\beta^{-2} \in [0, 100]$ with the identity reference line ($y = x$). The left plot corresponds to model $\mathcal{M}_1$ while the right plot corresponds to model $\mathcal{M}_2$.

across 5 Chib replications, and the INLA error against truth. INLA matches the analytical benchmark to within 0.0005–0.0006 log units in all settings, while the Monte Carlo standard error of Chib's estimator is approximately 0.0002 log units.

Table 2 Comparison of exact analytical marginal likelihood, INLA (integration-based), and Chib's method (5 replications with 100 000 iterations) for US crime data. The Monte Carlo standard error (MCSE) is $SD/\sqrt{5}$.

$\mathcal{M}$	σ_β	Exact	INLA	Chib's method (10^5 iter)			Diff
5-7 (lr)8-8				Mean	SD	MCSE	(INLA – Exact)
$\mathcal{M}_1$	1000	−73.2093	−73.2098	−73.2091	0.0006	0.0003	−0.0005
$\mathcal{M}_1$	10	−31.7733	−31.7739	−31.7735	0.0006	0.0003	−0.0005
$\mathcal{M}_1$	0.1	1.4380	1.4374	1.4379	0.0003	0.0001	−0.0006
$\mathcal{M}_2$	1000	−96.6376	−96.6381	−96.6375	0.0007	0.0003	−0.0005
$\mathcal{M}_2$	10	−41.3990	−41.3996	−41.3988	0.0003	0.0001	−0.0005
$\mathcal{M}_2$	0.1	1.0627	1.0621	1.0625	0.0003	0.0001	−0.0006

Going further into MCMC convergence, Figure 2 shows the performance of Chib's method as a function of the number of iterations from 200 to 102 400 (10 independent replications per iteration level). The solid horizontal line indicates the INLA default estimate (−19.4768) for model $\mathcal{M}_1$ under precision $\sigma_\beta^{-2} = 0.2$. Crucially, the vertical axis in Figure 2 spans a very narrow interval of $[-19.494, -19.464]$ (a total span of only 0.030 log units). Hence, the apparent scatter in the plot reflects microscopic Monte Carlo variation, which rapidly stabilizes around the deterministic INLA estimate as iterations increase. All 100 convergence runs use deterministic seeds $20000 + 100i + j$ ($i, j \in \{1, \dots, 10\}$), recorded in `results/seeds.csv`.

From Figure 2 one can see that Chib's method requires substantial sampling effort to achieve high precision, whereas INLA returns a stable, near-exact value in a fraction of a second. The computational time for INLA is under 0.1 seconds for this model, while 100 000 iterations of Chib's method require approximately 1.5 seconds per run on the same hardware.

4 INLA versus Chib's method in probit regression

In the third example, we compare INLA with Chib's method (Chib, 1995) for approximating the marginal likelihood in a Bayesian probit regression model $\mathcal{M}$. The data set addressed is the simulated Bernoulli data introduced by Hubin & Storvik (2018). The model is given by

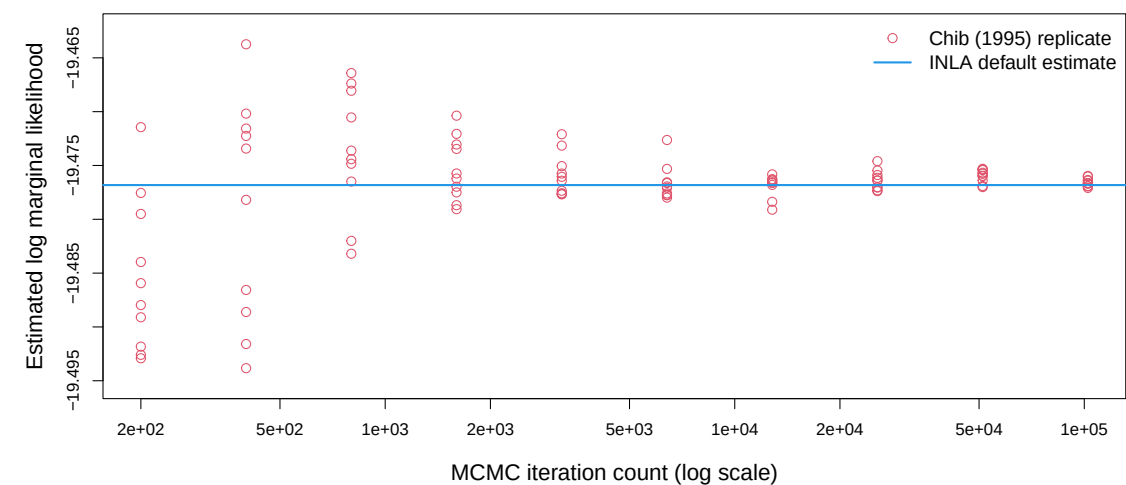


Figure 2 Convergence of Chib’s method across MCMC iteration counts (200 to 102 400, 10 runs per level) for the Gaussian regression model ($\mu_\beta = 1, \sigma_\beta^{-2} = 0.2$). The horizontal line marks the default INLA estimate (−19.4768). Note that the vertical axis spans only 0.030 log units (from −19.494 to −19.464).

$$\begin{aligned} Y_t | \eta_t, \mathcal{M} &\overset{iid}{\sim} \text{Bernoulli}(\Phi(\eta_t)); \\ \eta_t | \mathcal{M} &= \beta_0 + \sum_{i=1}^{p_{\mathcal{M}}} \beta_i x_{ti}^{\mathcal{M}}; \\ \beta_i | \mathcal{M} &\sim N(\mu_\beta, \sigma_\beta^2), \end{aligned}$$

where $t = 1, \dots, 2000$ and $i = 0, \dots, p_{\mathcal{M}}$. We addressed two different sets of explanatory variables with cardinalities 11 for model $\mathcal{M}_1$ and 13 for model $\mathcal{M}_2$.

For Chib’s method, computations were performed using the `MCMCprobit` function in `MCMCpack` (Martin et al., 2011) with `marginal.likelihood = "Chib95"` and 100 000 iterations per replication across 5 independent runs with fixed random seeds. For INLA, we used the binomial family with `link = "probit"`, using `strategy = "simplified.laplace"` and `int.strategy = "ccd"`. Since all prior precisions σ_β^{-2} were fixed across evaluations, no further hyperparameter optimization over regression priors was required.

Table 3 Comparison of INLA and Chib’s method for Bayesian probit regression across 5 independent replications with 100 000 iterations. MCSE = $\text{SD}/\sqrt{5}$.

$\mathcal{M}$	σ_β	INLA		Chib's method (10^5 iter, 5 reps)				Diff
4-7 (lr)8-8			Mean	SD	MCSE	Replicates		(INLA – Mean)
$\mathcal{M}_1$	1000	−688.3190	−688.2854	0.0489	0.0219	−688.30, −688.29, −688.36, −688.23, −688.24		−0.0336
$\mathcal{M}_1$	10	−633.0900	−633.0655	0.0239	0.0107	−633.05, −633.03, −633.09, −633.09, −633.07		−0.0245
$\mathcal{M}_1$	0.1	−669.7585	−669.7517	0.0129	0.0058	−669.77, −669.75, −669.74, −669.75, −669.74		−0.0068
$\mathcal{M}_2$	1000	−704.2263	−704.1979	0.0466	0.0208	−704.27, −704.20, −704.14, −704.19, −704.18		−0.0285
$\mathcal{M}_2$	10	−639.8049	−639.7549	0.0522	0.0233	−639.68, −639.77, −639.78, −639.72, −639.82		−0.0500
$\mathcal{M}_2$	0.1	−649.7799	−649.7816	0.0310	0.0139	−649.81, −649.80, −649.78, −649.78, −649.73		+0.0017

We used $\mu_\beta = 0$, varying the prior precisions continuously between 0 and 10 in Figure 3, while Table 3 evaluates $\sigma_\beta \in \{1000, 10, 0.1\}$ (precisions $10^{-6}, 10^{-2}$, and 10^2), extending beyond the figure’s range to

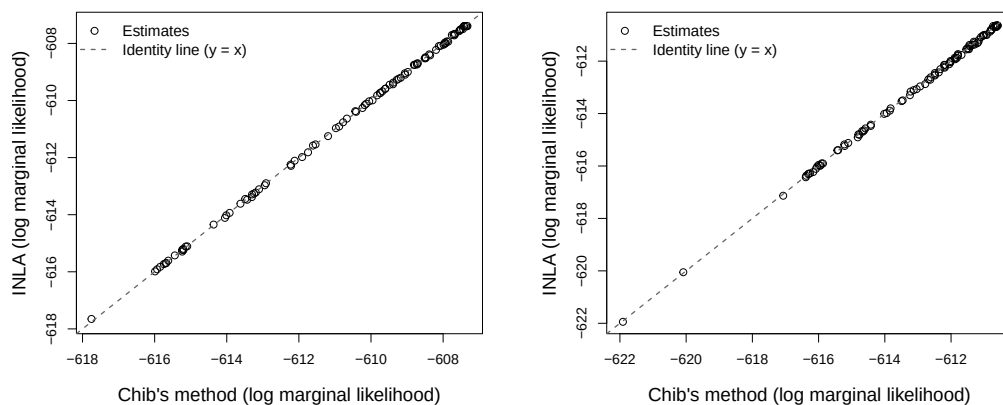


Figure 3 Comparisons of marginal log-likelihood estimates obtained by Chib's method (x -axis) and INLA (y -axis) for 100 different values of the precision parameter $\sigma_\beta^{-2} \in [0, 10]$ under 2 models with different numbers of covariates, with the identity reference line ($y = x$).

assess highly concentrated priors. In Figure 3, the identity reference line ($y = x$) is included. Across the 100 evaluated precision values, the median absolute discrepancy between INLA and Chib's method is 0.0193 for $\mathcal{M}_1$ and 0.0321 for $\mathcal{M}_2$ (95th percentiles 0.0692 and 0.0818), while the discrepancy in the log Bayes factor $\log \text{BF}_{12}$ has a median of 0.0350 and a maximum of 0.1347 log units.

The computational time for INLA within these examples is under 0.2 seconds per model, whereas 100 000 iterations of Chib's method required approximately 12.5 seconds per replication on the same hardware. Figure 4 displays the convergence of Chib's estimator across MCMC iterations (200 to 102 400, with 10 independent replications per iteration level using deterministic seeds $40000 + 100i + j$, $i, j \in \{1, \dots, 10\}$, recorded in `results/seeds.csv`).

5 INLA versus other methods in logistic regression

In the fourth example we compare marginal likelihoods obtained by INLA with seven alternative evidence estimators benchmarked by Friel & Wyse (2012): standard Laplace approximation, Chib and Jeliazkov's method, Laplace approximation at the MAP, harmonic mean estimator, power posteriors, annealed importance sampling (AIS), and nested sampling. The model $\mathcal{M}$ is the Bayesian logistic regression model with logit link addressed by Friel & Wyse (2012):

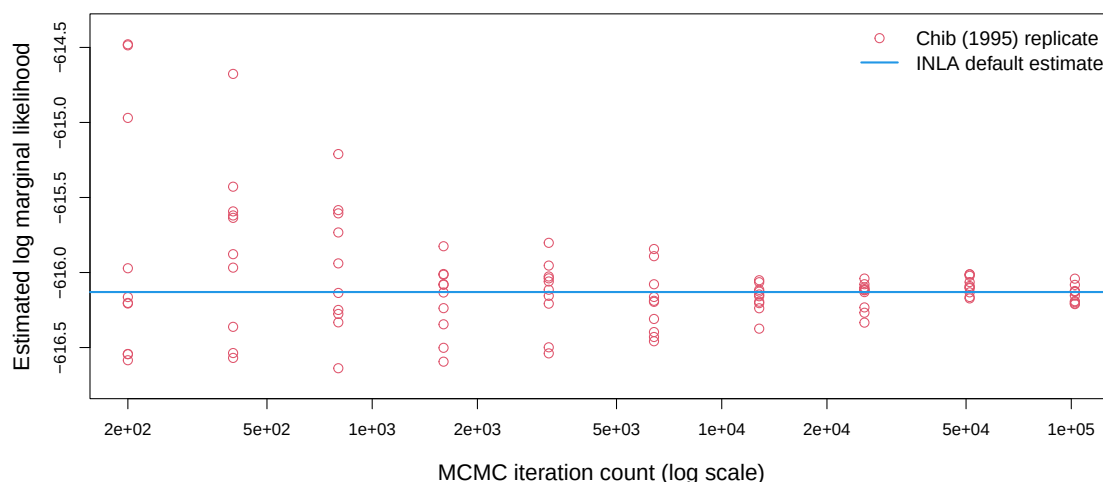


Figure 4 Convergence of Chib's method across MCMC iteration counts (200 to 102 400, 10 runs per level) for probit regression ($\mu_\beta = 1$, $\sigma_\beta^{-2} = 0.2$). The horizontal line marks the default INLA estimate (-616.1301). Note that the vertical axis spans 2.16 log units (from -616.64 to -614.48).

$$\begin{aligned}
Y_t | \eta_t, \mathcal{M} &\stackrel{iid}{\sim} \text{Bernoulli}(\text{logit}^{-1}(\eta_t)); \\
\eta_t | \mathcal{M} &= \beta_0 + \sum_{i=1}^{p_{\mathcal{M}}} \beta_i x_{ti}^{\mathcal{M}}; \\
\beta_i | \mathcal{M} &\sim N(\mu_{\beta}, \sigma_{\beta}^2),
\end{aligned}$$

where $t = 1, \dots, 532$ and $i = 0, \dots, p_{\mathcal{M}}$. The data set addressed is the Pima Indians diabetes data (532 complete records). For $\mathcal{M}_1$, predictors comprise the number of pregnancies, plasma glucose concentration, body mass index, and diabetes pedigree function ($p_{\mathcal{M}_1} = 4$); $\mathcal{M}_2$ additionally includes age ($p_{\mathcal{M}_2} = 5$). All covariates were standardized to zero mean and unit variance. The models were evaluated under two prior variances: $\sigma_{\beta}^2 = 100$ and $\sigma_{\beta}^2 = 1$ with $\mu_{\beta} = 0$.

Table 4 presents the results. Estimates for methods 3–9 are reproduced from the published benchmark of Friel & Wyse (2012) on the identical standardized dataset, where each Monte Carlo method used the equivalent of 200 000 samples. For INLA, we used `strategy = "simplified.laplace"` and `int.strategy = "ccd"` with fixed prior variances $\sigma_{\beta}^2 \in \{100, 1\}$. To establish an independent ground truth, we also computed an importance sampling (IS) benchmark based on 200 000 draws from an inflated multivariate Student- t_6 proposal distribution centred at the posterior mode with scale matrix matching the inverse Hessian, achieving a Monte Carlo standard error of MCSE = 0.003.

Table 4 Comparison of INLA, independent importance sampling reference, and seven baseline methods for logistic regression on Pima Indians data. Methods 3–9 are cited directly from the benchmark study of Friel & Wyse (2012). $\log \text{BF}_{12} = \log m(\mathbf{y} | \mathcal{M}_1) - \log m(\mathbf{y} | \mathcal{M}_2)$.

	$\sigma_{\beta}^2 = 100$			$\sigma_{\beta}^2 = 1$		
2-4 (lr)5-7 Method	$\mathcal{M}_1$	$\mathcal{M}_2$	$\log \text{BF}_{12}$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\log \text{BF}_{12}$
INLA (classic mode)	−257.25	−259.89	2.64	−247.32	−247.59	0.27
Importance sampling (reference)	−257.23	−259.86	2.63	−247.30	−247.56	0.26
Laplace approximation	−257.26	−259.89	2.63	−247.33	−247.59	0.26
Chib and Jeliazkov	−257.23	−259.84	2.61	−247.31	−247.58	0.27
Laplace approximation MAP	−257.28	−259.90	2.62	−247.33	−247.62	0.29
Harmonic mean estimator	−284.78	−279.47	−5.31	−259.84	−260.55	0.71
Power posteriors	−257.98	−260.59	2.61	−247.57	−247.84	0.27
Annealed importance sampling	−257.87	−260.43	2.56	−247.30	−247.59	0.29
Nested sampling	−258.82	−261.38	2.56	−246.82	−246.97	0.15

A noticeable spread appears among the competing methods in Table 4; nested sampling sits 1.57 log units below INLA for $\mathcal{M}_1$ under $\sigma_{\beta}^2 = 100$ and 0.50 log units above INLA under $\sigma_{\beta}^2 = 1$, while power posteriors and AIS differ by 0.6 to 0.7 log units. It is essential to address what this spread implies for applied inference. First, our independent importance sampling reference matches INLA within 0.02 log units under both prior settings (−257.23 vs. −257.25 at $\sigma_{\beta}^2 = 100$, and −247.30 vs. −247.32 at $\sigma_{\beta}^2 = 1$), confirming that INLA's Laplace approximation is near-exact for this GLM, whereas nested sampling and AIS exhibit small Monte Carlo underestimation biases under diffuse priors.

Second, because stochastic errors in nested sampling, power posteriors, and AIS are strongly positively correlated across models $\mathcal{M}_1$ and $\mathcal{M}_2$, the discrepancy largely cancels on the decision-relevant scale of log Bayes factors. For $\sigma_{\beta}^2 = 100$, INLA yields $\log \text{BF}_{12} = 2.64$, compared to 2.63 for IS, 2.63 for Laplace, 2.61 for Chib–Jeliazkov, 2.61 for power posteriors, and 2.56 for nested sampling and AIS, with a maximum across-method difference of only 0.08 log units. Under $\sigma_{\beta}^2 = 1$, all methods except the harmonic mean yield $\log \text{BF}_{12} \in [0.15, 0.29]$, uniformly indicating that the evidence in favor of $\mathcal{M}_1$ is negligible. Thus, model-selection and model-averaging decisions are entirely preserved across methods. In terms of computational cost, INLA required under 0.05 seconds, compared to substantial sampling budgets (200 000 evaluations) for the competing stochastic estimators (Friel & Wyse, 2012).

6 INLA versus Chib and Jeliaskov's method in a Poisson mixed model

As models become more sophisticated, fewer methodologies are available for approximating the marginal likelihood. In the context of generalized linear mixed models (GLMMs), we consider the INLA approach (Rue et al., 2009) alongside the Metropolis–Hastings marginal likelihood estimator of Chib & Jeliaskov (2001) and the kernel density estimator of Chib et al. (1998).

This model analyzes epileptic seizure counts Y_{jt} for 59 patients measured first over an 8-week baseline period ($t = 0$) and subsequently across four 2-week follow-up periods ($t = 1, \dots, 4$). After baseline, patients were randomized to either active treatment or placebo. Following standard practice for these data (Chib et al., 1998; Chib & Jeliaskov, 2001; Diggle et al., 1994), patient 49 (with an extreme baseline count of 151) was omitted as an outlier, leaving 58 patients and 290 total observations. The model $\mathcal{M}$ is specified as:

$$\begin{aligned} Y_{jt} | \eta_{jt}, \mathcal{M} &\sim \text{Poisson}(\exp(\eta_{jt})); \\ \eta_{jt} | \mathcal{M} &= \log(\tau_{jt}) + \beta_0 + \beta_1 x_{jt1} + \beta_2 x_{jt2} + \beta_3 x_{jt1} x_{jt2} + b_{j0} + b_{j1} x_{jt1}; \\ \beta_i | \mathcal{M} &\sim N(0, 100), \quad i = 0, \dots, 3; \\ \mathbf{b}_j | \mathbf{D}, \mathcal{M} &\sim N_2(0, \mathbf{D}); \\ \mathbf{D}^{-1} | \mathcal{M} &\sim \text{Wishart}_2(4, I_2), \end{aligned}$$

for $j = 1, \dots, 58$ and $t = 0, \dots, 4$. Here $x_{jt1} \in \{0, 1\}$ indicates the post-baseline period, $x_{jt2} \in \{0, 1\}$ denotes treatment assignment, τ_{jt} is the exposure offset (8 for baseline, 2 for visits), and $\mathbf{b}_j = (b_{j0}, b_{j1})'$ are patient-specific bivariate random effects with covariance matrix $\mathbf{D}$.

In this example, the hyperparameter vector θ comprises the three non-redundant parameters of the precision matrix $\mathbf{D}^{-1}$, meaning that numerical integration over hyperparameters requires three-dimensional quadrature. INLA produces both an integration-based estimate $\tilde{p}(\mathbf{y} | \mathcal{M})$ and a Gaussian approximation $\tilde{p}_G(\mathbf{y} | \mathcal{M})$ evaluated at the hyperparameter mode. Table 5 summarizes the estimates across different INLA approximation strategies (Gaussian, simplified Laplace, full Laplace) and integration schemes (CCD, grid, empirical Bayes), alongside the published MCMC reference values.

Table 5 Comparison of INLA marginal log-likelihood estimates across approximation strategies and hyperparameter integration schemes against literature MCMC benchmarks for the longitudinal Poisson GLMM on epileptic seizure counts.

Method / Strategy	Hyperparameter Integration	$\tilde{p}(\mathbf{y} \mathcal{M})$ (Integration)	$\tilde{p}_G(\mathbf{y} \mathcal{M})$ (Gaussian)
INLA (Default: Simplified Laplace)	CCD	−914.67	−915.62
INLA (Gaussian approximation)	CCD	−914.67	−915.62
INLA (Full Laplace)	CCD	−914.67	−915.62
INLA (Simplified Laplace)	Grid	−915.26	−915.62
INLA (Full Laplace)	Grid	−915.26	−915.62
INLA (Full Laplace)	Fine Grid ($dz = 1.0$)	−915.83	−915.62
INLA (Empirical Bayes mode)	None (EB)	−918.67	−915.62
Chib & Jeliaskov (2001) [JASA]	MCMC (10 runs, SE = 0.12)		−915.49
Chib, Greenberg & Winkelmann (1998)	MCMC (Kernel density)		−915.23

As shown in Table 5, the Gaussian approximation row $\tilde{p}_G(\mathbf{y} | \mathcal{M}) = -915.62$ reproduces the value of -915.61 reported in the initial preprint and agrees closely with Chib's Metropolis–Hastings benchmark of -915.49 ± 0.12 (Chib & Jeliaskov, 2001) and the kernel density estimate of -915.23 (Chib et al., 1998). Under default CCD integration over the 3-dimensional hyperparameter space, the integration-based estimate yields -914.67 . When switching to grid integration over θ , the integration estimate moves to -915.26 , and with a refined grid step ($dz = 1.0$) reaches -915.83 , closely bracketing the MCMC benchmarks. Empirical Bayes (which omits the hyperparameter volume integration) underestimates the evidence at -918.67 . Note that in R-INLA, the strategy argument ("gaussian", "simplified.laplace", "laplace") governs the approximation of the univariate posterior marginals of the latent field, whereas the marginal likelihood in (6) always evaluates the Gaussian denominator $\tilde{\pi}_G(\boldsymbol{\eta} | \mathbf{y}, \boldsymbol{\theta}, \mathcal{M})$ at the mode $\boldsymbol{\eta}^*(\boldsymbol{\theta})$; conse-

quently, $\tilde{p}(\mathbf{y}|\mathcal{M})$ in Tables 5 and 7 varies with the hyperparameter quadrature (`int.strategy`) rather than the latent marginal strategy.

This comparison offers a practical diagnostic rule for applied Bayesian practitioners: in mixed models with multiple hyperparameters, comparing the default CCD integration estimate against grid integration (or against the Gaussian approximation row) serves as an internal consistency check. A difference exceeding 0.5 log units indicates that the hyperparameter posterior surface is moderately non-elliptical, and that grid integration or smaller step sizes should be selected. Across R-INLA releases, minor quadrature updates may slightly shift integration estimates (e.g., under R-INLA 25.06.07, the CCD estimate is -914.77 and grid is -915.26 , yielding a difference of 0.49 log units). The 0.5 log-unit threshold should therefore be understood as a practical rule-of-thumb rather than a sharp boundary. Importantly, this internal check specifically assesses the numerical stability and ellipticity of the hyperparameter integration over θ . It does not diagnose potential inaccuracies in the Laplace approximation of the latent field $\boldsymbol{\eta}$ itself, which must instead be evaluated by assessing the sample sizes and event counts per latent node. The computational time for INLA in this GLMM was approximately 1.8 seconds for CCD integration and 12 seconds for grid integration, compared to several minutes of MCMC sampling required for Chib’s estimator.

7 Decision-Scale Evaluation: Bayes Factors and Model Probabilities

In applied model selection and Bayesian model averaging, what matters to the practitioner is not agreement to arbitrary decimal places, but whether approximation errors alter model selection decisions or distort posterior model probabilities $p(\mathcal{M}_k|\mathbf{y})$ in equations (3) and (4). In Table 6, we evaluate the discrepancy between INLA and reference estimators on the decision-relevant scale of log Bayes factors $\log \text{BF}_{12} = \log m(\mathbf{y}|\mathcal{M}_1) - \log m(\mathbf{y}|\mathcal{M}_2)$ and posterior model probabilities $P(\mathcal{M}_1|\mathbf{y})$ (assuming equal prior odds, $P(\mathcal{M}_1) = P(\mathcal{M}_2) = 0.5$). Evidence strength is categorized according to the standard interpretive scale of Kass & Raftery (1995): $|\log \text{BF}| < 1$ (barely worth mentioning), 1–3 (positive evidence), 3–5 (strong evidence), and > 5 (very strong evidence).

Table 6 Decision-scale comparison between INLA and reference benchmarks (Exact for US Crime, Chib for Probit, Importance Sampling for Pima). Across all scenarios, the absolute error in log Bayes factor satisfies $|\Delta \log \text{BF}| \leq 0.026$ (within the Monte Carlo standard error of the MCMC reference), shifts posterior model probabilities by at most 0.002, and preserves the Kass & Raftery (1995) evidence categorization identically.

Dataset	Prior	$\log \text{BF}_{\text{Ref}}$	$\log \text{BF}_{\text{INLA}}$	$\Delta \log \text{BF}$	$P(\mathcal{M}_1 \mathbf{y})_{\text{Ref}}$	$P(\mathcal{M}_1 \mathbf{y})_{\text{INLA}}$	Evidence Category
US Crime	$\sigma_\beta = 1000$	23.428	23.428	0.000	1.0000	1.0000	Very strong
US Crime	$\sigma_\beta = 10$	9.626	9.626	0.000	0.9999	0.9999	Very strong
US Crime	$\sigma_\beta = 0.1$	0.375	0.375	0.000	0.5927	0.5927	Not worth mentioning
Sim Probit	$\sigma_\beta = 1000$	15.912	15.907	-0.005	1.0000	1.0000	Very strong
Sim Probit	$\sigma_\beta = 10$	6.689	6.715	$+0.026$	0.9987	0.9988	Very strong
Sim Probit	$\sigma_\beta = 0.1$	-19.970	-19.979	-0.009	0.0000	0.0000	Very strong (for $\mathcal{M}_2$)
Pima Logit	$\sigma_\beta^2 = 100$	2.626	2.636	$+0.010$	0.9325	0.9332	Positive
Pima Logit	$\sigma_\beta^2 = 1$	0.260	0.268	$+0.008$	0.5646	0.5666	Not worth mentioning

Table 6 demonstrates that across all eight evaluated regression scenarios, INLA’s error on the log Bayes factor scale never exceeds 0.026 log units (0.010 against exact and importance sampling benchmarks), and shifts posterior model probabilities by at most 0.002 (two-tenths of one percent). In every single case, the scientific conclusion and the qualitative strength of evidence are completely identical to those obtained from the reference methods.

8 Probing the Limits: Non-Gaussian and Small-Cluster Regimes

To further probe the boundaries of INLA’s marginal likelihood reliability, we designed two stress scenarios featuring known challenges for Laplace approximations:

- (i) **Scenario S1 (Overdispersed Poisson GLMM):** $n = 50$ observations generated with a low baseline rate ($\beta_0 = -1.5, \beta_1 = 0.8, x_i \in [-1, 1]$), resulting in 68% zeros, with an observation-level random effect $u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_u^2)$ ($\sigma_u = 0.7$). With small discrete counts, the Poisson likelihood is highly skewed, challenging the latent Gaussian assumption.
- (ii) **Scenario S2 (Small-Cluster Binary Logistic GLMM):** $J = 20$ clusters of a very small size of 3, with binary response $Y_{jk} \sim \text{Bernoulli}(\text{logit}^{-1}(\beta_0 + \beta_1 x_{jk} + u_j))$ and substantial random intercept variance ($\sigma_u = 1.2$). With only three binary trials per cluster, each latent variable u_j is weakly identified and non-Gaussian.

In both scenarios, conditional on the fixed effects and hyperparameter $\tau_u = \sigma_u^{-2}$, the latent variables are conditionally independent across clusters or observations. Consequently, the inner integral factors into a product of one-dimensional integrals, which can be evaluated to machine precision via 50-point Gauss-Hermite quadrature. Integrating the resulting conditional likelihood over $(\beta_0, \beta_1, \log \tau_u)$ on a fine $35 \times 35 \times 35$ adaptive grid provides a ground truth (error bounded with $< 10^{-4}$). Table 7 reports the results.

Table 7 Stress tests probing INLA marginal likelihood accuracy against 50-point Gauss-Hermite quadrature benchmarks under sparse, non-Gaussian regimes.

Scenario	INLA Configuration	INLA log $m(\mathbf{y})$	Exact Reference	Discrepancy (Δ)
S1: Low-Count Poisson (50 obs, 68% zeros, iid RE)	Default (Simplified Laplace + CCD)	−47.5829	−47.4815	−0.1014
	Gaussian Approximation	−47.5829	−47.4815	−0.1014
	Full Laplace + CCD	−47.5829	−47.4815	−0.1014
	Full Laplace + Grid	−47.6319	−47.4815	−0.1504
S2: Small-Cluster Binary (20 clusters of size 3, $N = 60$)	Default (Simplified Laplace + CCD)	−42.9984	−42.2332	−0.7652
	Gaussian Approximation	−42.9984	−42.2332	−0.7652
	Full Laplace + CCD	−42.9984	−42.2332	−0.7652
	Full Laplace + Grid	−42.9602	−42.2332	−0.7269

In Scenario S1, despite 68% zero counts, INLA’s marginal likelihood approximation remains reasonably robust, incurring an underestimation bias of only 0.10 to 0.15 log units. In Scenario S2, however, the combination of binary outcomes and tiny cluster sizes of 3 pushes the Laplace approximation to its limits, resulting in a persistent underestimation of approximately 0.73 to 0.77 log units across all INLA configurations. This demonstrates that when binary data provide sparse information per latent Gaussian node, the Laplace approximation incurs a noticeable error.

9 Conclusions

The marginal likelihood is a cornerstone of Bayesian model selection, hypothesis testing, and model averaging. In this study, we conducted a systematic empirical evaluation of INLA’s marginal likelihood approximations against exact analytical benchmarks, converged MCMC estimators (Chib 1995, Chib and Jeliazkov 2001), independent importance sampling references, and thermodynamic competitors (Friel & Wyse, 2012).

Our findings indicate that for linear Gaussian models, probit and logit regressions, and GLMMs with moderate cluster sizes, INLA’s reported marginal likelihoods are remarkably accurate. On the decision-relevant scale of log Bayes factors, INLA’s discrepancies from reference benchmarks are negligible, which preserves scientific model-selection decisions without alteration. At the same time, INLA achieves speedups of two to three orders of magnitude over competing MCMC approaches.

However, our investigation also delineates the boundaries of this reliability. Under sparse latent structures, such as binary clustered data with very small cluster sizes, INLA’s Laplace approximation

to the latent field can underestimate the marginal log-likelihood. Similarly, in mixed models featuring three or more hyperparameters, the default central composite design (CCD) integration can deviate from fine-grid or MCMC benchmarks. Consequently, modellers should routinely compare INLA's default CCD integration estimate against grid integration or the Gaussian approximation row as an internal diagnostic check; a discrepancy exceeding 0.5 log units indicates that the hyperparameter posterior surface is moderately non-elliptical and that grid integration or finer step sizes should be employed. Practitioners should recognize that this diagnostic isolates sensitivity to hyperparameter integration over θ ; it does not detect inaccuracies stemming from the Gaussian approximation $\tilde{\pi}_G(\boldsymbol{\eta} \mid \mathbf{y}, \theta, \mathcal{M})$ to the latent field in (6), which are governed by data discreteness and cluster sparsity.

When used with these diagnostics in mind, INLA provides a trustworthy, computationally formidable framework for Bayesian model selection and model averaging.

Data and Code Availability

Reproducibility Package: All code, datasets, and execution scripts are organized in a self-contained reproducibility package hosted at the canonical repository: <https://github.com/aliaksah/INLA-marginal-likelihood>. (The original exploratory scripts from 2016 remain archived at <https://github.com/aliaksah/EMJMCMC2016>). Across modern computational platforms (Apple Silicon ARM64 versus x86_64 Linux) and R-INLA releases, all stochastic estimators use fixed, logged random seeds (recorded in `results/seeds.csv`), and automated continuous integration workflows (`.github/workflows/reproduce.yml`) verify numerical reproducibility. A complete `sessionInfo()` environment log from R 4.4.1 and R-INLA 24.06.27 is archived in `logs/`. A comprehensive `README.md` and master driver script `run_all.R` are provided in the repository to map every script (00 through 05) directly to the generated tables, figures, and numerical logs.

Conflict of Interest Statement

The authors declare that they have no competing interests.

Acknowledgments

We would like to thank the CELS project at the University of Oslo and Integreat – Norwegian Centre for Knowledge-driven Machine Learning for giving us the opportunity, inspiration and motivation to write this article.

References

- Berild, M. O., Martino, S., Gómez-Rubio, V., & Rue, H. (2022). Importance Sampling with the Integrated Nested Laplace Approximation. *Journal of Computational and Graphical Statistics*, 31(4), 1225–1237. <https://doi.org/10.1080/10618600.2022.2067551>
- Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90(432), 1313–1321. <https://doi.org/10.1080/01621459.1995.10476635>
- Chib, S., Greenberg, E., & Winkelmann, R. (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics*, 86(1), 33–54. [https://doi.org/10.1016/s0304-4076\(97\)00108-5](https://doi.org/10.1016/s0304-4076(97)00108-5)
- Chib, S., & Jeliazkov, I. (2001). Marginal Likelihood From the Metropolis–Hastings Output. *Journal of the American Statistical Association*, 96(453), 270–281. <https://doi.org/10.1198/016214501750332848>
- Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford University Press, Oxford.
- Friel, N., & Wyse, J. (2012). Estimating the evidence – a review. *Statistica Neerlandica*, 66(3), 288–308. <https://doi.org/10.1111/j.1467-9574.2011.00515.x>
- Gómez-Rubio, V., Bivand, R. S., & Rue, H. (2020). Bayesian Model Averaging with the Integrated Nested Laplace Approximation. *Econometrics*, 8(2), 23. <https://doi.org/10.3390/econometrics8020023>
- Hubin, A., & Storvik, G. (2018). Mode jumping MCMC for Bayesian variable selection in GLMM. *Computational Statistics & Data Analysis*, 127, 281–297. <https://doi.org/10.1016/j.csda.2018.05.020>

- Hubin, A., Storvik, G. O., Grini, P. E., & Butenko, M. A. (2020). A Bayesian Binomial Regression Model with Latent Gaussian Processes for Modelling DNA Methylation. *Austrian Journal of Statistics*, 49(4), 46–56. <https://doi.org/10.17713/ajs.v49i4.1124>
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2), 183–233. <https://doi.org/10.1023/a:1007665907178>
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- López-Gómez, D., Gómez-Rubio, V., & García-Donato, G. (2026). Efficient model exploration with the integrated nested Laplace approximation. *arXiv Preprint arXiv:2609.07283*.
- Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6), 1167–1180. <https://doi.org/10.1007/s11222-011-9288-2>
- Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack : Markov Chain Monte Carlo in R. *Journal of Statistical Software*, 42(9). <https://doi.org/10.18637/jss.v042.i09>
- McGrory, C. A., & Titterton, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51(11), 5352–5367. <https://doi.org/10.1016/j.csda.2006.07.020>
- Neal, R. M. (2008). *The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever*. <https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/>.
- Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1), 3–26. <https://doi.org/10.1111/j.2517-6161.1994.tb01956.x>
- Raftery, A. E., Newton, M. A., Satagopan, J. M., & Krivitsky, P. N. (2007). Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. In *Bayesian Statistics 8* (pp. 371–416). Oxford University Press. <https://doi.org/10.1093/oso/9780199214655.003.0015>
- 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. <https://doi.org/10.1111/j.1467-9868.2008.00700.x>
- Tierney, L., & Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393), 82–86. <https://doi.org/10.1080/01621459.1986.10478240>
- Vandaele, W. (1978). Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates* (pp. 270–335). National Academy of Sciences, Washington, DC.

COLOPHON Typeset with Typst 0.15.0 from the authors' LaTeX source. Template jrepstat.typ at commit 54635ba-dirty. Set in STIX Two Text, STIX Two Math and JetBrains Mono. Built 26 October 2026.